University of Southampton

Faculty of Environmental and Life Sciences

School of Biological Science

Genomic constraints on domestication: the role of plasticity and transposable elements

by

Anne Jeanette Romero

ORCID ID 0000-0003-3700-7225

Thesis for the degree of Doctor of Philosophy (PhD)

September 2024

University of Southampton

Abstract

Faculty of Environmental and Life Sciences

School of Biological Sciences

Thesis for the degree of Doctor of Philosophy (PhD)

Genomic constraints on domestication: the role of plasticity and transposable elements

by

Anne Jeanette Romero

Crop domestication is an important evolutionary process that transforms wild plants into cultivated crops, facilitating our shift from foraging to agriculture. Despite our reliance on domesticated crops, only a small proportion of edible plant species are domesticated. Is there a genomic constraint on domestication that allows the domestication of certain species over others? Our understanding of crop domestication has focused on the selection that transformed the progenitor into the domesticated crop, whereas little is known about the selection between wild species in early domestication. Here, we investigate the role of plasticity and transposable elements (TEs) on the selective advantage of the tomato progenitor over never-domesticated wild species (referred here as 'wilds'). Plasticity is the ability of an organism to respond to new environments. Phenotypic and gene expression plasticity were assessed in domesticated, progenitor and wild species. A greater number of traits and genes were plastic in the progenitor than in the wild species, linked to important fruit traits and plant processes. Underlying genetic diversity may have contributed to this enhanced plasticity. The ability of TEs to move from one location of the genome to another makes them a great contributor to diversity generation. Annotation of single nucleotide polymorphism (SNP) and transposon insertion polymorphism (TIP) to characterise genetic diversity revealed greater nucleotide and TIP diversity in the progenitor than in wild species with evidence of TIPs associated with genes that were putatively selected during domestication. Since mutation rates underpin the maintenance of high genetic diversity, we employed mutation accumulation (MA) lines to estimate the haploid mutation rate for single nucleotide variants (SNVs), indels and TE insertions. SNV and indel mutation rates were higher in the progenitor than in wild species, although there was no detectable difference in TE insertion rates. We provide the first mutation accumulation experiment to estimate mutation rates in tomatoes. Overall, we found evidence for the role of plasticity, genetic diversity and mutation rates in the domestication of the tomato progenitor. Uncovering genomic mechanisms that facilitate domestication could identify adaptive variation in crop wild relatives and could be important in crop breeding to tackle food security challenges.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

List of Accompanying Materials

Supplementary Tables A

Metadata of the plasticity pot experiment and RNA sequence processing; phenotypic traits measured and statistical outputs; differentially expressed genes (DEGs) between species and significant Gene Ontology (GO) terms and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways associated with DEGs between species. This includes Tables A1 to A26.

Supplementary Tables B

Resequencing details with quality and mapping statistics; Genome size estimation. Single nucleotide polymorphism (SNP) statistics; PopoolationTE2 outputs; Extensive de novo TE Annotator (EDTA) outputs; Summary of Transposon insertion polymorphism (TIP) analysis outputs; Gene ontology (GO) analysis outputs. This includes Tables B1 to B17.

Supplementary Tables C

Summary of sequencing and mapping statistics; Summary of mutation counts and rates; Lists of single nucleotide variants (SNVs), indels and Transposable elements (TE) insertions identified. This includes Tables C1 to C5.

Research Thesis: Declaration of Authorship

Print name: Anne Romero

Title of thesis: Genomic constraints on domestication: the role of plasticity and transposable elements

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature: Date: 20/03/2025

Acknowledgements

Thank you.

My whole PhD experience has been a privilege.

First and foremost, I would like to thank my supervisor Professor Mark Chapman for encouraging me to pursue a PhD and believing that I could do it. I am extremely grateful for all your positive support and guidance. I extend my deepest appreciation to my other supervisors Professor Adam Eyre-Walker and Professor Tom Ezard. Thank you all for the advice, invaluable insights, and exceptional proofreading, that shaped this thesis.

Additional thanks to the technical team. Research would not run smoothly without you and a special thanks to Mike for always being happy to help me in the glasshouse. I acknowledge the use of the IRIDIS High-Performance Computing Facility and associated support services for their technical expertise.

I would also like to thank my funders BBSRC through the SoCoBio DTP and the University of Southampton. A yearly highlight of my PhD was the SoCoBio summer school that did not just provide incredible training in business, science communication and industry but also delivered a whole lot of fun. They have provided opportunities for personal and professional development, especially with my PIPs which set me up with future opportunities.

To my lab group, past and present, your support and encouragement have made this journey joyful. It was a pleasure to do my PhD alongside so many enthusiastic researchers. Shout out to the ‘Lv6 Heroes’!

I am grateful for the constant support and love of my friends and family. To JV, Kate and Jimboy, you all have been such a rock, thank you for all the take-away incentives to keep me going! A special mention to my parents, I owe this academic achievement to both of you, thank you for teaching me the value of hard work and never failing to encourage me to go further.

Lastly, to my ever-patient partner, V, who has been through it all with me. Thank you for sharing your expertise in Excel, storytelling and proofreading. Your love, humour and endless pep talk got me through my PhD. I can’t wait to celebrate with you and the peeps!

Definitions and Abbreviations

bp ..... Base pair

DEGs ..... Differentially expressed genes

DNA ..... Deoxyribonucleic acid

Kb ..... Kilobases

M ..... Million

Mb ..... Megabases

PCA ..... Principal component analysis

RNA ..... Ribonucleic acid

SNP ..... Single Nucleotide Polymorphism

subsp. ..... Subspecies

TE ..... Transposable elements

TIP ..... Transposon Insertion Polymorphism

var. ..... Variety

Chapter 1 Introduction

1.1 Domestication

Crop domestication is an evolutionary process that transforms wild plant species into domesticated species, adapted for human cultivation and use (Gepts and Papa, 2001). Darwin (1868) used domestication as a model to understand evolutionary processes such as variation and selection in domesticated crops. This has also served as a model in genetic and recent evolutionary adaptations (Ross-Ibarra *et al.*, 2007; Andersson and Purugganan, 2022).

A mutualistic relationship forms as humans domesticate wild plant species to adapt to anthropogenic environments: humans take control of the plant's reproduction and create a controlled environment for them to thrive in, in turn, plants provide valuable resources (Purugganan, 2022). There is evidence that hundreds of wild plants were collected and cultivated as a food source in early domestication, but later abandoned (Wallace *et al.*, 2019). So, even though half of the 390,000 plant species on Earth are edible (Willis, 2017), only a few hundred have been domesticated (Zeven and De Wet, 1982). What's more, we rely on only 15 crops to provide 90% of our calorie intake (FAO, 2017). This dependence on a few domesticated crops is rooted in the limited number of species domesticated in the first place. All this begs the question: Why are some species domesticated and others are not?

The selection of plant species for domestication could have been a conscious decision by early farmers, it could have been influenced by natural selection where crop progenitors had an advantage over non-domesticated wild relatives prior to cultivation. Selection during domestication occurred in two stages: selection between wild species and the selection that transformed the crop progenitor into domesticated crops (Jones *et al.*, 2021). A lot is known about the latter however, our understanding of the former is limited.

1.1.1 Emergence of domesticated plants

Barley, wheat and a few pulses are among the first crops thought to be domesticated approximately 12,000 years ago (Purugganan and Fuller, 2009). They were domesticated in the Fertile Crescent, one of many centres of plant domestication, where wild plants were domesticated independently and synchronously. Others include the Mesoamerica (notably maize, squash and common bean), the Andean region (notably potato and tomato), and South East Asia (notably rice, millet and soybean).

The emergence of domesticated plants marks the transition from wild species into a phenotypically different taxon, accumulating genetic changes through conscious and unconscious selection (Zohary, 2004). These characteristics selected during domestication are called the 'domestication syndrome', they result in differentiation between wild progenitors and domesticated crops (Hammer, 1984; Fuller, 2007). Common phenotypes evolved independently through convergent evolution across many domesticated species, including traits associated with loss of seed dispersal, increase in seed and fruit size and seasonality control (Fuller *et al.*, 2014). Genetic architectures of many domestication phenotypes were revealed to be more complex and influenced by multiple genes (Xue *et al.*, 2016; Ishikawa *et al.*, 2022), with some traits controlled by one or a few genes with large effects (Doebley *et al.*, 2006). Although these resulted in the increased fitness of domesticated species due to higher production under cultivation, there were also undesirable consequences of unintended selection of traits and genes, such as reduced plant immunity, as well as the organoleptic quality of the tomato fruit being much higher in the domesticated than in the wild species (Singh and van der Knaap, 2022). Early domesticates are genetically and phenotypically heterogeneous, known as landraces, associated with traditional farming systems and adapted to local conditions and food preferences (FAO, 2019).

Due to the success of domesticated crops, evident by their spread across the planet, human dependence on these few species increased. This brings benefits such as continuous food supply and various crop breeding strategies, which increase food production. Domesticated species represent a small fraction of the diversity found in wild species, as domestication often comes with a reduction in genetic diversity. The lack of evidence of genetic loss in ancient samples, suggests a domestication bottleneck occurred more recently (Blanca et al., 2015; Allaby et al., 2022). Domestication bottleneck involves the reduction and recovery in population size that lowers genetic diversity and increases mutation load as a result of genetic drift (Moyers *et al.*, 2018). Reduced diversity in modern cultivars is a consequence of more recent human cultivation practices (Trucci 2021), and the positive selection of several alleles that resulted in limited population size (Gaut *et al.*, 2018).

Low genetic diversity has led agriculture to become more vulnerable to climate change, endangering food security. Genetic diversity is the range of genetic traits within the species which gives rise to genetic variation among individuals, vital for improving crop resilience to biotic and abiotic stresses (Swarup *et al.*, 2020). Wild relatives including progenitors of domesticated crops are a useful source of this, they have maintained high genetic diversity that aids their survival in various environmental conditions (Dempewolf *et al.*, 2017, Zhang *et al.*, 2017, Vincent *et al.*, 2019). Understanding the evolutionary processes that help crop progenitors adapt to their environment can support the development of new crop varieties that can

sustainably tolerate current and future environmental challenges. Identifying and incorporating useful genetic traits is the aim of many breeding programs focussing on crop improvements.

1.1.2 Selection between wild species in early domestication

The emergence of domestication was fraught with many environmental changes. Approximately 10-12,000 years ago, around the time when crops were first domesticated, there was an increase in atmospheric CO2 and temperature (Shakun *et al.*, 2012). These climatic changes would have affected plant growth and could induce genetic changes important for adaptation. The transition of humans from hunter-gatherers, foraging wild plants for their fruits and seeds, to early farmers, was also a significant change for wild plants (Purugganan and Fuller, 2009). Early stages of domestication involved transporting wild plants from their natural environment into human-modified environments. These would grow conspicuously in these environments, giving way to their intentional planting for human consumption or use. The maintenance of wild populations would have been influenced by foraging and the growth of multiple species together (Rowley-Conwy and Layton, 2011). These early anthropogenic environments are thought to be disturbed by events such as seasonal fires and flooding (Wood and Lenné, 2018). Domesticated landscapes such as those managed for human use could also impose selective pressures on plant species even in the absence of cultivation (Alam and Purugganan, 2024).

Natural selection was the main driving force in early domestication leading to a protracted process (Purugganan, 2019), contrary to previous studies suggesting it was a rapid process from strong artificial selection (Abbo *et al.*, 2011; Rindos, 2013). Archaeological records revealed that domestication phenotypes took longer to arise and be fixed in domesticated cereal crops (Tanno and Willcox, 2006; Purugganan and Fuller, 2011). Population genomics studies also support this idea revealing a gradual decline in effective population size, instead of a sudden population bottleneck expected with rapid domestication (Gaut *et al.*, 2018; Allaby *et al.*, 2022). The lack of evidence for genetic diversity loss associated with domestication bottleneck has been reported in rice (Cubry *et al.*, 2018), maize (Kistler *et al.*, 2018), barley (Mascher *et al.*, 2016), wheat (Scott *et al.*, 2019), sorghum (Smith *et al.*, 2019) and common beans (Trucchi *et al.*, 2021). Selection can act on standing genetic variation or on novel mutations. Many crop domestication traits may be linked to multiple genes of small effect, resulting in longer fixation time in a population (Stetter et al., 2017a). This suggests weak selection pressures acting for a prolonged period on large, multiple, interconnected populations, countering the effect of drift to generate a sufficient variation for selection to act on (Allaby *et al.*, 2019; Alam and Purugganan, 2024). This indicates that wild populations in domesticated landscapes are phenotypically

similar but genetically diverse, highlighting the importance of generating variation during domestication.

Evidence of incomplete domestication has been reported, with wild species related to maize, grain amaranth and squash being cultivated but not domesticated (Vallebueno-Estrada et al., 2016; Stetter et al., 2017b; Wallace et al., 2019; Clement et al., 2021). Competition between wild plants would favour those best suited to these domesticated landscapes and those able to develop a mutualistic relationship with humans. Those that are less well-adapted would tend to be excluded. This suggests that wild species may have different domestication potential. Several factors that hinder or promote domestication have been discussed in the literature. The most obvious advantages would be characteristics favourable to early farmers such as larger edible parts (e.g. fruits or seeds), favourable taste, and ease of cultivation. Domestication of annual crops has been favoured over perennials during domestication, this is attributed to their greater reproductive allocation resulting in high seed productivity (Van Tassel *et al.*, 2010). Milla (2023) also pointed out that resource acquisition and growth rates are important phenotypes influenced by response to domestication. Moreover, Salman-Minkov *et al.* (2016) proposed that polyploid species were more likely to be domesticated. Other genetic constraints have also been linked to incomplete domestication such as the genetic architecture of domestication traits, lack of standing genetic variation, and increased genetic load (Stetter, 2020).

1.1.3 Studying crop wild relatives

Domestication has been used as a model for exploring evolutionary processes of socio-economically important crops capitalising on the availability of genomic data (Gregory, 2009; Larson *et al.*, 2014). In recent decades, there has been a push to increase research and conservation of crop wild relatives, as they offer valuable resources for crop improvements (Brozynska *et al.*, 2016; Kapazoglou *et al.*, 2023; Mammadov *et al.*, 2018; Dempewolf *et al.*, 2014).

The question of why domestication was limited to so few species was first posed by Diamond (2002), who noted that often only one member of the closely related groups of species became domesticated, suggesting that the hindrance lay with the species itself, not the early farmers. This was supported by evidence of many wild plant species that were harvested but not domesticated (Wallace *et al.*, 2019), and the independent domestication that occurred worldwide (Purugganan and Fuller, 2009). Furthermore, the failure of modern breeding techniques to add to the short list of domesticates, also highlights that there are factors that prevent the domestication of many species (Diamond, 2002).

The term ‘domesticability’ was adopted to describe a species' ability to generate heritable and adaptive variation that can be selected during domestication (Romero et al., 2025). Investigating the mechanisms that allowed the initial domestication of crops is one of the core questions in domestication, however, major constraints hinder this research (Zeder, 2015). There are few or missing preserved plant remains from archaeological sites to enable the association of morphological data to plants available in early domestication. Genetic data is also limited by the availability of ancient DNA to enable associated genetic changes during domestication. Consequently, current research on the initial domestication of crops relies on modern domesticates and their living crop wild relatives. These extant crop wild relatives are not identical to the ‘true’ progenitor that gave rise to our modern crops, as these species also have evolved since the divergence of the crop. Here, the term ‘progenitor’ will be used as a proxy to describe the extant wild species that evolved from the ancestral progenitor.

The majority of studies that explore the important mechanisms in domestication only focus on the changes between domesticates and wild progenitors. Research on crop wild progenitors and other wild species that were not domesticated is few and often limited to phenotypic assessments. In grasses and legumes, crop progenitors have been shown to have larger seeds and faster germination compared to other wild relatives, but there are no significant differences in total seed yield and plant biomass (Cunniff *et al.*, 2014; Preece *et al.*, 2015; Preece *et al.*, 2018; Preece *et al.*, 2021). Another study on 27 crop species showed that progenitors had thicker, but less dense, roots compared to other wilds, suggesting their adaptation to fertile soil conditions, enabling greater nutrient acquisition (Martín-Robles *et al.*, 2018). These demonstrate the progenitor’s competitiveness and potential pre-adaptation in a cultivated environment. There is a need for more comparative analyses between progenitors and other wild species that were never domesticated (never-domesticated wild species) and focus on genomic differences between these species.

Uncovering traits linked to domesticability would suggest that domesticated species are a selected set of species and comparing progenitors and other wild species is a more appropriate model to study evolvability (Romero et al., 2025). Furthermore, understanding processes and aspects of plant genomes that promote domestication can facilitate future domestication of new crops and aid the efforts of crop improvements to tackle current and future climate challenges.

1.2 Plasticity

Plasticity is the ability of an organism to respond to different environments (Schneider, 2022). The change in environment triggers a plastic response, from changes at the genomic level to changes in physiology and morphology (Pfennig, 2021). Species vary in the plasticity in their traits and are affected by their environment in different ways (Palacio-López *et al.*, 2015; Schneider and Lynch, 2020). Plasticity can generate genetic and phenotypic variation, important for selection in early domestication. Differences in plasticity between crop wild relatives may have given certain species an advantage over others, allowing their domestication.

1.2.1 History of plasticity research

The history of plasticity and its importance in evolution was fraught with controversies (Sommer, 2020). The ”Baldwin effect” was the first example of plasticity in 1898 by James Baldwin; this describes how learned behaviour, such as response to a new stressor might affect fitness and consequently selection (Baldwin, 1896). In 1909, Richard Woltereck introduced the term “reaction norm” to describe the effects of different environments on phenotypic expression in *Daphnia* (Woltereck, 1909). It was only in 1911 that Wilhelm Johannsen recognised the difference between genotype and phenotype (Johannsen, 1911), therefore introducing the concept of genotype-environmental interaction.

A few decades later, Schmalhausen and Waddington introduced the theory of “stabilising selection” (Schmalhausen, 1949) and genetic assimilation (Waddington, 1977), both arguing that adaptive plastic traits can be fixed in the population. To distinguish environmentally induced phenotypic variation from genetic polymorphism, which is genetically determined phenotypic variation, Ernst Mayr coined the term “polyphenism” in 1963 (Mayr, 1963). Anthony Bradshaw observed alternative phenotypes in plants as a response to extreme environmental conditions, realising that close relatives from the same genus can have differing plastic responses, inferring a genetic basis to differences in phenotypes (Bradshaw, 1965). At this point, critiques were still arguing that plasticity may hinder evolution and not promote it (Williams, 1966; Charlesworth *et al.*, 1982).

There was still a lack of empirical evidence for the role of plasticity in evolution. The first compelling claim came in 1989 by Mary Jane West-Eberhard (West-Eberhard, 1989). She offered an extensive collection of alternative phenotypes and plasticity in nature, that can be selected and lead to novel traits and adaptation (West-Eberhard, 1989; West-Eberhard, 2003). The

eventual identification of molecular mechanisms of phenotypic response led the way for full acceptance of plasticity as an agent of evolutionary change (Uller *et al.*, 2018).

1.2.2 The role of plasticity in adaptive evolution

The role of plasticity in facilitating adaptive evolution has long been proposed (Pfennig *et al.*, 2010; Gilbert *et al.*, 2015). Even though the role of plasticity in evolutionary biology is still controversial, many have argued that it could facilitate adaptation (Laland *et al.*, 2014; Pfennig, 2021). This has been referred to as the 'plasticity-first hypothesis' where plasticity generates variants under stressful conditions with enhanced fitness (West-Eberhard, 2003; Schwander and Leimar, 2011; Levis and Pfennig, 2016).

Pfennig (2021) reviewed the different ways plasticity could facilitate evolution. First was the 'buying time', which describes how plasticity can allow time for genetic adaptive evolution to arise, enabling the population's persistence in novel environments (Fox *et al.*, 2019). Standing genetic variation in novel and changing environments could facilitate adaptive plasticity, buying time for the population to evolve through new genetic mutations or speeding up the process as plasticity itself can evolve (Pfennig, 2021). Secondly, the 'plasticity-led evolution' is where environmental changes trigger adaptive phenotypic evolution that simultaneously alters selection patterns, creating new trait variants upon which selection can act. Plasticity could also expose 'cryptic' genetic variation for selection; standing genetic variations are usually hidden from selection but environmentally induced genetic changes (i.e. gene expression changes) can facilitate phenotypic plasticity (Levis and Pfennig, 2016). Thirdly, 'non-genetic inheritance' describes how plasticity can aid evolution through non-genetic inheritance systems such as epigenetic changes (Bonduriansky, 2012). Gene regulation during developmental plasticity was a driving force of evolutionary diversification and novel traits (Piperno, 2011). Beneficial phenotypes could be selected and be fixed through genetic assimilation; over multiple generations, if this new environmental condition persists; traits induced by changes in the environment increase in fitness and become canalised, losing their plasticity (Wood *et al.*, 2023). This has been proposed for several crops (Piperno *et al.*, 2019; Diggle and Miller, 2013; Belcher *et al.*, 2023). These concepts significantly overlap with each other, such as the 'buying-time' and 'plasticity-led' mechanisms resulting in beneficial phenotypes allowing the evolution of plasticity, resulting in their adaptation and fixation in the population by 'non-genetic inheritance'.

Levis and Pfennig (2016) argued that plasticity may have greater evolutionary potential than previously thought, compared to phenotypic change induced by mutations. Environmentally induced phenotypic change affects multiple individuals simultaneously with newly induced traits generated in diverse genotypes, increasing the chance of genetic accommodation. Also, adaptively directed plasticity allows for the immediate selection of new traits and faster refinement of their expression, further supported by the fact that plasticity can also uncover cryptic genetic variation (Levis and Pfennig, 2016).

1.2.3 Plasticity in domestication

Early domestication resulted in changes in the availability of resources such as nutrients, water and light; these environmental changes can induce plasticity (Milla *et al.*, 2014; Jones *et al.*, 2021). Plant roots and leaf morphology and physiology are sensitive to local growing conditions (Sultan *et al.* 2015), especially root traits which are known to be highly plastic (Schneider and Lynch 2020). A study on water and nutrient plasticity in seven different crops and progenitors found that under high water and nutrient conditions, plants produced larger and thinner leaves, with an increase in leaf area and stomatal conductance and there was no reduction in plasticity as a result of domestication. (Matesanz and Milla, 2018). This may suggest pre-adaptation to fertile and well-watered soils expected under cultivation. Plastic response to novel environments has been linked to local adaptation (Noble *et al.*, 2019; Johansson *et al.*, 2020; Radersma *et al.*, 2020), indicative of plastic traits as an important driver in the selection of progenitors over never-domesticated wild relatives under cultivation.

Competition between and among wild species in human-modified fields can also trigger plasticity. The crop progenitor of erect knotweed grown, instead of in its natural habitat, in a garden, with reduced interspecific competition, triggered a shift in plant architecture, from a small herb with minimal branching to a small shrub with more branches and seeds (Mueller *et al.* 2017). This was a response to light availability, where a small change in cultivation practice by early farmers induced an increase in yield within a single growing season. Other factors, such as changes in herbivory could also promote plasticity in phenotypes relating to plant defence (Wang *et al.*, 2023a).

Changes in the environment in early domestication would have triggered a plastic response in wild plants. This can result in multiple phenotypes from a single genotype, allowing the cultivation of crop progenitors in different environmental conditions. For example, an increase in CO_2 can increase yield in most species (Cunniff *et al.*, 2017; Sage, 1995), but this varies significantly in different genotypes (Dingkuhn *et al.*, 2020; Bishop *et al.*, 2015). Several studies

have found that crop progenitors were more plastic than their domesticated counterparts (Matesanz and Milla, 2018; Piperno *et al.*, 2019). Adaptive plasticity in early domestication could have enabled progenitors to quickly adapt to human-modified fields. Changes in growing conditions have exposed cryptic variations in the morphology of crop progenitors (Piperno *et al.*, 2015; Mueller, 2017; Piperno *et al.*, 2019).

Studies on the progenitor of maize, teosinte, indicate that changes in climate in Early Holocene (during early domestication) induced maize-like phenotypes (Piperno *et al.*, 2015; Piperno *et al.*, 2019). The beneficial phenotypic traits in teosinte revealed by Piperno *et al.* (2019) were coupled with high gene expression plasticity compared to maize. Wild plants with adaptive variation that allowed their success in anthropogenic environments, such as greater stress resistance or yield, could be selected upon in early domestication allowing their domestication over other species. However, there is a lack of plasticity data between crop progenitors and never-domesticated wild relatives.

1.2.3.1 Studying plasticity

A plastic response is measured through reaction norms; this describes the pattern of phenotypic expression by a single genotype across a range of different environments (Pigliucci *et al.*, 1998). Researchers can manipulate environmental signals to assess the response of interest in a plasticity experiment. Plastic responses can be discrete or continuous, with the former resulting in alternative phenotypes (Sommer, 2020). In plants, these can range from changes in traits, life-history outcomes, biochemical and impacts on gene expression. An interdisciplinary approach that combines multiple metrics of plasticity can provide extensive insight into changes in plastic response. A plastic response can be considered adaptive if it's statistically associated with fitness, but this can be biased as both trait and fitness are affected by environmental conditions (Stinchcombe *et al.*, 2002). Alternatively, an adaptive response can be inferred based on ecological or functional outcomes, for example, thermoregulatory morphological phenotypes that minimise heat/cold stress or longer root systems in plants growing in nutrient-rich soils (Sultan, 2015). The underlying mechanisms of plastic response can also be explored such as changes in DNA methylation (Song *et al.*, 2012; Dar *et al.*, 2022) and gene expression (Rivera *et al.*, 2021).

Challenges can arise when investigating the role of plasticity in evolution; it can operate at different time scales and levels of biological organisation and be revealed through evolutionary time frames by assessing population changes across generations (Wund, 2012). Several study systems can be used to overcome this problem: (i) experimental model system, (ii) natural

populations under environmental change, and (iii) natural systems with existing ancestral and derived populations. The latter is commonly used in assessing changes during domestication by using modern domesticated crops and their extant wild progenitors. However, as previously discussed, living progenitor species are not identical to the ancient progenitor, as they too have evolved. In investigating processes such as plasticity, extant crop wild relatives are a valuable resource for indirect evidence to make inferences on the differences between these species in early domestication.

Plasticity is important in cultivation and has been characterised for several progenitors, that result in better crop architecture and reduce germination inhibitors (Piperno *et al.*, 2015; Mueller *et al.*, 2023). Environment plays two key roles in evolution, first affecting how genetic variation translates into phenotypic variation and second, its influence on the selection among these phenotypes (Wund, 2012). Therefore, it is vital to assess both genetic and phenotypic plasticity in progenitors and never-domesticated wilds when performing a comparative analysis of the role of plasticity in early domestication. If plasticity played a role in domestication, we expect (i) plasticity to be greater in the progenitor compared to the never-domesticated species; (ii) phenotypes and/or gene expression that are plastic are more likely to be divergent between the progenitor and the domesticated species; (iii) reduction in plasticity in the domesticates compared to the progenitor may be an indication of genetic assimilation.

1.3 Mutations

A main source of genetic diversity or genetic polymorphism is mutations, these are changes in the DNA sequence that can alter proteins and subsequently their functions. This can occur at different levels and have varying effects on the genome. There are many different types of mutation: single nucleotide polymorphisms (SNPs), indels (insertions and deletions), transposable elements (TEs; discussed below), chromosomal mutation (fragments of chromosomes rearranged), and copy number variation (CNV) (sections of chromosomes duplicated or lost). This generates genetic variability, whether it is new or standing genetic variation, that can promote adaptive or favourable phenotypes that early farmers consciously and/or unconsciously selected. This can generate genetic variation, which can be important for selection under changing growing conditions in early domestication.

1.3.1 Transposable elements

Barbara McClintock challenged the early concepts of genes with her discovery of mobile elements that can change their position within a genome. This can result in changes in gene expression (Ravindran, 2012). These mobile elements are also known as transposons or transposable elements (TEs) and McClintock's findings were first summarised in "The Origin and behaviour of mutable loci in maize" (McClintock, 1950). This comprised the discovery of a TE named "Ds" or dissociation locus that could change location within the chromosome with the presence of "Ac" or activator, a second locus, required for transposition of both loci. It was later revealed that Ac encoded a transposase enzyme, whilst Ds lacked the protein that encoded this protein (Fedoroff *et al.*, 1983). The insertion of these elements could result in reversible mutations affecting gene function, illustrating the role of TEs in switching genes on and off. Another discovery by McClintock was the TE Suppressor-Mutator (Spm), which could alternate between an inactive and active form. This complex regulation is now known to be due to methylation (Cui and Fedoroff, 2002). It took the discovery of TEs in other eukaryotes for McClintocks' discovery to be widely recognised and in 1983 McClintock was awarded the Nobel Prize in Physiology or Medicine (Ravindran, 2012). Now, TEs are recognised as a major portion of the DNA in many species, from 44.7% in the human genome to 85% in the maize genome (International Human Genome Sequencing Consortium, 2004; Vicient, 2010).

1.3.1.1 Classification of TEs

Wicker *et al.*, (2007) introduced the first universal classification system that took into account structural characteristics and mode of transposition (Table 1.1). It proposed a naming system with a three-letter code denoting class, order and superfamily, enabling comparative studies on TEs from various species.

There are two classes of TEs that differ in their proliferation within the genome: Class I (retrotransposons) utilises the RNA transcription via a 'copy and paste' mechanism; and Class II (DNA transposons) utilises DNA replication via 'cut and paste' mechanism (Feschotte, 2008; Bourque *et al.*, 2018). TE order is the second hierarchal grouping, illustrating major differences in insertion mechanism, overall organisation and enzymology (Wicker *et al.*, 2007). Next, superfamilies within an order share replication strategy, differentiated by features such as protein structure, non-coding domains, and target site duplication (TSD). TSD is a short repeat sequence produced upon TE insertion.

Superfamilies are divided into families based on the conservation of DNA sequence in their coding region, internal domain or terminal repeat region. High sequence similarity was specified as the "80-80-80" rule, where TEs belonging to the same family share ≥ 80% of their sequence identity, ≥ 80% of their coding or internal domain, with an alignment of ≥ 80 bp long (Wicker *et al.*, 2007). Depending on the presence of the coding sequence for transposase, that allows transposition, TEs can also be grouped into autonomous (presence) and non-autonomous (absence) (Kazazian, 2004). Non-autonomous TEs require the presence of an autonomous TE for its activation, however, these autonomous partners do not have to be from the same family, therefore cross-activation of TEs is possible (Wicker *et al.*, 2007). This hierarchy demonstrates the abundance and diversity of TEs and with the rise of novel TEs emerging, a robust identification system is vital in identification and annotation of TEs.

Table 1.1: Transposable element (TE) classification (Wicker *et al.*, 2007).

Class	mode of transposition	transposition intermediate	Order	transposition	Superfamily
I (Retrotransposons)	RNA transcription via 'copy and paste' mechanism	RNA	LTR	autonomous	Copia, Gypsy, Bel-Pao, Retrovirus, ERV
			LINE	autonomous	R2, RTE, Jockey, L1, I
			SINE	non-autonomous	tRNA, 7SL, 5S
			DIR	autonomous	DIRS, Ngaro, VIPER
II (DNA transposons)	DNA replication via 'cut and paste' mechanism	DNA	TIR	autonomous	CACTA, hAT, Merlin, Mutator, PIF-Harbinger, P-element, PiggyBac, Tc1-Mariner, Transib
			MITEs	non-autonomous	Stowaway, Tourist, Emigrant
			Helitron	autonomous/ non-autonomous	Helitron
			Crypton	autonomous/ non-autonomous	Crypton
			Maverick	autonomous/ non-autonomous	Maverick

1.3.1.2 TE detection tools

The advancements in sequencing technologies in the past decades have led to an increase in the availability of genome sequences and tools to study the dynamics of transposable elements in many species (Ramakrishnan *et al.*, 2022). However, despite the diversity, abundance and significance of TEs in many species, they remain poorly annotated and studied only in model organisms (Ou *et al.*, 2019). Due to their repetitive nature and high copy number, TE annotation poses several analytical challenges and often requires the use of specialised tools. TEs could also have complex nesting structures, where a new TE inserts into an existing TE sequence.

With the availability of whole genome sequencing (WGS) data, many tools have been developed to detect TE insertions from Illumina paired-end short reads (Ramakrishnan *et al.*, 2022). There are two main strategies for TE discovery and annotation: annotation with databases and de novo annotation (Table 1.2). These tools can differ greatly in their TE annotation strategies, with previous reviews proposing the application of multiple tools complementarily to obtain exhaustive annotations (Lerat, 2010; Goerner-Potvin *et al.*, 2018).

Therefore, the development of a pipeline such as an Extensive de-novo TE Annotator (EDTA) that produces a comprehensive TE library for TE annotation using structural and homologous approaches is valuable (Ou *et al.*, 2019). Ou *et al.* (2019) benchmarked structural and homology annotators including general repeat annotators and LTR, non-LTR, TIR and Helitron annotators assessing sensitivity, specificity, accuracy, precision and false discovery rate. The EDTA pipeline integrated the best-performing tools including the following: LTR annotation tools, LTR_FINDER (Xu and Wang, 2007), LTRharvest (Ellinghaus *et al.*, 2008) and LTR_retriever (Ou and Jiang, 2018); the TIR and MITE annotation tools, Generic Repeat Finder (Shi and Liang, 2019) and TIR-Learner (Su *et al.*, 2019) with TIR candidates that are less than 600 bp classified as MITEs; the Helitron annotation tool, HelitronScanner (Xiong *et al.*, 2014). This pipeline also included filtering scripts and RepeatModeler (Flynn *et al.*, 2020) to identify non-LTRs and any unclassified TEs. EDTA annotates TEs in a reference genome producing a non-redundant TE library for annotation of structurally intact and fragmented TEs.

Table 1.2: Main tools used for Transposable element (TE) analysis.

Categories	Tools	Function	Reference
Databases	Repbase	database of consensus sequences for eukaryotic genomes	Bao *et al.* (2015)
	Dfam	open database of DNA repeat families	Wheeler *et al.* (2012)
	RepetDB	TE database for plants and fungi	Amselem *et al.* (2019)
	PlanTE-MIR DB	Plant specific TE database	Lorenzetti *et al.* (2016)
	PlaNC-TE	Plant specific TE database	Pedro *et al.* (2018)
Broad spectrum annotators	RepeatMasker	identify, classify and mask repetitive elements in a genome	Smit *et al.* (2015)
	RepeatModeler	de novo TE annotation in genomes	Flynn *et al.* (2020)
	dnaPipeTE	de novo TE annotation in raw reads	Goubert *et al.* (2015)
	EDTA	combines multiple tools to identifies and annotates TEs in genomes	Ou *et al.* (2019)
TE-specific annotators	LTR annotator	identification and annotation of LTRs	You *et al.* (2015)
	LTR_FINDER	predicts structure naad location of full-length LTRs	Xu and Wang (2007)
	LTRharvest	identification of LTRs	Ellinghaus *et al.* (2008)
	LTR_retriever	identification of LTRs	Ou and Jiang (2018)
	TIR-Learner	homology and structure based identification of TIRs	Su *et al.* (2019)
	Generic Repeat Finder	identifies TEs including MITEs	Shi and Liang (2019)
	HelitronScanner	identifies Helitrons	Xiong *et al.* (2014)
Transposon insertion polymorphisms (TIP) detection	Jitterbug	identifies non-referenced TEs in sequenced samples	Hénaff *et al.* (2015)
	PopoolationTE2	identifies reference and non-referenced TEs in sequenced samples	Kofler *et al.* (2016)
	Teflon	identifies reference and non-referenced TEs in sequenced samples	Adrion *et al.* (2017)
	T-lex3	identifies reference and non-referenced TEs in sequenced samples	Bogaerts-Márquez *et al.* (2020)
	McClintock	combines multiple tools to identify reference and non-reference TEs	Chen *et al.* (2023)

With the availability of reference genomes and the annotation of TE sequences in genomes, TE insertion polymorphisms (TIPs) can be explored. TIPs are the presence or absence of a TE at a

particular locus when compared with the reference genome, that can be studied in multiple individuals from the same species. Several tools specialise in TIP detection that can detect all TE types for short-read data, including T-lex3 (Bogaerts-Márquez *et al.*, 2020), PopoolationTE2 (Kofler *et al.*, 2016), and Jitterbug (Hénaff *et al.*, 2015).

TE detection can be addressed by one or a combination of the following approaches: (i) using split-read (SR), where one segment of a read maps to the reference and the other maps to a TE sequence; (ii) employing discordant read pairs (DRP) with paired-end reads, where one read maps onto the reference genome and the other maps onto a TE sequence; (iii) identification of TE-specific motifs such as TSDs (Goerner-Potvin *et al.*, 2018). A benchmark study on 12 widely used TE detection tools found that PopoolationTE2 was the overall best performer (Vendrell-Mir *et al.*, 2019). It was the best-performing broad-spectrum tool, with the highest sensitivity in LTR detection; performing as the best tool at low coverage and showing the best balance between sensitivity and precision at high coverage. Additionally, it also had ease of use, fast run time and using real data, PopoolationTE2 was the best at detecting both heterozygous and homozygous insertions in *Drosophila* and human datasets (Vendrell-Mir *et al.*, 2019).

1.3.2 Mutations in domestication

There are many domestication or diversification traits are a result of new mutations (Meyer and Purugganan, 2013). During domestication, these mutations were selected over time (Wright *et al.*, 2005). Independent domestication of different crops across the world suggests that progenitors experienced similar selection pressures leading to convergent evolution and resulting in domestication syndrome (Smýkal *et al.*, 2018). Evidence for convergence evolution in many crops was the result of loss of function mutations in genes important in domestication: reduced seed shattering was due to a mutation in the YABBY transcription factor Shattering1 (Sh1) orthologues in rice, maize, wheat and sorghum (Lin *et al.*, 2012; Katkout *et al.*, 2015); improved plant architecture were due to a mutation in the Terminal Flower1 (TFL1) orthologues in the common bean and pigeon pea (Repinski *et al.*, 2012; Mir *et al.*, 2014). Mutations in different genes that occupy analogous networks have also resulted in identical phenotypes, such as the repression of flowering with FRIGIDA and FLC were downregulated by vernalisation (Alonso-Blanco *et al.*, 2009). These highlight the importance of mutation to the emergence of beneficial phenotypes in domestication.

SNPs are the simplest and most frequent form of DNA polymorphism in plant genomes. They can be found at high densities in tomato with an average of 6.1 SNP/kb (Kim *et al.*, 2014). These are common DNA markers and our empirical understanding of SNP's contribution to trait

variation and selection is extensive (McNally *et al.*, 2009; Blanca *et al.*, 2012; Huq *et al.*, 2016; Contreras-Soto *et al.*, 2017). However, there is still a lack of understanding of how different types of mutations contribute to the creation and maintenance of variation.

A large proportion of plant genomes is composed of many repetitive sequences, most of which are Transposable Elements (TEs); with LTRs in class I and MITEs in class II as the most prevalent in plant genomes (Cho, 2021). The proportion of TEs within crop genomes varies, not just in totality but also in the proportion of different types of TEs (Vitte *et al.*, 2014). In maize, where TEs were first discovered, 85% of its genome is attributed to repetitive sequences with 63% recognised structurally as TEs (Schnable *et al.*, 2009; Jiao *et al.*, 2017); with wheat and rice it is at least 85% and 40% respectively (International Rice Genome Sequencing Project, 2005; Wicker *et al.*, 2018). TEs are a significant component of crop genomes, not just because of their abundance but also their effect on genes. The impact of TEs can differ depending on their site of integration; transposition into up/downstream regions, introns and exons can cause proximal effects such as knock-outs, changes in gene expression and activation of flanking genes (Gill *et al.*, 2021). Some of the TE insertions reported in crops and their impact are listed in Table 1.3. These studies highlight the variety of TE insertion sites and their consequences on gene regulation and expression, with TEs found near or in genic regions having the largest impact on gene function.

Table 1.3: Effect of transposable elements (TEs) in crops.

Crop	TEs	TE location	Type of impact	Traits	Reference
barley	*Sukkula*-like retrotransposon	upstream of *HvHMA3* gene	upregulation	low cadium barley variety	Lei *et al.* (2020)
foxtail millet	various TEs	intron and exon of *GBSS1* gene	disruption or loss of function	waxy traits	Kawase *et al.* (2005)
foxtail millet	*Harbinger*	exon of *sh1* gene	loss of function	loss of seed shattering	Liu *et al.* (2022a)
maize	*Hopscotch*	regulatory region of *tb1*	upregulation	increased apical dominance repressing branching	Studer *et al.* (2011)
maize	various TEs	exon of *waxy* gene	loss of function	waxy trait	Xiaoyang *et al.* (2017)

maize	*Ac/Ds*	*Bronze* gene	loss of function	loss of purple pigmentation	McClintock (1953)
morning glory	Tip100 (*Ac/Ds* family)	intron of *CHS-D* gene	loss of function	flower variegation	Habu *et al.* (1998)
morning glory	*En/Spm*-related TE	intron of *DUPLICATED* gene	loss of function	double flower phenotype	Nitasaka (2003)
rapeseed	MITE	upstream region of *BnFLC.A10*	upregulation	vernalization requirement	Hou *et al.* (2012)
rapeseed	Copia	upstream of *SHATTERPROOF 1* gene	downregulation	resistance to pod shattering	Liu *et al.* (2020b)
rice	Dasheng	exon of *Wx* gene	loss of function	glutinous trait	Hori *et al.* (2007)
rice	LTR	upstream of the *Pit* gene	upregulation	fungal pathogen resistance	Hayashi and Yoshida (2009)
tomato	Rider	intron of the Solyc12g038510 gene	nonsense mutation	jointless fruit stem phenotype	Soyk *et al.* (2017)
tomato	Rider	Intron of *sun* locus	gene duplication	elongated fruit shape	Xiao *et al.* (2008)
tomato	Rider	*PSY1* gene	loss of function	yellow flesh fruits	Jiang *et al.* (2012)

1.3.2.1 Population dynamics of mutations

Most mutations are neutral or detrimental, with beneficial mutations expected to be very rare (Eyre-Walker and Keightley, 2007). The most apparent consequence of a mutation is its effect on genes causing its loss of function and/or changes in gene expression. This also applies to TEs as they are generally thought to be deleterious (Lee, 2022). Integration of TEs into a genome can be harmful to the hosts' genome indirectly in several ways: (i) disrupting the coding sequence leading to loss of function or indirectly by disruption in promoter or enhancer regions, changing gene expression (Hirsch and Springer, 2017); (ii) as a way of TE repression, DNA methylation suppresses TE expression but it can also result in the downregulation of nearby genes (Hollister and Gaut, 2009); (iii) TEs can introduce new regulatory elements upon their insertion (Chuong *et al.*, 2017); (iv) TEs can induce ectopic recombination which can lead to the duplication and deletion of genomic regions (Almojil *et al.*, 2021). However, these genomic disruptions have also given rise to advantages, because TEs generate significant genomic changes, that can drive the

development of phenotypic diversity which can enable populations to adapt to changing environments (Oliver and Greene, 2012; Ramakrishnan *et al.*, 2021).

The fate of mutations will depend on a number of factors but its effect on the host's genome greatly influences this. Neutral mutations can persist in the population by change and change frequency over time due to genetic drift (Kimura, 1985). For mutations that affect fitness, the mutation-selection balance model describes the equilibrium reached in the population between the rate of new mutations and their removal from the population through selection (Crow, 2017). The most effective selection is usually for strongly deleterious mutations (Keightley and Eyre-Walker, 2010). These negatively impact the host's fitness and are selected against resulting in their removal from the population over time. However, mutations can also increase in frequency or be fixed in a population through genetic drift, selection and/or genetic hitchhiking (Loewe and Hill, 2010). This is especially true for beneficial mutations, where the probability of fixations increases. Changes in mutation rates or mutation-selection balance can affect the genetic variation in a population.

The contribution of mutations, such as TEs, on adaptation for evolution in general and during early domestication is largely unknown. Populations can adapt to novel growing conditions (i.e. cultivated fields) through selection of standing genetic variation or the selection of new mutations. Greater standing genetic variation in a population can facilitate adaptation because of the greater capacity to adapt to environmental changes and the ability to mitigate the effects of harmful mutations (Barrett and Schluter, 2008). This results in greater potential for individuals to have beneficial traits suited for the new environment that allows for faster adaptation.

Furthermore, evidence of increased TE activity in stress conditions has been documented for several TE families (Kimura *et al.*, 2001; Salazar *et al.*, 2007; Woodrow *et al.*, 2011; Ito *et al.*, 2013; Roquis *et al.*, 2021). Activation of TEs can trigger random genetic variation, vital for adaptation through natural selection (Schrader and Schmitz, 2019). Therefore, an increase in TE activity could mean more novel phenotypes arise in certain species resulting in promoting their domestication through the accumulation of beneficial mutations at the expense of others.

There are a limited number of mutations that can be under selection at any one point. Estimates suggest that no more than 100 genes can be under positive selection during domestication simultaneously (Allaby *et al.*, 2015). In maize, 7.6% of the genome showed signs of selection during domestication (Hufford *et al.*, 2012). The benefits of a higher mutation rate to the speed of adaptation have been extensively studied in micro-organisms (Desai and Fisher, 2007; Desai *et al.*, 2007; Wiser *et al.*, 2013). A more likely consequence of a faster mutation rate is the faster production of deleterious mutations. However, with the reduced stress conditions in cultivation (i.e. availability of nutrients and reduced competition), purifying selection would be more

relaxed, allowing the persistence of populations even with increased mutation load as reported in several domesticated species (Smith *et al.*, 2019; Renaut and Rieseberg, 2015; Wang *et al.*, 2021a). Furthermore, deleterious mutations that reduce fitness under one condition (e.g. natural conditions) can become advantageous in a different environment (e.g. under domestication) (Dwivedi *et al.*, 2023), such as the *sh1* loss of function mutation was vital in the domestication of maize, rice and sorghum (Lin *et al.*, 2012). So it might be the case that some wild species in early domestication had a large reservoir of genetic variation, a high rate of transpiration or a high rate of mutation induced by domestication.

1.3.2.2 Studying mutations

To understand the distribution of mutational effects, experimental and population genetics approaches are applied (Loewe and Hill, 2010). In recent decades, the distribution of mutational effects has been inferred through genomic data of populations (Chen *et al.*, 2022). Mutations affect DNA diversity through the introduction of new genetic variants. Nucleotide diversity (π) can be estimated by annotating SNPs using the formula by Nei and Li (1979) or using one of the bioinformatics tools such as VCFtools (Danecek *et al.*, 2011). Other mutations, such as TEs, can be detected and annotated using the tools discussed above (Table 1.2). Furthermore, transposon insertion polymorphisms (TIPs) can also be employed to detect TE diversity and activity. This scores the presence and/or absence of a TE in a particular locus. TIPs have been characterised in a number of crops such as rice (Castanera *et al.*, 2023), sorghum (Tao *et al.*, 2021), eggplant (Gramazio *et al.*, 2019) and carrots (Macko-Podgórni *et al.*, 2019). Comparative genomics of TIPs has been performed between domesticates and their progenitors (Dominguez *et al.*, 2020; Liu *et al.*, 2020c; Castanera *et al.*, 2021; Gao and Fox-Fogle, 2023). These highlight changes in TE composition between progenitors and domesticates, for example, greater diversity of TE families in tomato progenitors compared to domesticated varieties (Dominguez *et al.*, 2020) and changes in TIP frequency between domesticated rice and its progenitor (Castanera *et al.*, 2021). However, we are lacking empirical data on the differences between crop progenitors and other related wild species. Extending the identification of polymorphisms to TEs will greatly expand our understanding of their contribution to genetic diversity and their role in early domestication.

In eukaryotes, nucleotide diversity is found to be positively correlated with mutation rates (Wang and Obbard, 2023). Wang and Obbard (2023) performed a meta-analysis of mutation rates experiments in eukaryotes covering 134 species, from unicellular eukaryotes to plants and mammals. They found higher mutation rates in species with longer generation times, larger genomes and smaller effective sizes. Mammals and plant species have higher mutation rates

than arthropods and unicellular eukaryotes (Wang and Obbard, 2023). Mutation rates of single nucleotide mutations (SNMs) are generally higher than the indel mutation rate, with short deletions more frequent than insertions (Wang and Obbard, 2023). Mutation rates of SNMs and indels have been characterised in several plant species such as *Arabidopsis* (Monroe *et al.*, 2022), maize (Yang *et al.*, 2017), rice (Ichikawa *et al.*, 2023) and duckweed (Sandler *et al.*, 2020). There is a large difference in mutation rates reported between duckweed (0.22×10^{-9}) and other plant species (19.8×10^{-9}), attributed to the lack of segregated germline in most plant species (Lanfear, 2018).

TE insertion rates are predominantly studied in unicellular organisms (Sousa *et al.*, 2013; Hénault *et al.*, 2020) or multicellular model organisms such as *Caenorhabditis elegans* (Bégin and Schoen, 2006) and *Drosophila melanogaster* (Adrion *et al.*, 2017; McCullers and Steiniger, 2017). TE insertion rates (per site per generation) can differ dramatically between different types of TEs, such as in *Escherichia coli* from 4.0×10^{-8} to 1.15×10^{-5} (Sousa *et al.*, 2013). TE insertion rates can also differ between populations, for example in *D. melanogaster*, European populations have higher transposition rates than West African with 23.36×10^{-5} and 8.99×10^{-5} copy^{-1} generation^{-1} respectively (Wang *et al.*, 2023b). The overall rate of transposition was 4.93×10^{-9}, higher than the SNM rate estimate of 3.30×10^{-9} in *D. melanogaster* (Wang *et al.*, 2023b). Activity of specific TE families and under varying conditions has been studied in Arabidopsis (Tsukahara *et al.*, 2009), rice (Nakazaki *et al.*, 2003), maize (Alleman and Freeling, 1986), and sunflower (Vukich *et al.*, 2009), however, whole-genome transposition rates in plants have yet to be reported.

For direct estimates of per-generation mutation rates, mutation accumulation (MA) experiments and parent-offspring (PO) sequencing are usually performed. In MA experiments, a single inbred or asexual genome accumulates spontaneous mutations over multiple generations, minimising the effectiveness of natural selection (Halligan and Keightley, 2009). On the other hand, PO sequencing utilises mutations arising over a single-generation (Yang *et al.*, 2015). This has several limitations with the number of mutations highly influenced by (i) sequencing and/or mapping errors, (ii) the genetic variation of heterozygotes, and (iii) pre-existing heterozygous sites in parental genomes, that can lead to false positive and false negative mutation calls (Wang and Obbard, 2023). Multiple generations in MA experiments combat these limitations, providing better long-term estimates of mutation rates. In plants, MA experiments have been performed on the model organism *Arabidopsis thaliana*, for estimates of mutation rates based on single nucleotide variants (SNVs) and indels (Ossowski *et al.*, 2010; Weng *et al.*, 2019; Lu *et al.*, 2021). Other SNV estimates have also been reported for Chinese cabbage (Park *et al.*, 2019) and two species of duckweed (Sandler *et al.*, 2020). No estimates of transposition rates from MA experiments have been reported in plants. Combined with whole-genome sequencing

(WGS), this experimental system can capture extensive mutation rates and distribution patterns. This can be employed to investigate the transposition rate in progenitors and never-domesticated species to assess mutation rates of different types of mutations among these species.

1.4 Study system: Tomato

Tomato (*Solanum lycopersicum* L.) is found in the Solanaceae family, this family consists of more than 300 species, including agronomically important crops like aubergine, pepper and potato (Knapp, 2002). This makes the tomato an ideal model system for fruit development as well as other fundamental processes such as response to biotic and abiotic stresses (Liu *et al.*, 2022b). Tomato is one of the most important crop species widely consumed, with world production totalling over 180 million tons in 2019 (FAO, 2021). They are easy to grow, with a relatively small genome size (950 Mb) and a vast availability of genomic resources (The Tomato Genome Consortium, 2012).

1.4.1 Domestication history

Figure 1.1

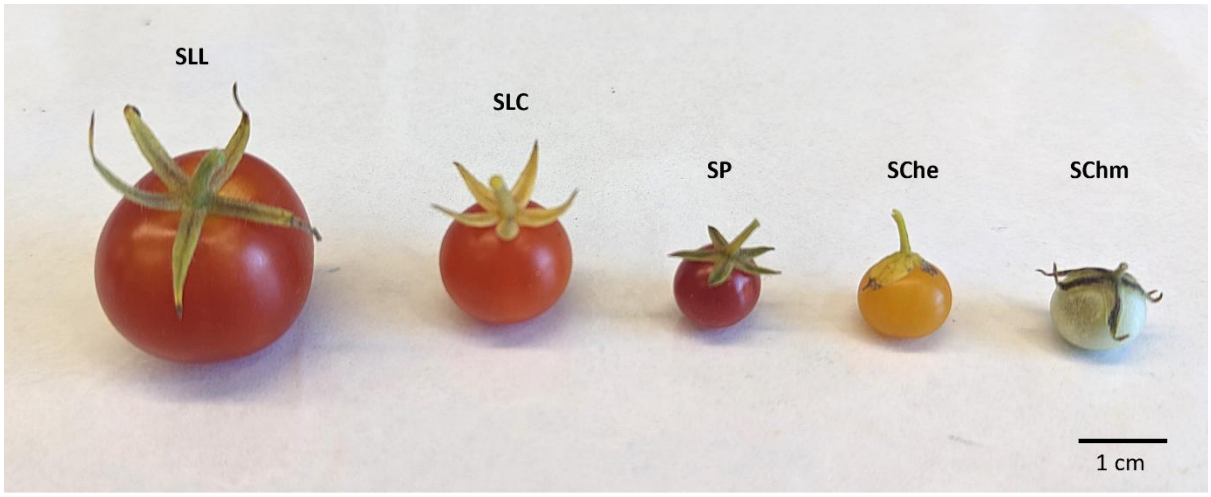


Figure 1.1: Images of different species of tomatoes.

This illustrates the size difference between the domesticated tomato *Solanum lycopersicum* (SLL; accession LA0395), the intermediate *S. lycopersicum* var. *cerasiforme* (SLC; LA1324), the progenitor *S. pimpinellifolium* (SP; LA1578) and other never-domesticated wild species, *S. cheesmaniae* (SChe; LA1039) and *S. chmielewskii* (SChm; LA2663).

Generally, the domestication history of tomato is described as a two-step process accompanied by an increase in fruit size (Figure 1.2), with the domestication of the progenitor *S. pimpinellifolium*, giving rise to *S. lycopersicum* var. *cerasiforme*, a semi-domesticated intermediate, before its subsequent improvement that brought about the domesticate *S. lycopersicum* var. *lycopersicum* (Lin et al., 2014; Blanca et al., 2015). Many details of this process are still contested, particularly regarding the role of *S. lycopersicum* var. *cerasiforme* in the process. Apart from being described as an intermediate, *S. lycopersicum* var. *cerasiforme* has also been described as an admixture following the hybridisation between *S. pimpinellifolium* and *S. lycopersicum* (Ranc *et al.*, 2008). A recent study by Razifard *et al.* (2020) suggests that many traits associated with cultivated tomatoes arose in South *American S. lycopersicum var. cerasiforme* (predating domestication) but were lost with its spread northwards as it was domesticated. They estimate *S. lycopersicum var. cerasiforme* diverging from *S. pimpinellifolium* around 78,000 years ago, with significant gene flow evident between these species (Razifard *et al.*, 2020). This challenges the linear process of tomato domestication with significant gene flow from wild relatives revealing complexities in tomato domestication.

Modern tomato cultivars are mostly hybrids, where agronomically important traits from multiple parents are combined through breeding programs (Bai and Lindhout, 2007). Artificial selection in cultivated tomatoes has resulted in reduced genetic diversity. Therefore, genetic diversity in wild tomatoes has been studied intensively. The introgression of cultivated tomatoes with wild relatives has resulted in increased abiotic tolerance and improved fruit quality and yield (Hobson and Grierson, 1993; Semel *et al.*, 2006; Ikeda *et al.*, 2013). This produced a wide range of phenotypic variation in modern cultivars, making crop wild relatives a great source of genetic diversity with great potential for tomato breeding.

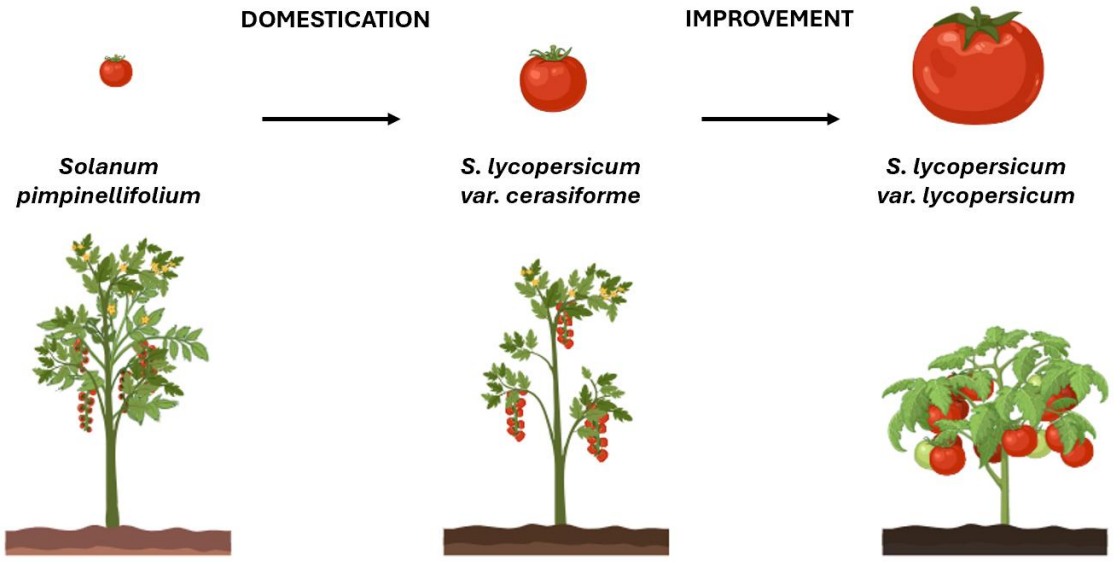


Figure 1.2: Illustration of tomato domestication history

Tomato domestication involved a two-step process accompanied by changes in plant morphology, specifically an increase in fruit size. The wild progenitor *S. pimpinellifolium* was domesticated to give rise to *S. lycopersicum var. cerasiforme*, which was improved to *S. lycopersicum var. lycopersicum*, the domesticated tomato.

1.4.2 The tomato genome

The first complete tomato genome (Heinz 1706) was sequenced in 2012, providing insight into its genome components and evolution (The Tomato Genome Consortium, 2012). Its genome size is roughly 950 Mb and assembled into 12 chromosomes, with 34,727 protein-coding genes. The first WGS of eight accessions of domesticated tomatoes compared against the reference genome identified a total of 4,290,679 unique SNPs (80,000 to 1.5 million SNPs per accession) and 128,000 indels (Causse *et al.*, 2013). Intergenic regions had significantly higher SNP frequency in the tomato genome than in genic regions (Aflitos *et al.*, 2014). For the progenitor, *S. pimpinellifolium*, WGS data of two accessions yielded 4,680,647 SNPs (Kim *et al.*, 2014). Recent admixture of domesticated and the progenitor, gene pools as a result of breeding led to segments of chromosomes 4,9,11 and 12, more closely related to the progenitor than the domesticated. The estimated divergence between domesticated tomato and its progenitor is 0.6% accounting for 5.4 million SNPs (The Tomato Genome Consortium, 2012).

Estimates suggest that TEs account for 63% of the genome, mostly comprising retrotransposon (The Tomato Genome Consortium, 2012). Mehra *et al.* (2015) found that 99% of *S. lycopersicum*

genes overlap with TEs in their genic regions or 5kb upstream region, illustrating the insertion of TE families within hAT, PIF-Harbinger and Helitron, SINE and LTR/ERV having preferentially within genic regions. Other studies have focused on specific TE superfamily/family and their distribution and influence on agronomic traits of tomato. Rider retrotransposon has been shown to contribute to plant architecture (Soyk *et al.*, 2017), fruit shape (Xiao *et al.*, 2008) and colour (Jiang *et al.*, 2012). Benoit *et al.* (2019) demonstrated that Rider is drought-inducible through MYB transcription factors involved in drought-stress response, highlighting its potential for crop breeding for drought tolerance.

1.4.3 Previous research on tomato

As an important agricultural crop, research efforts have focussed on understanding the phenotypic response of tomatoes under different stresses (Sato *et al.*, 2002; Zaki and Yokoi, 2016; Gerszberg and Hnatuszko-Konka, 2017; Lupo and Moshelion, 2024). Domestication research on tomatoes is extensive (Bai and Lindhout, 2007; Blanca et al., 2015), but only a few include other tomato wild relatives apart from the progenitor (Tohge *et al.*, 2020; Pereira *et al.*, 2021; Salazar-Mendoza *et al.*, 2023). Many studies have explored plasticity divergence between domesticated crops and their progenitors (Milla *et al.*, 2014; Matesanz and Milla, 2018; Piperno *et al.*, 2019), assessment of plasticity between crop progenitor and a never-domesticated wild relative are yet to be performed. In addition, comparative assessments of phenotypic and genetic plasticity are rarely done, with the focus usually on one or the other.

Many of the studies on wild tomato species highlight their diversity in morphology, metabolites and other fruit phenotypes (Iijima *et al.*, 2013; Schwahn *et al.*, 2014; Yeats *et al.*, 2012). Genetic diversity in tomatoes suggests that there is a reduction in nucleotide diversity during domestication (Lin *et al.*, 2014; Sauvage *et al.*, 2017; Razifard *et al.*, 2020; Fuentes *et al.*, 2022). Estimates of nucleotide diversity between tomato wild species may be influenced by the species' mating system, such that SI species have higher nucleotide diversity than SC species (Roselius *et al.*, 2005). Whole-genome sequence diversity based on SNPs suggests that the number of SNPs increases from the domesticates to the closely related *S. pimpinellifolium*, *S. cheesmaniae* and *S. galapagense* and then more sharply with wild species more distantly related (Aflitos *et al.*, 2014). Structural variants (SV) have also been explored in tomatoes, these are large insertions, deletions, duplications such as TEs, and chromosomal rearrangements (Alonge *et al.*, 2020). This found that *S. pimpinellifolium* had more SV diversity than *S. lycopersicum* and that most SVs identified were TEs and repeated sequences. A TIP analysis that included at least one species from each tomato species was performed by Dominguez *et*

al. (2020), this suggested that *S. pimpinellifolium* was the most genetically diverse group compared to the other wild species and the domesticated species. However, this fails to capture the short TE sequences such as MITEs, therefore it is important to apply TE identification tools that can identify all TE types. Most of the TE insertion identified were present at low frequency in the population, suggestive of recent transposition (Dominguez *et al.*, 2020). This might suggest that there is high TE activity in the tomato progenitor. There are currently no estimates of rates of mutation in tomatoes. Comparative estimates of mutation rates that include SNVs, indels and TEs in tomato progenitor and other never-domesticated wilds would give a broad insight into the role of mutation in early domestication.

1.5 Summary

Understanding the mechanisms at play in early domestication is important and will give us insights into the evolutionary processes acting on crop wild relatives. There is a lack of comparative research on crop progenitors and other wild species, with the majority focusing only on phenotypic assessments. Here, we aim to explore the role of plasticity, genetic diversity and mutation rates, particularly transposable elements (TEs), in the evolutionary advantage of the tomato progenitor, Solanum *pimpinellifolium*, that allowed its domestication and not the other wild tomato species. We follow the prevailing idea that *S. lycopersicum* and *S. lycopersicum* var *cerasiforme* are domesticated from the progenitor *S. pimpinellifolium*. This will provide more insight into the genetic diversity of crop wild relatives and hopefully encourage further comparative research into other mechanisms that may have been advantageous in early domestication.

1.6 Aims and objectives

This work aims to explore (i) the contribution of different mechanisms to the selective advantage of *S. pimpinellifolium* over never-domesticated wild tomato species, and (ii) how these mechanisms may have changed throughout domestication. These will be addressed through the following objectives:

1. The role of plasticity in tomato domestication (Chapter 2).
   Hypothesis: Tomato progenitor has more plastic phenotypic traits and genes than never-domesticated wilds.

To assess phenotypic and gene expression plasticity between the tomato progenitor (*S. pimpinellifolium*) and never-domesticated wild species (*S. cheesmaniae* and *S. chmielewskii*).

2. The role of transposable elements (TEs) on tomato domestication (Chapter 3).
   Hypothesis: Tomato progenitor has higher genetic diversity than never-domesticated wilds.
   To assess single nucleotide polymorphism (SNP) and transposon insertion polymorphism (TIP) diversity within the progenitor (*S. pimpinellifolium*) and never-domesticated species (*S. cheesmaniae* and *S. galapagense*) and compare these between species to determine whether the progenitor has higher genetic diversity than never-domesticated species.

3. The role of mutation rate on tomato domestication (Chapter 4).
   Hypothesis: Tomato progenitor has a higher mutation rate than never-domesticated wilds.
   To assess single nucleotide variant (SNV), indel and TE mutation rates between the progenitor (*S. pimpinellifolium*) and never-domesticated species (*S. cheesmaniae*).

Chapter 2 The role of plasticity in tomato domestication

2.1 Abstract

Plasticity is the ability of an organism to respond to new environments and can occur at different biological levels. Here we present the first phenotypic and genetic plasticity assessments of crop progenitor and never-domesticated relatives in order to explore whether plasticity can promote the domestication of some species over others. If plasticity promotes the domestication of only certain wild species, we hypothesise that (i) the progenitor has greater phenotypic and gene expression plasticity than the never-domesticated wild relatives, and (ii) traits (including gene expression) that have diverged between wild and domesticated species will be more likely to be plastic in the progenitor than those traits that have not diverged. Changes in phenotypic traits that are important in domestication and gene expression between the domesticated *Solanum lycopersicum* and *S. lycopersicum* var. *cerasiforme*, their progenitor *S. pimpinellifolium*, and the never-domesticated wild species *S. cheesmaniae* and *S. chmielewskii* (referred here as 'wilds') were assessed. Plants of each were grown in a glasshouse under three experimental treatments, broadly reflective of changes in growing conditions in early domestication.

We identified traits and genes that were plastic and/or divergent between species. The progenitor had earlier flowering, greater fruit yield and larger seeds than the wilds. During domestication, there was a reduction in height and an increase in fruit weight, yield, size, and seed size. More traits, including fruit weight and size, were plastic in the progenitor than in the wilds (11 vs 7 out of 16). RNA-sequencing of leaves and fruits from the domesticated (*S. lycopersicum)*, the progenitor (*S. pimpinellifolium*) and the wild (*S. cheesmaniae*) species was performed, followed by differential gene expression analysis. More genes were plastic in the progenitor compared to the domesticated and the wild, and these were enriched for a pathway involved in plant hormone signal transduction, related to important plant processes such as germination and stress response. We identified more plastic genes in the progenitor than in the other two species, supporting the first hypothesis. Genes that were divergent between the progenitor and the wild were more likely than expected by chance to be plastic in the progenitor, supporting the second hypothesis. This association of plastic and divergent genes was found in the domesticated species as well as the wilds, which may suggest that plasticity is

shared across species, but that the overall greater plasticity in the progenitor served as a 'fast track' to the evolution of novel phenotypes for selection to act on.

2.2 Introduction

Domestication is an evolutionary process that transforms the wild progenitor into cultivated crops reliant on humans for resources and reproduction (Zeder, 2015). Half of the ~390,000 plant species on earth are edible (Willis, 2017), but only a few hundred have been domesticated (Zeven and De Wet, 1982) with 15 providing 90% of our calories (FAO, 2017). This begs the question, why are so few species domesticated? Answering this question will reveal mechanisms contributing to crop progenitors' evolutionary advantage. Identification of adaptive variation in crop wild relatives is important in the adaptation to novel environments. These could be utilised in the crop improvement of existing crops as well as the development of novel crops that can sustainably tolerate current and future environmental challenges.

Many environmental changes occurred during crop domestication (Purugganan and Fuller, 2009; Larson *et al.*, 2014; Purugganan, 2022), which could have acted as a selection pressure on wild species during early domestication. However, early human settlers foraged wild plants in their natural environments and transported these to human-modified settlements. This transition in growing conditions could have involved changes in soil quality with an increase in nutrient and water inputs, and reduced competition for space and nutrients relative to the wild environment, as well as movement away from natural pests and herbivores (Bogaard *et al.*, 2013; Araus *et al.*, 2014). The growth of wild plants alongside human communities led to their subsequent cultivation of some species as a result of selection between plant individuals with favourable characteristics for human consumption or use (Purugganan, 2022). Human cultivation itself may have affected which species were domesticated, with desirable traits such as tastier fruits and/or ease of cultivation selected for in early domestication.

Plasticity is the ability of the organism to acclimate to different environments (Schneider, 2022). The change in environment triggers a plastic response, through changes in gene expression levels which can result in changes in morphology or physiology. Even though the role of plasticity in evolutionary biology is still controversial, many have argued that it could facilitate adaptation (Laland *et al.*, 2014; Pfennig, 2021). Plasticity can allow time for genetic adaptive evolution to arise, enabling the persistence of the population in novel environments (Fox *et al.*,

2019). Adaptive plasticity in early domestication could have enabled progenitors to quickly adapt to human-modified fields. Cunniff et al. (2014) revealed that cereal crop progenitors were more resilient to competition and disturbance compared to other related wild grasses. Similarly, Martín-Robles *et al.* (2018) highlighted thicker and more dense roots in progenitors of diverse crops compared to other wild relatives, suggestive of pre-adaptation to agricultural conditions. Plasticity could also expose 'cryptic' genetic variation for selection; standing genetic variations are usually hidden from selection but environmentally induced changes in gene expression facilitate phenotypic plasticity (Levis and Pfennig, 2016). A few garden experiments exploring changes in growing conditions in early domestication have exposed cryptic variations in the morphology of crop progenitors (Piperno *et al.*, 2015; Mueller, 2017; Piperno *et al.*, 2019). Gene regulation induced by plasticity is a driving force of novel traits and evolutionary diversification (Piperno, 2011), providing the genetic basis of phenotypic plasticity explored through gene expression studies (Piperno *et al.*, 2019; Campbell-Staton *et al.*, 2021; Kenkel and Matz, 2016). Under the persistence of a new environmental condition over multiple generations, beneficial phenotypes could be selected on and be fixed through genetic assimilation, i.e. the process by which traits induced by changes in the environment become canalised, losing their plasticity (Wood *et al.*, 2023), as has been proposed for several crops (Piperno *et al.*, 2019; Diggle and Miller, 2013; Belcher *et al.*, 2023).

Identification of adaptive variation can be utilised in crop improvement to better adapt to current and future climate challenges, aiding food security (Brooker *et al.*, 2022). Investigating these mechanisms which allowed the initial domestication of crops is one of the core questions in domestication. Current research on the initial domestication of crops relies on modern domesticates and their living crop wild relatives, but these are few and often limited to phenotypic assessments (Cunniff et al., 2014; Preece et al., 2015; Preece et al., 2021). To explore the role of plasticity in early domestication between crop wild relatives, we need studies that consider phenotypic and genetic plasticity in never-domesticated wild relatives parallel to the progenitor.

Tomato (*Solanum lycopersicum* L.) is found in the Solanaceae family, consisting of agronomically important crop species including aubergine, pepper and potato (Knapp, 2002). Tomato is widely consumed with world production reaching over 180 million tons in 2019 (FAO, 2021). The domesticated tomato species was estimated to have originated 7,000 years ago (Razifard *et al.*, 2020). 12 wild tomato species originate from the Andean region (Bergougnoux, 2014), that vary in their morphology from the red-fruited progenitor to orange and green-fruited

wild tomato species (Gonzali and Perata, 2021). The domestication history of tomato has been generally described as a two-step process accompanied by an increase in fruit size, with the domestication of the progenitor *S. pimpinellifolium*, giving rise to *S. lycopersicum* var. *cerasiforme*, an intermediate, before its subsequent improvement that brought about the domesticate *S. lycopersicum* var. *lycopersicum* (Lin et al., 2014; Blanca et al., 2015).

In this study, we investigated the role of plasticity in the selective advantage of the progenitor over never-domesticated wilds in early domestication. We quantify the number of plastic traits and genes in the domesticates, progenitor and never-domesticated wild species. Assaying gene expression plasticity is an objective analysis of thousands of traits. We hypothesise that there are a greater number of plastic traits and genes in the progenitor compared to never-domesticated tomato species and that traits and genes divergent between domesticates and progenitor are more likely to be plastic than expected by chance. To test this, we assess (i) the difference in traits and gene expression between these species under control conditions; (ii) the plasticity in traits and gene expression within each species; (iii) divergence in trait and gene expression plasticity during domestication. These will give insights into whether plasticity played a role during the early domestication of tomatoes, with implications for future food security.

2.3 Materials and Methods

2.3.1 Plant material and growth conditions

Self-compatible tomato accessions (Table A1) were grown in large pots with a soil mix of 2:1 Levington compost (F2+sand) and vermiculite in the glasshouses at the University of Southampton for one generation to reduce maternal effects. For the following generation, a plasticity pot experiment was set up (Figure 2.1A) that included the domesticated tomato (D) *Solanum lycopersicum* (SLL; number of accessions (n)=2) and *S. lycopersicum* var *cerasiforme* (SLC; n=2), the progenitor (P) *S. pimpinellifolium* (SP; n=2) and the never-domesticated tomato wild species (W) *S. cheesmaniae* (SChe; n=2) and *chmielewskii* (SChm; n=2). Seeds were obtained from Tomato Genetic Resource Centre (TGRC; https://tgrc.ucdavis.edu/) and Centre for Genetic Resources the Netherlands (CGN) Wageningen University (https://cgngenis.wur.nl/). Accessions were grown in three experimental treatments: control, root crowding and low nutrient (Figure 2.1B): (i) control treatment had large pots with a soil mix

of 2:1 Levington compost (F2+sand) and vermiculite; (ii) root crowding treatment had small pots with the same soil mix as the control; (iii) low nutrient treatment had large pots with low nutrient soil mix composed of 1:1 Levington compost (F2+sand) and vermiculite. Four replicates per treatment per accession (120 plants) were grown. The metadata for this experiment is provided in Table A1.

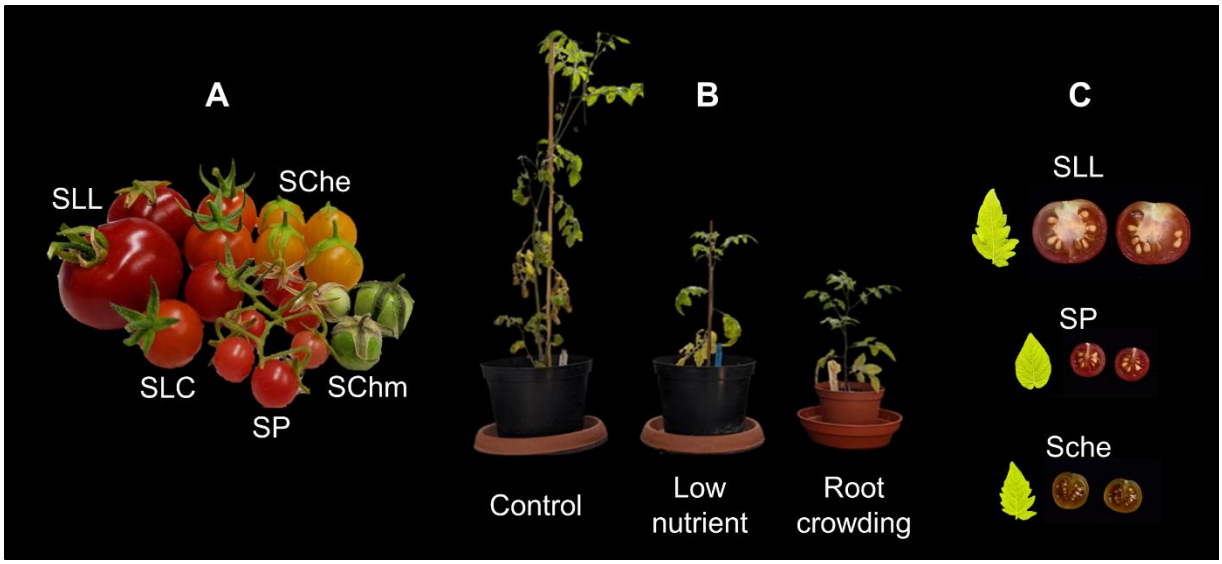


Figure 2.1: Plasticity pot experiment.
Tomato species, including the domesticates *Solanum lycopersicum* (SLL), and *S. lycopersicum* var. *cerasiforme* (SLC), the progenitor *S. pimpinellifolium* (SP), and the never-domesticated wilds, *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm) (A), were grown under control, low nutrient and root crowding treatments in a pot trial. We follow the prevailing idea that *S. lycopersicum* and *S. lycopersicum* var *cerasiforme* are domesticated from the progenitor *S. pimpinellifolium*. (B) Plant morphology of accession SLC, LA1324 for different treatments. (C) Representative leaf and fruit samples from domesticated (SLL, LA0395), progenitor (SP, LA1578) and never-domesticated wild (SChe, LA1039) tomato accessions.

2.3.2 Phenotypic analyses

Sixteen traits were measured for each plant in the pot experiment (Table 2.1). The first mature fruit from each plant was harvested, cut longitudinally, and scanned using a Canon CanoScan LiDE 300 and analysed using Tomato Analyzer version 4.0 software (Rodriguez *et al.*, 2010) at 600 dpi with units of cm (Figure 2.2). Seeds were collected from the first mature fruits and

cleaned (Appendix Methods A. 1). Seed number and size were calculated using ImageJ (Schneider *et al.*, 2012).

Table 2.1: Phenotypic traits affected by domestication.

Phenotypic traits	measurements	references
Developmental stages	days to flowering days to fruiting fruit development (days)	Grandillo and Tanksley (1996)
Plant height	height at flowering (cm) height at fruiting (cm)	Grandillo and Tanksley (1996); Pnueli *et al.* (1998)
Fruit traits	number of fruits fruit weight (g) fruit yield (g/plant)	Frary *et al.* (2000)
Fruit morphology (Figure 2.2)	fruit perimeter (mm) fruit area (mm^2) fruit maximum width (mm) pericarp area (mm^2) pericarp thickness (mm)	Tanksley (2004)
Seed traits	number of seeds seed size (mm^2) seed yield (seed/plant)	Doganlar *et al.* (2000)

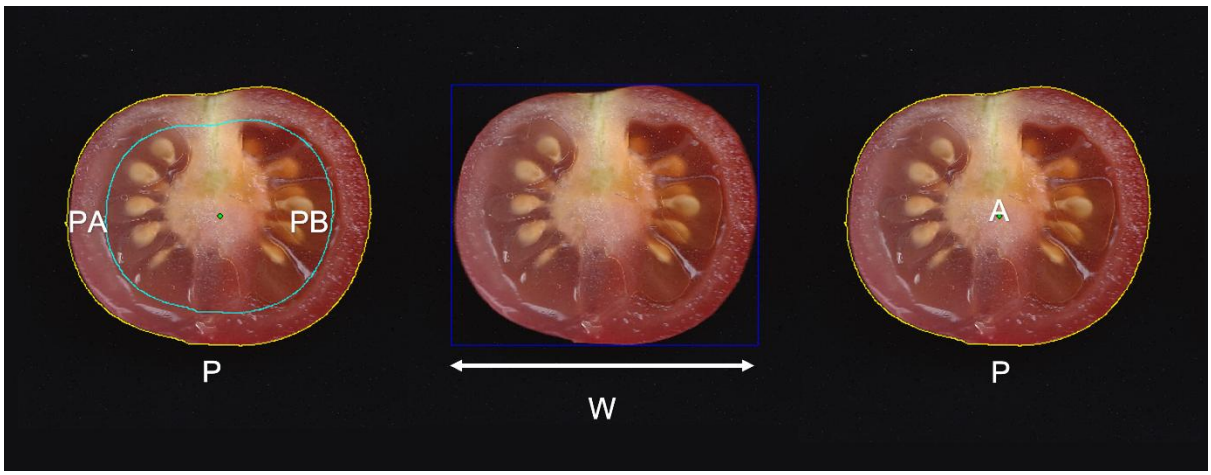


Figure 2.2: Fruit morphology measurements with Tomato Analyzer 4.0.
The following were measured: Fruit perimeter (P); fruit area (A); maximum width (W); pericarp area (PA; the area between the pericarp boundary [PB] and P); Pericarp thickness (PT; PA divided by the average of PB and P).

2.3.2.1 Statistical analysis of phenotypic traits

Statistical analyses of phenotypic traits were split into: (i) traits measured under control treatments only, to explore traits which are divergent between species ["interspecific": 3 species x 1 treatment]; (ii) traits measured under control, root crowding and low nutrient treatments to explore phenotypic plasticity in all species ["plasticity": 3 species x 3 treatments]; (iii) the overlap between traits with plasticity in either domesticates, progenitor or wilds and divergent between the domesticates and the progenitor ["plasticity divergence": 3 species x 3 treatment].

To test if the number of divergent (compared to non-divergent) or plastic (compared to non-plastic) traits in each species and/or group was statistically significant at a probability of 0.5, a two-way binomial test was performed. To test for significant differences between species under control, a one-way ANOVA (analysis of variance) was implemented if homogeneity and normality assumptions were satisfied, diagnostic plots were checked. Homogeneity assumption was tested with Levene's Test for Homogeneity of Variance. The normality assumption was tested with the Shapiro-Wilk normality test. The ANOVA was followed by the Tukey test for multiple comparisons of means. For parameters that did not satisfy the assumptions for ANOVA, Generalised Linear Models (GLMs) with an identity link function and variance^2 error distribution were performed to normalise residuals. A statistically significant difference in treatment effect between pairs of treatments suggests detectable plasticity. To identify the relationship between traits, a correlation matrix using the Spearman rank correlation test was generated. To test for the overlap between divergent and plastic genes, Fisher's exact test was performed between divergent traits between the domesticated and the progenitor species and traits plastic in (i) the domesticated species, (ii) the progenitor species, (iii) the wild species. All statistical tests and graphs were generated using R v.4.2.1 (R Core Team, 2022).

2.3.3 Gene expression analyses

2.3.3.1 RNA extraction and sequencing

We only used species which were closely related to *S. lycopersicum*; this included the domesticated *S. lycopersicum*, the progenitor *S. pimpinellifolium* and the wild species, *S. cheesmaniae*.

Leaf and fruit samples were taken from the domesticated, the progenitor and the wild species (Figure 2.1C; Table A1). For the domesticated and, the progenitor species, two accessions x three treatments each were sampled for both leaf and fruit tissues (24 samples; accessions are considered as replicates). For the wild species, two accessions with three treatments were sampled for leaves but only control plants produced fruits (8 samples). To reduce variation among samples due to developmental stage, leaf samples were taken at the 8-leaf stage and the fruit samples were the first mature fruits. Sampling was done at the same time of day to reduce variation. Tissue samples were ground in liquid nitrogen and RNA extraction was performed using RNeasy Plant Mini Kit (Qiagen) according to the protocol for the purification of total RNA from plant cells and tissues. To remove any contamination of genomic DNA, the On-Column DNase Digestion protocol was performed. The quality of RNA extraction was checked using a NanoDrop Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Samples were sent to Novogene Bioinformatics Institute (Cambridge, UK) and after confirming RIN > 6.0 were sequenced using the Illumina NovaSeq 6000 Sequencing System generating at least 6GB of data per sample.

2.3.3.2 Differential gene expression analysis

Raw RNA sequences were quality-checked with FASTQC v. 0.11.9 (Andrews, 2010). Sequences were then trimmed and filtered using Trimmomatic v0.32 (Bolger *et al.*, 2014b) to remove poor-quality bases and reads: Illumina adapters were removed; leading and trailing bases with quality below 5 were removed; reads shorter than 72bp were removed; sliding window trimming was performed for a window size of 4 with the required quality of 20. Quantification of gene expression data was performed using STAR v2.7.10a (Doblin *et al.*, 2013) and RSEM v1.3.1 (Li *et al.*, 2011). STAR genome index was generated with the reference *Solanum lycopersicum* SL4.0 genome (Hosmani *et al.*, 2019). Alignments were then performed using the trimmed reads and

RSEM uses the transcript coordinates from STAR to quantify gene expression and generate a matrix of the results. Transcriptome processing statistics can be found in Table A1.

For differential expression analysis, the R package DESeq2 v1.36.0 (Love *et al.*, 2014) was used in R v.4.2.1 (R Core Team, 2022). DESeq2 is designed to statistically test count data in gene expression with small numbers of replicates (Seyednasrollah *et al.* 2015). Three datasets were created to identify the following: (i) differentially expressed genes (DEGs) between species in leaves and fruits under control treatments [hereafter "interspecific" analysis: 3 species x 1 treatment x 2 tissues], (ii) genes in each species with plasticity in leaves (i.e. DEGs between treatments) ["plasticity" analysis: 3 species x 3 treatments x 1 tissue], (iii) overlap between genes with plasticity in the domesticated, progenitor or wild species (in leaves and fruits) and divergent between the domesticated and the progenitor ["plasticity divergence": 3 species x 3 treatments x 2 tissues]. The 'plasticity divergence' analysis uses the 'plasticity' dataset for leaf analysis and a 'fruit' dataset of the domesticated and the progenitor species (2 species x 3 treatments x 1 tissue) because plasticity analysis for the wild species cannot be performed due to lack of fruits from root crowding and low nutrient treatments. Leaf and fruit samples were analysed separately. DEGs were filtered with an adjusted p-value < 0.05 and $\log_2$ Fold Change > 0.58, which corresponds to a 1.5-fold change in expression.

2.3.3.3 Gene Ontology (GO) analysis

To understand the biological processes associated with the DEGs, TopGO (Alexa and Rahnenfuhrer, 2022) was used to test the over-representation of gene ontology (GO) terms with genes of interest identified from the differential gene expression analysis. SlimGO annotation of ITAG4.0 *Solanum lycopersicum* was used in a gene-to-GO format (http://systemsbiology.cau.edu.cn/agriGOv2/download.php). Fisher's exact tests compared DEGs and the background list of all genes associated with their GO annotation, focusing on biological processes as the ontology of interest with a minimum of five genes per GO term. P-values were adjusted using the Benjamini & Hochberg (1995) method to control the false discovery rate with a p-value < 0.05 was considered significant. The enrichment ratio per GO term refers to the number of annotated DEGs relative to the expected number by chance.

2.3.3.4 Pathway analysis

To identify significant biological pathways associated with DEGs, the KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology Based Annotation System (KOBAS 3.0) (Bu *et al.*, 2021) was performed. FASTA nucleotide sequences of genes of interest were extracted from the SL4.0 tomato reference (Hosmani *et al.*, 2019) using seqkit (Shen *et al.*, 2016). This input gene list was annotated with the KEGG pathway database of *Solanum lycopersicum* (tomato) and employed Fisher's exact tests to statistically identify significant pathways with Benjamini and Hochberg (1995) FDR correction. KEGG pathways with adjusted p-value < 0.05 and containing five or more genes were considered significant. The enrichment ratio refers to the number of DEGs relative to the total number of genes in the pathway.

2.3.3.5 Focus on known fruit domestication genes

To assess whether known domestication genes demonstrated a change in gene expression during domestication, a literature search to identify domestication genes related to fruits was carried out and we identified eleven genes (Table 2.2). For each domestication gene, we assessed whether the gene (i) showed divergence in expression between species in the fruit datasets, and (ii) showed divergence between the domesticated and the progenitor species and was plastic in these species.

Table 2.2: List of domestication genes related to fruit transcriptome changes.

Locus	Gene description	Protein	Phenotype affected	Reference
Solyc02g083950	*wuschel*	Transcription factor	Fruit size	Li *et al*. (2017a)
Solyc02g090730	*FW2.2* (fruit weight 2.2)	Cell number regulator	Fruit size	Frary *et al*. (2000)
Solyc11g071380	*CLAVATA3*	Transcription factor	Fruit size	Chu *et al*. (2019)
Solyc11g071810	*fasciated*	Transcription factor	Fruit size	Huang *et al*. (2013)
Solyc05g005240	*YABBY*	Transcription factor	Fruit shape	Huang *et al*. (2013)
Solyc01g006540	*TomloxC* (lipoxygenase C)	Lipoxygenases	Lipids and volatiles	Gao *et al*. (2019)
Solyc01g079620	*MYB12*	Transcription factor	Flavonoids	Zhu *et al*. (2018)

Solyc09g010080	*Lin5* (invertase 5)	Beta-fructofuranosidase	Sugar	Tieman *et al.* (2017)
Solyc09g089580	*E8*	1-Aminocyclopropane-1-carboxylate oxidase	Volatiles	Tieman *et al.* (2017)
Solyc10g085230	ripening-related mRNA	UDP-glycosyltransferase	Steroidal alkaloids	Zhu *et al.* (2018)
Solyc12g055730	*LIP1*	Lipase	Lipids and volatiles	Garbowicz *et al.* (2018)

2.3.3.1 Differential gene expression

To test if the number of DEGs in each species was statistically significant at a probability of 0.5, a two-way binomial test was performed. To test for the difference between DEGs in leaf and fruit, a paired t-test was performed with each species comparison. To test for association a chi-squared test was performed for leaf and fruit data between (i) DEGs between the domesticated and progenitor species and genes plastic in the domesticated species; (ii) DEGs between the domesticated and progenitor species and genes plastic in the progenitor species; (iii) DEGs between the domesticated and progenitor species and genes plastic in the wild species. All statistical tests and graphs were generated using R v.4.2.1 (R Core Team, 2022).

2.4 Results

2.4.1 Phenotypic analyses

We are interested in whether species that were domesticated had more phenotypic plasticity than those which were not domesticated, and whether this plasticity aided the process of domestication. We therefore compared the phenotypes of 16 traits in domesticated tomato (*S. lycopersicum* and *S. lycopersicum* var. *cerasiforme*) against its wild progenitor (*S. pimpinellifolium*) and two other closely related wild species (*S. cheesmaniae* and *S. chmielewskii*), that were not domesticated. Plasticity of a trait was inferred when a significant difference between any pair of treatments within a species was identified. Note that not all plants produced fruits and no fruits were obtained from the wild species, *S. cheesmaniae*, under low nutrient treatments.

Several phenotypes were highly correlated (Table A2; Appendix Figure A. 1). Fruit weight, yield, perimeter, area, width, pericarp area, pericarp thickness and seed size were all positively correlated (rho > 0.70, p < 0.01). Many fruit-related characters were negatively correlated with vegetative characters (i.e., smaller plants had more and/or larger fruits). A summary of the phenotypic data and statistical outputs are reported in Table A3 to A10.

2.4.1.1 Interspecific analysis of traits

To assess how domestication traits differ among domesticated, progenitor and wild species, interspecific analysis with individual control plants from each species were measured for 16 traits to identify divergent traits between species (5 species x 1 treatment). Principal Component Analysis (PCA) and the first two principal components (Appendix Figure A. 2), showed that variability between individuals within species was larger across PC1 (with seed yield as the highest contributor) compared to PC2, with 94.2% and 3.9% respectively (Table A7; Appendix Figure A. 3).

The number of divergent traits between species (i.e., significantly different traits between species) reflected the genetic relatedness of species (Table A8; Figure 2.3A), with the greatest number of divergent traits between the domesticated tomato and other wild tomato species (12 out of 16; binomial: p = 0.077); whilst the lowest number of divergent traits was between the two domesticated species, as well as the progenitor and other wild species, *S. cheesmaniae* (7 out 16; binomial: p = 0.804).

Divergent traits (i.e., significantly different between at least one member of each group) were identified between domesticates, progenitors and wilds (Table 2.3; Figure 2.3B). The greatest number of divergent traits were identified between the domesticates and progenitor (13 out of 16; binomial: p = 0.021) as well as between the domesticates and the wilds (13 out of 16; binomial: p = 0.021). The lowest number of divergent traits was between the progenitor and the wilds (11 out of 16; binomial: p = 0.210). The majority of traits were divergent between all groups (Figure 2.3C).

Table 2.3: Summary of divergent traits between groups.

Divergent traits between domesticates (D; *S. lycopersicum* and *S. lycopersicum* var *cerasiforme*), progenitor (P; *S. pimpinellifolium*) and never-domesticated wilds (W; *S. cheesmaniae* and *S. chmielewskii*) under control conditions.

Traits	D vs P	D vs W	P vs W
Days to flowering	×	✓	✓
Days to fruiting	✓	✓	✓
Fruit development	✓	✓	×
Height at flowering	✓	✓	✓
Height at fruiting	✓	✓	✓
Fruit number	×	×	✓
Fruit weight	✓	✓	✓
Fruit yield	✓	✓	✓
Fruit perimeter	✓	✓	×
Fruit area	✓	✓	✓
Fruit width	✓	✓	×
Pericarp area	✓	✓	✓
Pericarp thickness	✓	✓	✓
Number of seeds	✓	×	×
Seed size	✓	✓	✓
Seed yield	×	×	×
Total	13	13	11

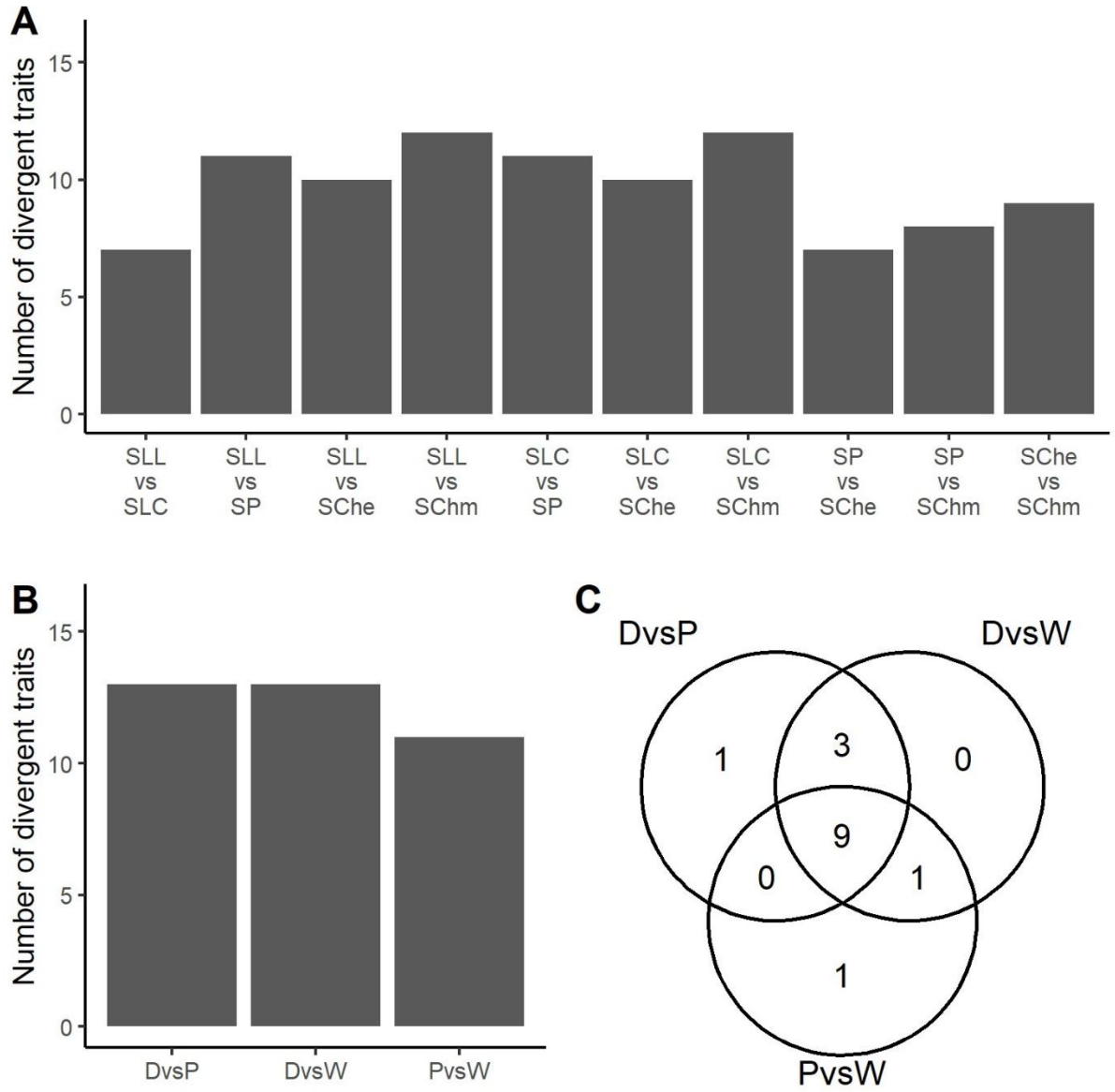


Figure 2.3: Divergent traits in the interspecific analysis.

Divergence of traits between domesticates (D; *S. lycopersicum* and *S. lycopersicum* var *cerasiforme*), progenitor (P; *S. pimpinellifolium*) and never-domesticated wilds (W; *S. cheesmaniae* and *S. chmielewskii*) under the control treatment. (A) The number of divergent traits between each pair of species. (B) The number of divergent traits between each group. (C) Venn diagram of shared divergent traits between each pair of groups. Note that seed yield was not divergent between any pairs of tomato groups.

We find that the domesticated species was detectably different at the 95% significance level from the wild progenitor and the two other wild species for most traits particularly noteworthy

were the increase in fruit weight (Figure 2.4G), fruit yield (Figure 2.4H), all fruit morphology traits (Figure 2.4I-M) and seed size (Figure 2.4O) compared to the progenitor and other wild species. Comparing progenitors to domesticates tells us that tomato domestication incurred a reduction in height at flowering and fruiting, as well as an increase in fruit weight, fruit size and seed size. These traits have made tomatoes a successful domesticated species and are linked to the increase in fruit size and harvest index during domestication.

The progenitor species was detectably different from both wild species at the 95% significance level for days to flowering (Figure 2.4A), fruit yield (Figure 2.4H) and seed size (Figure 2.4O). A fewer number of days to flowering, greater fruit yield and greater seed size in the progenitor compared to the wilds, could have given the progenitor a competitive advantage over other wilds (i.e. more fruits would be more attractive to early farmers).

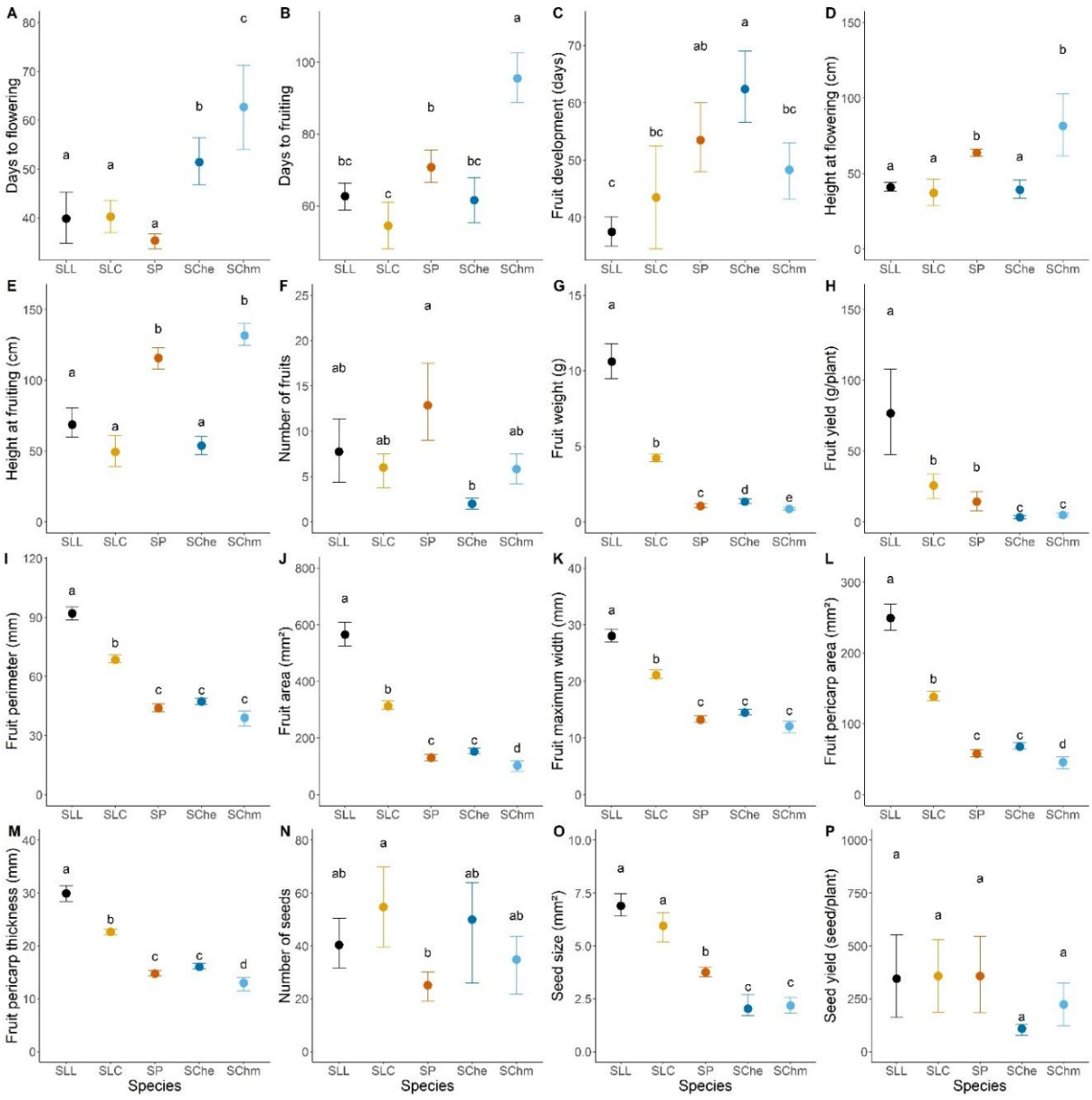


Figure 2.4: Interspecific analysis of phenotypic traits under control treatments. Panels represent (A) days to flowering, (B) days to fruiting, (C) length of fruit development, (D) height at flowering, (E) height at fruiting, (F) fruit number, (G) fruit weight (g), (H) fruit yield (g/plant), (I) fruit perimeter (mm), (J) fruit area (mm^2), (K) fruit maximum width (mm), (L) fruit pericarp area (mm^2), (M) fruit pericarp thickness (mm), (N) seed number, (O) seed size (mm^2), (P) seed yield (seed/plant). Species are abbreviated as *S. lycopersicum* (SLL), *S. lycopersicum* var. *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm). Letters represent statistically significance at p = 0.05 (1-way ANOVA or GLM). Species with the same letter do not significantly differ in their means. If two species have different letters, their means differ significantly. Dots and bars are the mean and standard error, respectively.

2.4.1.2 Plasticity analysis of traits

The analyses of phenotypic plasticity for each trait were performed per species (5 species x 3 treatments). Evidence of plasticity was a detectable difference in a phenotypic trait between at least one pair of treatments (summarised in Table 2.4; full test outputs in Table A9; Figure 2.5). Principal Component Analysis (PCA; Appendix Figure A. 4) demonstrates that the variability in the dataset was mostly explained by PC1 (93.2%) that PC2 (4.8%; Table A10; Appendix Figure A. 5).

Table 2.4: Summary of plastic traits in each species and group.
Traits are given per species for *Solanum lycopersicum* (SLL), *S. lycopersicum* var. *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm). These are summarised for the domesticates (D), progenitors (P) and never-domesticated wilds (W).

Traits	Plastic in SLL	Plastic in SLC	Plastic in SP	Plastic in SChe	Plastic in SChm	Plastic in D	Plastic in P	Plastic in W
Days to flowering	✓	✓	✓	×	✓	✓	✓	✓
Days to fruiting	✓	✓	×	×	×	✓	×	×
Fruit development	×	✓	×	×	✓	✓	×	✓
Height at flowering	✓	✓	✓	×	×	✓	✓	×
Height at fruiting	✓	✓	✓	✓	✓	✓	✓	✓
Fruit number	✓	×	✓	×	✓	✓	✓	✓
Fruit weight	✓	×	✓	×	×	✓	✓	×
Fruit yield	✓	×	✓	×	✓	✓	✓	✓
Fruit perimeter	×	×	×	×	×	×	×	×
Fruit area	×	×	✓	×	×	×	✓	×
Fruit width	✓	×	✓	×	×	✓	✓	×
Pericarp area	×	×	✓	×	×	×	✓	×
Pericarp thickness	×	×	✓	×	×	×	✓	×
Number of seeds	×	×	×	✓	×	×	×	✓
Seed size	×	✓	×	×	×	✓	×	×
Seed yield	✓	×	✓	✓	✓	✓	✓	✓
Total	9	6	11	3	6	11	11	7

Per species (Figure 2.5A and B), more traits were plastic in the progenitor (11 out of 16; binomial: p = 0.210) than in the domesticates (SLL = 9 [binomial: p = 0.804]; SLC = 6 [binomial: p = 0.455]), and other wilds (SChe = 3 [binomial: p = 0.021]; SChm = 6 [binomial: p = 0.455]).

Several traits were plastic in multiple species with height at fruiting being the sole trait plastic in all species. Per tomato group (Figure 2.5C and D), a plastic trait in a group implies the trait was plastic in at least one species in the group. Of the 16 traits measured, 11 were plastic in the domesticates and progenitor (binomial: p = 0.210) and 7 in wilds (binomial: p = 0.804), and many traits were plastic in two or three tomato groups (Table 2.4).

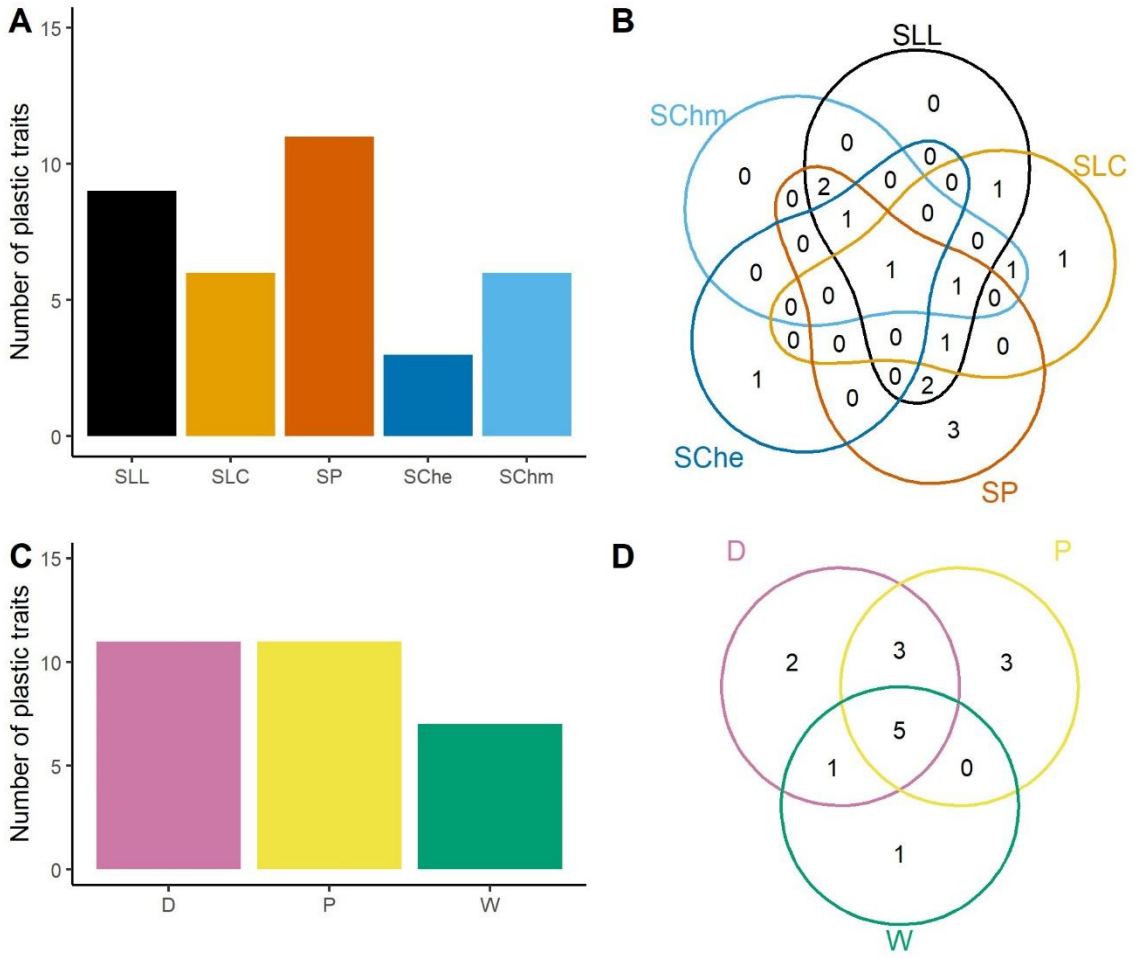


Figure 2.5: Plasticity within species in the plasticity analysis.
Traits were measured for five species, *S. lycopersicum* (SLL), *S. lycopersicum* var. *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm) under three treatments (control, root crowding and low nutrient). Plastic traits are those which are significantly different between at least one pair of treatments. (A) Number of plastic traits in each species. (B) Venn diagram demonstrating the overlap in plastic traits between species. (C) Number of plastic traits within groups. (D) Venn diagram of shared plastic traits between groups. Note that fruit perimeter was not plastic in any species.

Each species responded differently to different treatments resulting in variability in plasticity within tomato groups (Figure 2.6). Traits plastic in the progenitor and not in other wilds include height at flowering (Figure 2.6D), fruit weight (Figure 2.6G), fruit area (Figure 2.6J), fruit width (Figure 2.6K), pericarp thickness (Figure 2.6L) and area (Figure 2.6M). With fruit area, pericarp area and thickness only plastic in the progenitor. This greater plasticity in the progenitor for fruit traits and morphology may have provided important phenotypic variation that gave the progenitor an advantage in early domestication (i.e. bigger fruits in resource-rich human-modified fields). No plasticity was detected in the domesticates for the fruit area (Figure 2.6J), fruit width (Figure 2.6K), pericarp thickness (Figure 2.6L), pericarp area (Figure 2.6M) and seed number (Figure 2.6N) compared to progenitors or other wilds. Lack of plasticity in these fruit and seed traits may benefit farmers to obtain consistently high yields for each harvest.

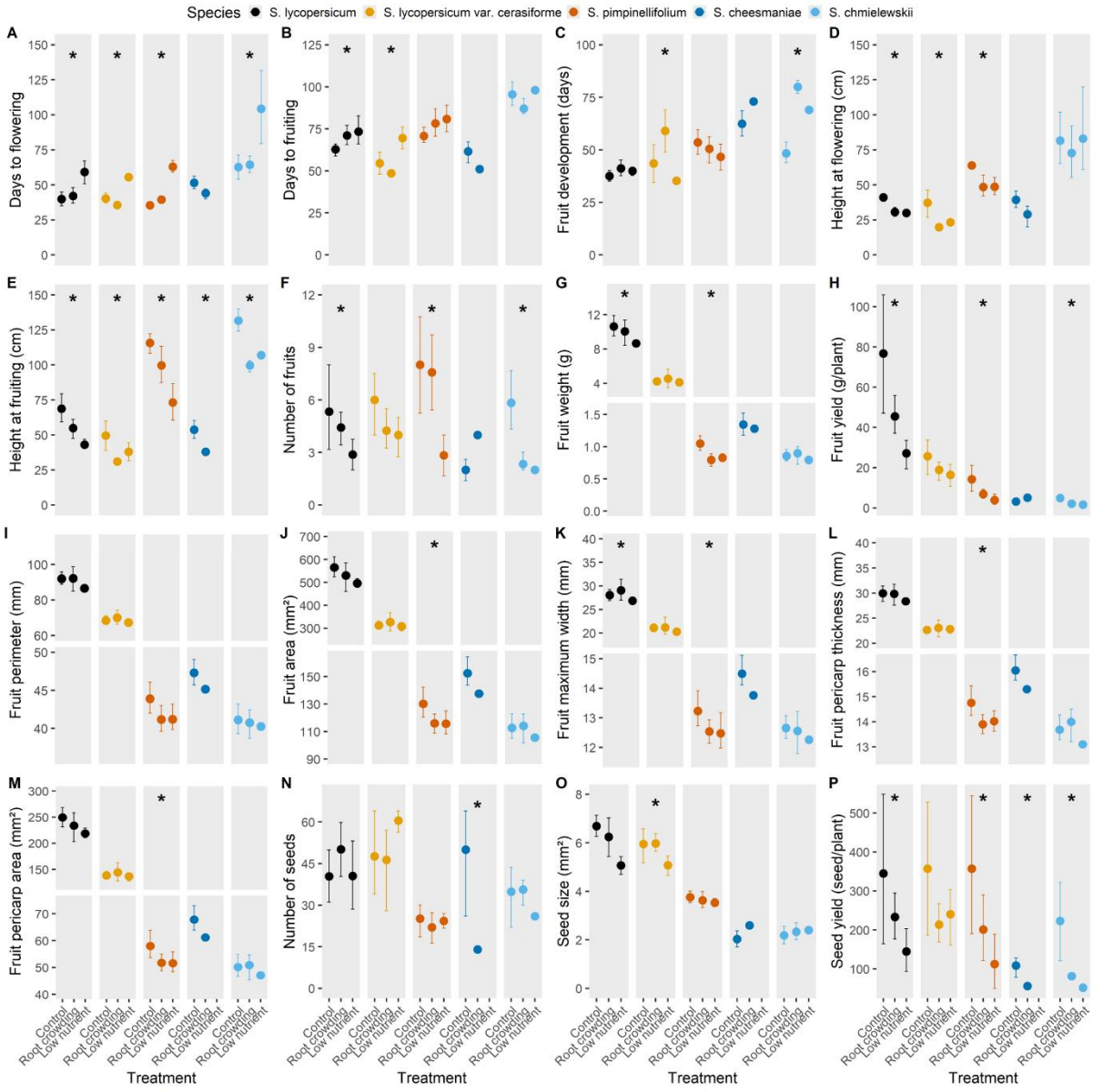


Figure 2.6: Phenotypic plasticity of 16 traits in the plasticity analysis.

Phenotypes are (A) days to flowering, (B) days to fruiting, (C) length of fruit development, (D) height at flowering, (E) height at fruiting, (F) fruit number, (G) fruit weight (g), (H) fruit yield (g/plant), (I) fruit perimeter (mm), (J) fruit area (mm^2), (K) fruit maximum width (mm), (L) fruit pericarp area (mm^2), (M) fruit pericarp thickness (mm), (N) seed number, (O) seed size (mm^2), (P) seed yield (seed/plant). Species are *S. lycopersicum* (SLL), *S. lycopersicum* var. *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm) and treatments are control, root crowding and low nutrient. * indicates plasticity detected in the species. Note: no fruit were obtained from SChe under low nutrient treatment. Dots and bars are the mean and standard error, respectively. Noteworthy are many fruit traits (F-M) that are plastic in the progenitor and not in the wilds. Refer to Table A9 for full statistical outputs.

2.4.1.3 Plasticity divergence analysis of traits

The overlap between plasticity and divergence could tell us whether plastic traits are those which were selected on. The greatest overlap in divergence between the domesticates and the progenitor species, and plasticity in species was detected in the progenitor followed by the domesticated species, and the wild species (Table 2.5; Appendix Figure A. 6). Height at fruiting was the only trait divergent and plastic in all species. Overlap in divergence and plasticity in the progenitor included height measurements, fruit weight, fruit yield, fruit area, width, pericarp area and thickness. There was no significant deviation from that expected by chance between plastic and divergent traits for all the species (Fisher's Exact test: $p > 0.05$), apart from the divergence between the domesticated (SLC) and the progenitor and plasticity in a wild species (SChm; Fisher's Exact test: $p = 0.001$). This was due to the greater number of divergent traits between the domesticated and the progenitor that were not plastic in the wild (SChm). The association between divergent traits and plasticity was only detected in the wild (SChm), i.e. divergent traits were more likely to not be plastic. Given that these plastic traits were under selection during domestication, they may have played a role in the early domestication of the progenitor.

Table 2.5: Summary of divergent and plastic traits between groups.

Overlap between plastic traits in *S. lycopersicum* (SLL) and *S. lycopersicum* var *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe), and *S. chmielewskii* (SChm) that are also divergent between domesticates (SLL and SLC) and the progenitor (SP).

Traits	Divergent between SLL vs SP and plastic in SLL	Divergent between SLL vs SP and plastic in SLC	Divergent between SLL vs SP and plastic in SP	Divergent between SLL vs SP and plastic in SChe	Divergent between SLL vs SP and plastic in SChm	Divergent between SLC vs SP and plastic in SLL	Divergent between SLC vs SP and plastic in SLC	Divergent between SLC vs SP and plastic in SP	Divergent between SLC vs SP and plastic in SChe	Divergent between SLC vs SP and plastic in SChm
Flowering time	×	×	×	×	×	×	×	×	×	×
Fruiting time	×	×	×	×	×	✓	✓	×	×	×
Maturation time	×	✓	×	×	✓	×	×	×	×	×
Height at flowering	✓	✓	✓	×	×	✓	✓	✓	×	×
Height at fruiting	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fruit number	×	×	×	×	×	×	×	×	×	×
Fruit weight	✓	×	✓	×	×	✓	×	✓	×	×
Fruit yield	✓	×	✓	×	✓	×	×	×	×	×
Fruit perimeter	×	×	×	×	×	×	×	×	×	×
Fruit area	×	×	✓	×	×	×	×	✓	×	×
Fruit width	✓	×	✓	×	×	✓	×	✓	×	×
Pericarp area	×	×	✓	×	×	×	×	✓	×	×
Pericarp thickness	×	×	✓	×	×	×	×	✓	×	×
Number of seeds	×	×	×	×	×	×	×	×	✓	×
Seed size	×	✓	×	×	×	×	✓	×	×	×
Seed yield	×	×	×	×	×	×	×	×	×	×
Total	5	4	8	1	3	5	4	7	2	1

2.4.2 Gene expression analyses

In addition to phenotypic traits, we are also interested in whether species that were domesticated have more gene expression plasticity than those which were not domesticated, aiding the process of domestication. To investigate this we compared gene expression of leaves and fruits from domesticated tomato (*S. lycopersicum*) against its wild progenitor (*S. pimpinellifolium*) and one other closely related wild species (*S. cheesmaniae*). Differential expression between any pair of treatments is used as evidence of plasticity for that gene (Differentially expressed genes; DEGs). Raw sequencing data resulted in an average of 24.1Mb (+/-0.44; SE) reads per sample. Data cleaning and filtering retained on average 93.08% of the reads per sample, with 92.22% to 96.47% of the cleaned data uniquely mapped onto the SL4.0 reference genome. No significant difference in the percentage of mapped reads was observed between species (ANOVA: $p > 0.05$). RNA filtering and mapping statistics are reported in Table A1.

2.4.2.1 Interspecific analysis of gene expression

To assess interspecific differences in gene expression, differential gene expression analysis for leaf and fruit transcriptomes was performed between the domesticated, the progenitor and the wild species grown under control conditions (3 species x 1 treatment x 2 tissues). PCA shows a clear distinction between species for both tissues (Figure 2.7A and B), with greater relative variability between individuals within species for the leaf than fruit transcriptomes. There were more DEGs between species in fruits compared to leaves (Table 2.6), but this was not statistically significant ($t = 3.9201$, $df = 2$, p-value = 0.059). The greatest number of DEGs from both leaf and fruit comparisons were identified in the progenitor vs wilds, followed by domesticates vs wild and then domesticated vs progenitor (Table 2.6; Figure 2.7C&D). DEGs shared between different comparisons were the greatest between domesticated vs wilds and progenitor vs wilds, for both leaves (Figure 2.7E) and fruits (Figure 2.7F). A full list of DEGs is reported in Table A11.

The number of DEGs between domesticated vs wilds was unexpectedly less than progenitor vs wilds in leaves (DvW=294; PvW=399) and fruits (DvW=1637; PvW=2006), which suggests that the domesticated transcriptome had evolved during domestication to be more wild-like, and indeed this was the case. In leaves, all 68 DEGs had the same direction of expression difference in domesticated vs progenitor and wilds vs progenitor (Figure 2.7E) and in fruits, 338 DEGs out of

366 had the same direction of expression difference (Figure 2.7F). This seems to account for the fewer DEGs in domesticated vs wilds compared to progenitor vs wilds.

Table 2.6: DEGs and associated GO terms and KEGG pathways in the interspecific analysis. Differentially expressed genes (DEGs) in leaf and fruit between domesticated, progenitor and wild species under control treatments, with their associated significant GO terms and KEGG pathways (full list of genes, GO and KEGG terms can be found in Tables A11, A12 and A13).

Species	Tissue	DEGs		GO terms	KEGG pathways
Domesticated vs Progenitor	Leaf	151	(0.720%)	9	0
Domesticated vs Wild	Leaf	294	(1.402%)	9	3
Progenitor vs Wild	Leaf	399	(1.903%)	9	0
Domesticated vs Progenitor	Fruit	750	(3.578%)	74	8
Domesticated vs Wild	Fruit	1637	(7.809%)	94	19
Progenitor vs Wild	Fruit	2006	(9.569%)	32	14

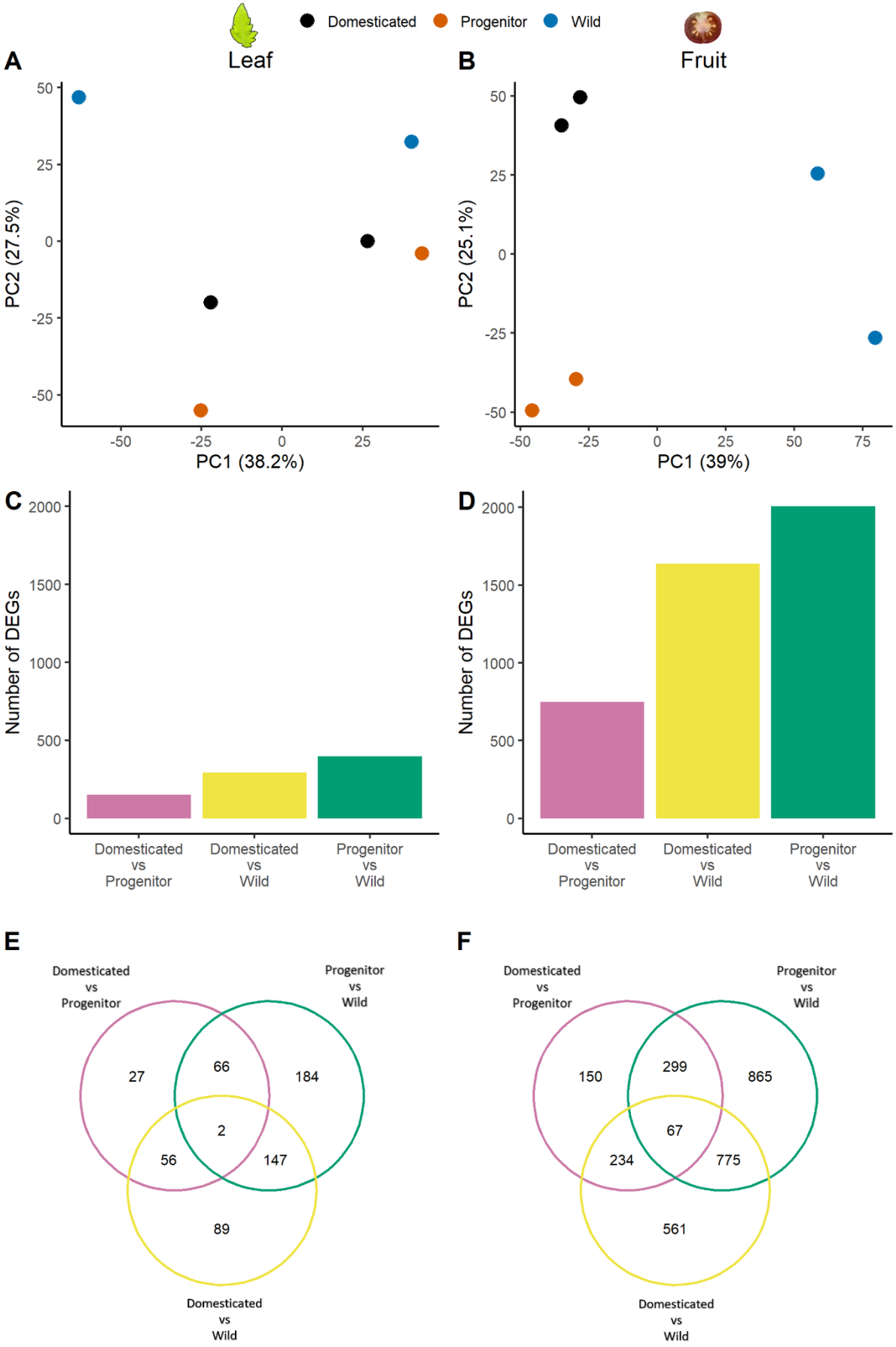


Figure 2.7: Gene expression analysis of the interspecific dataset.

Principal component analysis (first two principal components) of (A) leaf and (B) fruit transcriptomes. Differentially expressed genes (DEGs) between domesticated, progenitor and wild species under control treatments in (C) leaf and (D) fruit. Venn diagrams of DEGs shared between each species comparisons for (E) leaf and (F) fruit.

Significantly enriched biological process GO terms (Table A12) and KEGG pathways (Table A13) from DEGs were identified. For leaves, the DEGs were significantly enriched in nine GO terms for all three comparisons (Table 2.6), five of which were significantly enriched in all three comparisons and these all relate to response to biotic stimuli, especially to other organisms (Figure 2.8A). For DEGs between domesticated vs progenitor and domesticated vs wild, many enriched GO terms were linked to the same set of seven genes including glycosyltransferases and terpene synthases. DEGs between progenitor vs wild were enriched in GO terms related to the biotic environment resulting in 67 genes including methylesterase genes. Specific to DEGs between progenitor vs wild, GO terms related to defence response were enriched. No enriched KEGG pathways were identified in the DEGs between domesticated vs progenitor and between progenitor vs wild; three KEGG pathways were enriched in the DEGs between domesticated vs wild.

In fruits, the greatest number of enriched GO terms was reported between domesticated vs wild followed by domesticated vs progenitor and then progenitor vs wild (Table 2.6). Enriched GO terms between domesticated vs wild included response to stimulus and response to nutrient levels (Table A12); enriched GO terms between domesticated vs progenitor included defense response, and response to abiotic stimulus (Table A12); enriched GO terms between progenitor vs wild included photosynthesis, light harvesting and response to stress (Table A12). Between domesticated vs progenitor and domesticated vs wild, 41 enriched GO terms were shared (shared GO terms not all shown in Figure 2.8B); notably GO terms associated with lipid metabolism, secondary metabolite biosynthetic process, and response to abscisic acid, but these were not significant between the progenitor vs wild, indicating that the difference in gene expression may be due to domestication. There were 7 enriched GO terms specific between progenitor vs wild that included terms related to photosynthesis, cell wall metabolic processes, and response to various stress.

The greatest number of significantly enriched KEGG pathways were identified between domesticated vs wild, followed by progenitor vs wild and then domesticated vs progenitor with 19, 14 and 8 respectively (Table 2.6). Enriched KEGG pathways of DEGs in all fruit transcriptome comparisons were the following (Figure 2.8C): “Biosynthesis of secondary metabolites”,

“Carbon metabolism”, “Carotenoid biosynthesis”, and “Metabolic pathways”. Domesticated vs wild and progenitor vs wild comparisons both had enriched KEGG pathways in “Carbon fixation in photosynthetic organisms”,” Galactose metabolism”, “Glycolysis / Gluconeogenesis”, “Photosynthesis”, “Starch and sucrose metabolism” and “Sulfur metabolism”. These pathways may be important in domestication, differentiating the domesticated tomato from the wild relatives. Several pathways were only significant between progenitor vs wild, including “Amino sugar and nucleotide sugar metabolism”, “Biosynthesis of amino acids”, “Photosynthesis - antenna proteins”, and “Protein processing in endoplasmic reticulum”, indicating its importance in the differentiation between progenitor and wild. Enriched KEGG pathways between domesticated vs progenitor and domesticated vs wild comparisons were “Fructose and mannose metabolism”, which could indicate the importance of this pathway in tomato fruit domestication. “Cutin, suberin and wax biosynthesis”, “Phenylpropanoid biosynthesis”, and “Proteasome” pathways were only enriched in domesticated vs progenitor comparison; these could be pathways altered during domestication.

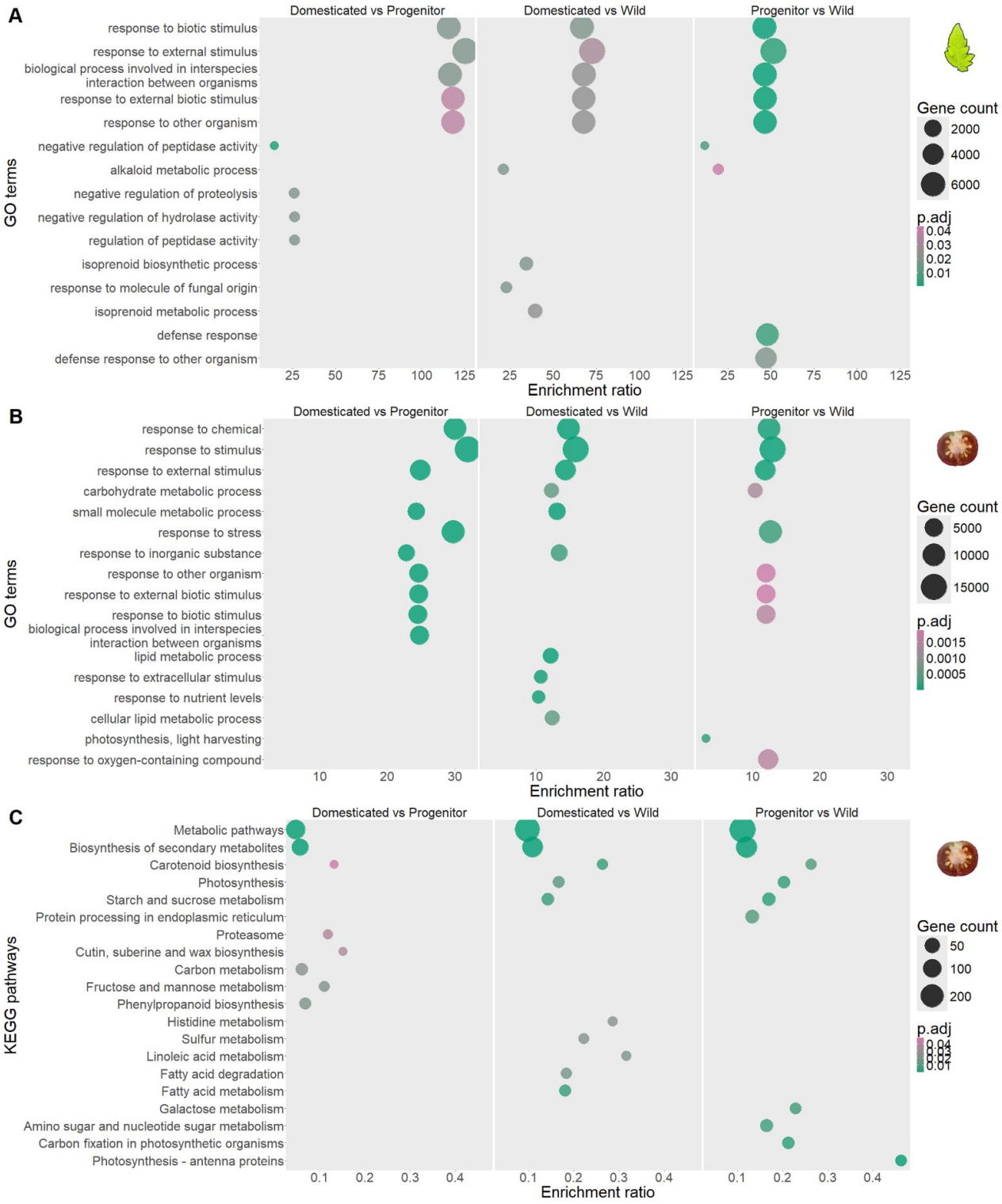


Figure 2.8: GO terms and KEGG pathways in the interspecific analysis.

Gene Ontology (GO) analysis of (A) leaf (all GO terms shown) and (B) fruit (top 10 GO terms based on p.adj) transcriptomes between domesticated, progenitor and wild species, with enrichment ratio and adjusted p-value (p.adj) for each GO term. (C) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of fruit transcriptomes between domesticated, progenitor and wild species, with enrichment ratio, gene counts and adjusted p-value (p.adj) for each KEGG pathway. All significant pathways are shown for Domesticated vs Progenitor and the top 10 pathways (based on p.adj) are shown for Domesticated vs Wild and Progenitor vs Wild.

Domestication genes (Table 2.2) were identified in fruit samples and five were DEGs in at least one interspecific comparison (*TomloxC*, *MYB12*, *LIN5*, UDP-glycosyltransferase, and *LIP1*; Figure 2.9). *TomloxC* had greater expression in the progenitor than in the domesticated species but *LIN5*, UDP-glycosyltransferase and *LIP1* had greater expression in the domesticated than in the progenitor species. *LIN5* and UDP-glycosyltransferase had greater expression in the domesticated than in the wild whereas *TomloxC* and *MYB12* had greater expression in the wild than in domesticated species. These genes have functions related to metabolic changes. Only one domestication gene, *LIP1* linked to fruit metabolites, had greater expression in the wild compared to the progenitor (Figure 2.9E).

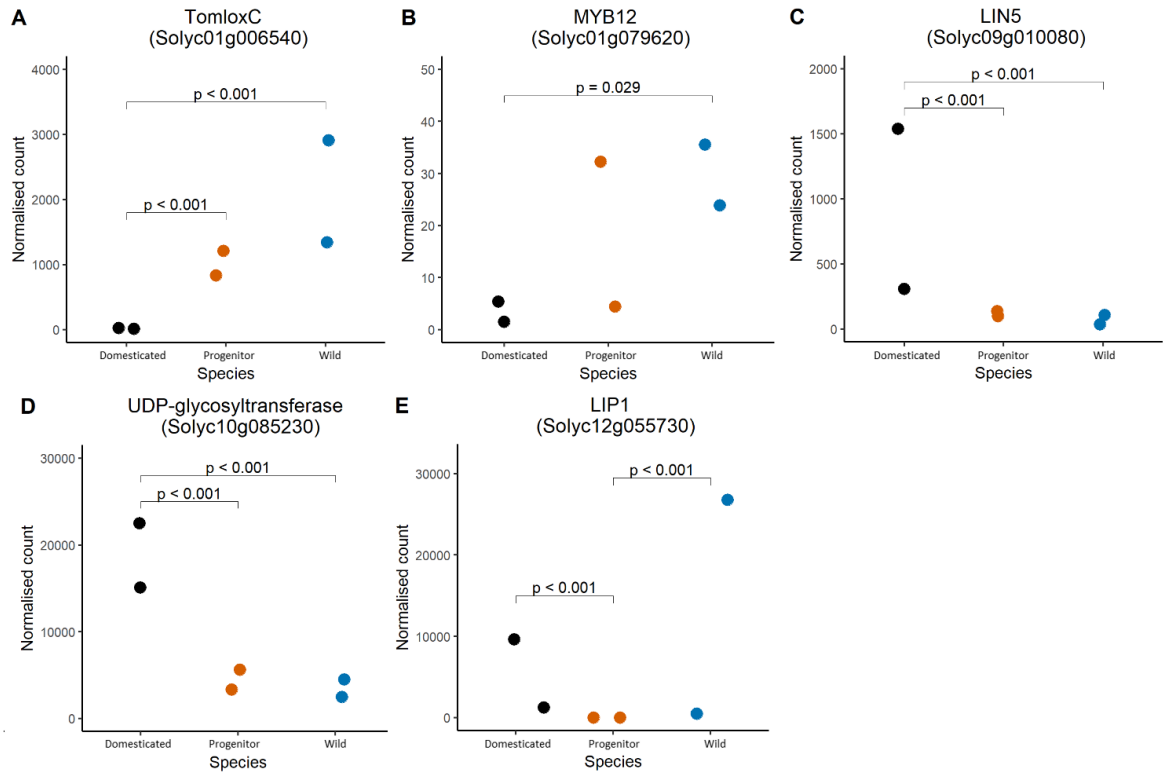


Figure 2.9: Domestication genes in the interspecific analysis.
Gene expressions based on normalised counts of (A) *TomloxC*, (B) *MYB12*, (C) *LIN5*, (D) *glycosyltransferase*, and (E) *LIP1* in the between domesticated, progenitor and wild species.

2.4.2.2 Plasticity analysis of gene expression

To examine plasticity, this dataset included three treatments for leaves, therefore resulting in a different set of DEGs to the interspecific comparison. Both interspecific divergence and

plasticity are considered here ('plasticity' dataset: 3 species x 3 treatments x 1 tissue). Principal component analysis (PCA) shows a distinction between domesticated, progenitor and wild species, with domesticated and progenitor clustering (Figure 2.10A), this parallels their genetic relatedness. The divergence between species (Table 2.7) revealed similar patterns to the results in the 2.4.2.1 'interspecific' dataset (full results on Table A14 to A18) therefore are only discussed in the Appendix Results A. 1.

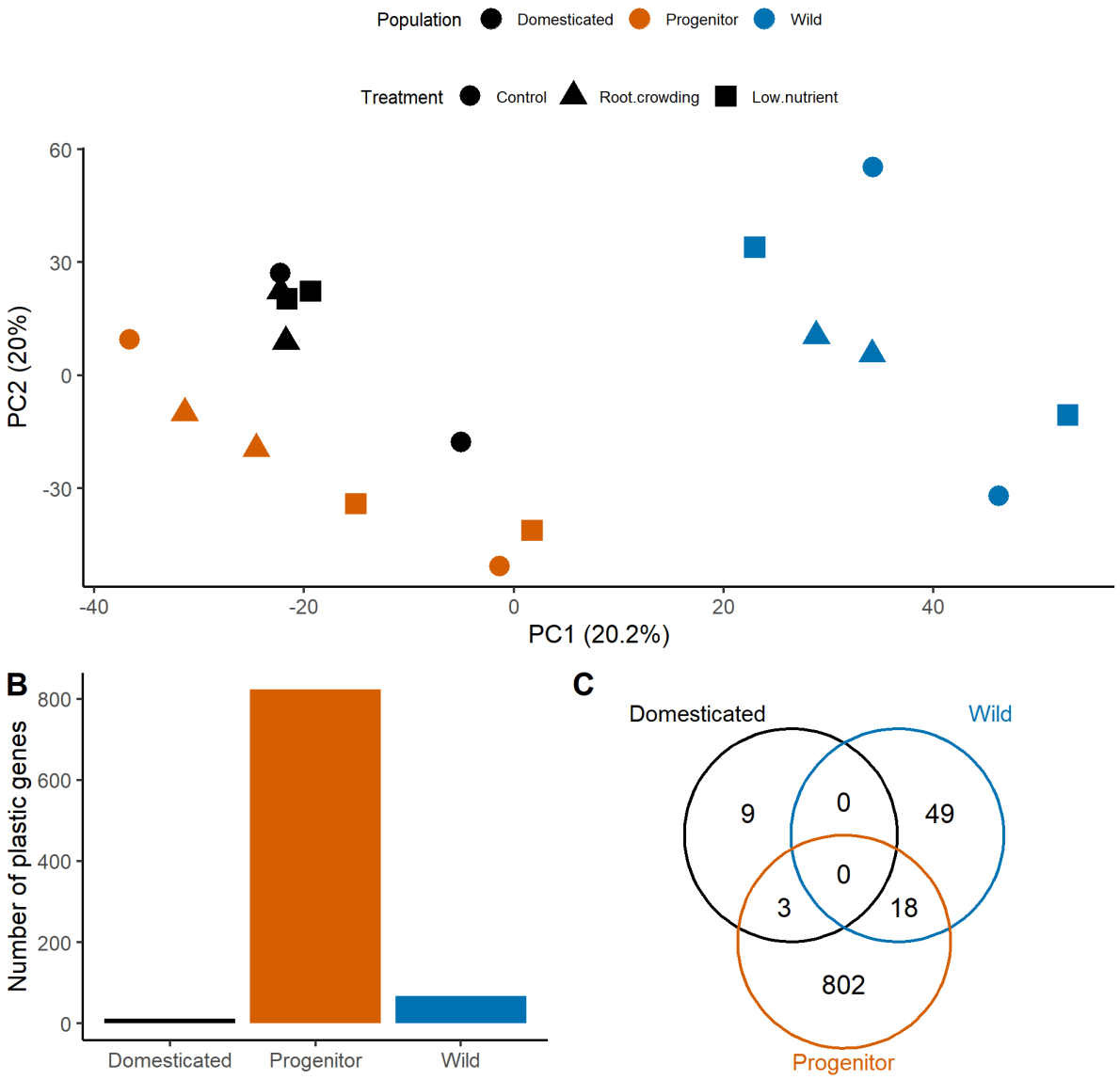


Figure 2.10: Gene expression analysis of the plasticity dataset.

Principal component analysis (first two principal components) of (A) leaf transcriptomes. Leaf transcriptomes were analysed for differentially expressed genes (DEGs) between treatments (i.e. plastic genes) within domesticated, progenitor and wild species. (B) Number of plastic genes in each species. (C) Venn diagram of shared plastic genes between species.

Table 2.7: DEGs and associated GO terms and KEGG pathways in the plasticity analysis. Differentially expressed genes (DEGs) in leaf between domesticated, progenitor and wild species and between treatments (i.e., plasticity) under control, root crowding and low nutrient treatments with their associated significant GO terms and KEGG pathways (full lists in Table A15 to A20).

Species	Tissue	DEGs		GO terms	KEGG pathways
Domesticated vs Progenitor	Leaf	256	(1.210%)	0	0
Domesticated vs Wild	Leaf	505	(2.387%)	16	1
Progenitor vs Wild	Leaf	732	(3.461%)	55	0
Plastic in Domesticated	Leaf	12	(0.057%)	0	0
Plastic in Progenitor	Leaf	823	(3.891%)	338	3
Plastic in Wild	Leaf	67	(0.317%)	42	0

Plastic genes were DEGs between at least one pair of treatments within each species. There was an order of magnitude more plastic genes in the progenitor than in the wild and the domesticated with 823 (3.891% of genes), 67 (0.317%) and 12 genes (0.057%), respectively (Table 2.7; Figure 2.10B; Table A18). At least 70% of the genes that were plastic in each species were species-specific (Figure 2.10C). There were no plastic genes that were shared between all three species, although 3 were shared between domesticated and progenitor species, and 18 between the progenitor and wild species (Figure 2.10C). These plastic genes resulted in 42, 338, and zero significant GO terms in the progenitor, wild and domesticated species respectively (Figure 2.11). There was no overlap in the top 10 GO terms identified in the plastic genes in the progenitor and wild species (Figure 2.11), however, there were 16 significant GO terms shared between progenitor and wild species (Table A19), including response to nutrient levels and response to starvation; however, Genes plastic in the domesticated and wild species were not significantly enriched for any KEGG pathways. Plastic genes in the progenitor species were enriched for three KEGG pathways: “Ribosome”, “RNA transport” and “Ubiquitin mediated proteolysis” (Table A20). Genes only plastic in the progenitor species resulted in five significant KEGG pathways, including the same three pathways mentioned above, as well as “Plant hormone signal transduction” and “Protein export”. Plasticity in “Plant hormone signal transduction” could have a link to various plant processes including plant growth, fruit ripening, stress response and disease resistance (Appendix Figure A. 7).

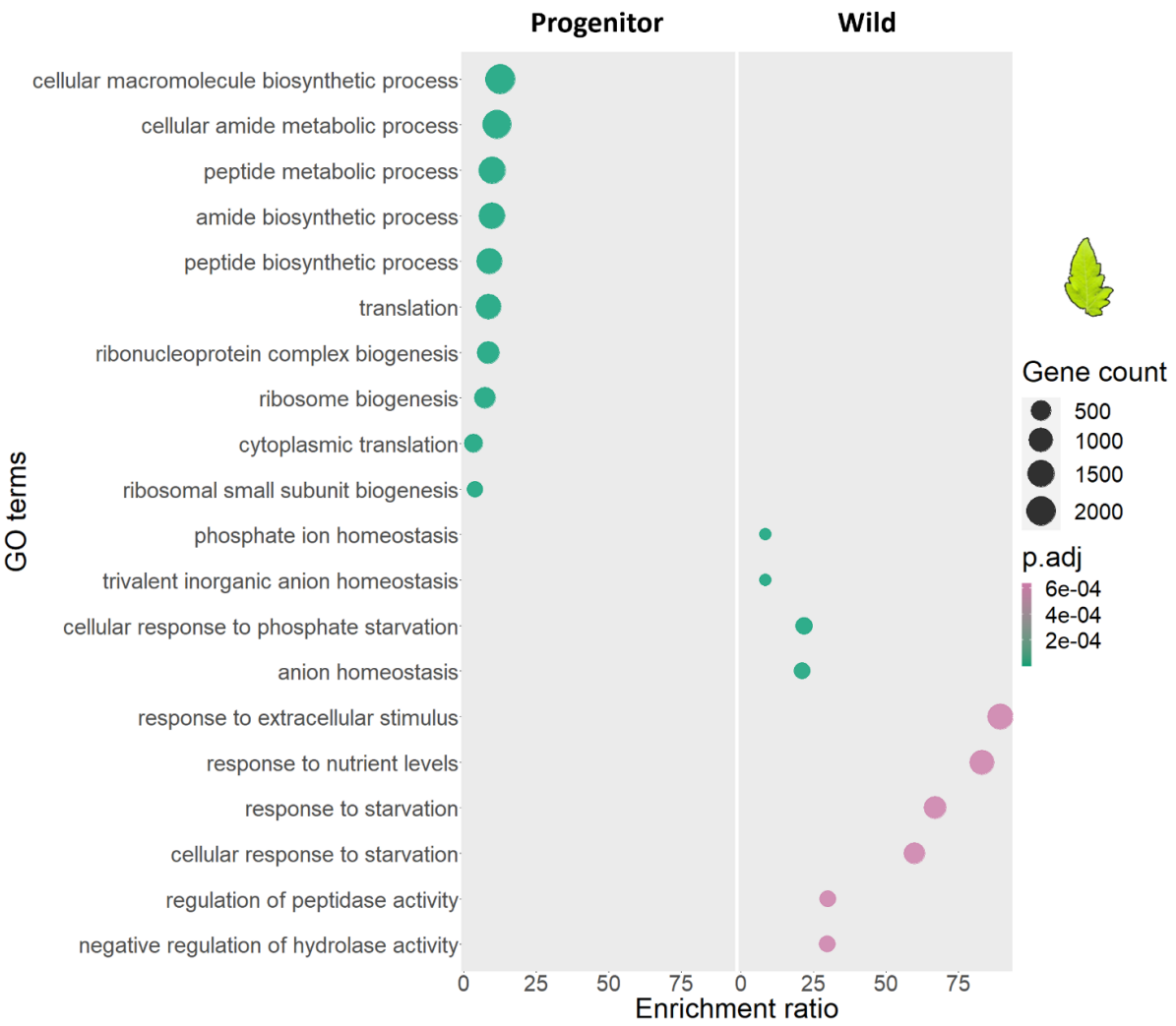


Figure 2.11: Gene Ontology (GO) terms in the plasticity analysis.

Top 10 significant terms from Gene Ontology (GO) analysis of plastic genes in progenitor and wild species, with enrichment ratio and adjusted p-value (p.adj) for each GO term. No significant GO terms were associated with plastic genes in the domesticated species.

2.4.2.3 Gene expression plasticity and divergence during domestication

To explore the overlap between plasticity and the divergence in gene expression during domestication, the 'plasticity' dataset for leaves was utilised along with a fruit dataset ('fruit' dataset: 2 species x 3 treatments x 1 tissue). Principal component analysis (PCA) of the fruit dataset shows a distinction between the domesticated and the progenitor species (Figure 2.12A) and there were 1655 DEGs between the domesticated and progenitor in fruits (Table 2.8). The divergence between species revealed similar patterns to the results in the other datasets (Table 2.8; Table A21 to A23; see Appendix Results A. 1 for full discussion).

The overlap in DEGs between the domesticated and progenitor and the genes plastic in the domesticated, the progenitor and the wild species were already discussed in the previous section (2.4.2.2 'plasticity' analysis). For fruit transcriptomes, more genes were plastic in the progenitor than in the domesticated species (Figure 2.12B), 26 and 73, respectively (Table 2.8; Table A24). No significant GO terms or KEGG pathways were identified apart from "Metabolic pathways" enriched in genes plastic in the progenitor species.

Table 2.8: DEGs and associated GO terms and KEGG pathways in the plasticity divergence analysis.

The overlap in differentially expressed genes (DEGs) between domesticated and progenitor species and genes that were plastic in domesticated, progenitor and wild species with their associated significant GO terms and KEGG pathways in leaves. For fruit, DEGs between domesticated and progenitor species and genes plastic in domesticated and progenitor species were identified with their associated significant GO terms and KEGG pathways (full lists in Table A21 to A26).

Description	Tissue	DEGs		GO terms	KEGG pathways
DE & plastic genes in the domesticated	Leaf	2	(0.009%)	0	0
DE & plastic genes in the progenitor	Leaf	38	(0.180%)	0	0
DE & plastic genes in the wild	Leaf	4	(0.019%)	0	0
Doemsticated vs Progenitor	Fruit	1655	(8.224%)	115	5
Plastic genes in the domesticated	Fruit	26	(0.129%)	0	0
Plastic genes in the progenitor	Fruit	73	(0.363%)	0	1
DE & plastic genes in the domesticated	Fruit	7	(0.035%)	0	0
DE & plastic genes in the progenitor	Fruit	30	(0.149%)	0	1

The overlap in DEGs between the domesticated and the progenitor and are plastic in the progenitor species could reveal whether plasticity was selected during tomato domestication (Table A25 and A26). In the domesticated species, two of the genes (0.78%) differentially expressed between the domesticated vs the progenitor species were plastic in leaves, more than expected by chance ($\chi^2 = 12.8$, df = 1, $p < 0.001$; Appendix Figure A. 8A). In the domesticated, 26 (1.57%) DEGs between the domesticated vs the progenitor were plastic in fruits, more than expected by chance (χ^2 9.7064, df = 1, p = 0.002; Appendix Figure A. 8B). For the progenitor, 38 (14.84%) DEGs between the domesticated vs the progenitor were plastic in leaves, more than expected by chance ($\chi^2 = 80.194$, df = 1, $p < 0.001$; Appendix Figure A. 8C). In the progenitor fruit, 30 (1.81%) DEGs between the domesticated vs the progenitor were plastic, more than expected by chance ($\chi^2 = 100.56$, df = 1, $p < 0.001$; Appendix Figure A. 8D). The plastic

and divergent genes were not overrepresented for any GO terms except for “Metabolic pathways”. Compared to the wild species, four (1.56%) DEGs between the domesticated vs the progenitor were plastic in leaves, again more than expected by chance ($\chi^2 = 9.0556$, df = 1, p = 0.003; Appendix Figure A. 8E). Overall, genes divergent between the progenitor and the domesticated species tended to be plastic within species. SP had the greatest overlap between genes divergent between the domesticated vs the progenitor and also plastic. The greatest contribution to χ^2 for all comparisons was the positive association between divergent and plastic genes (Appendix Figure A. 8).

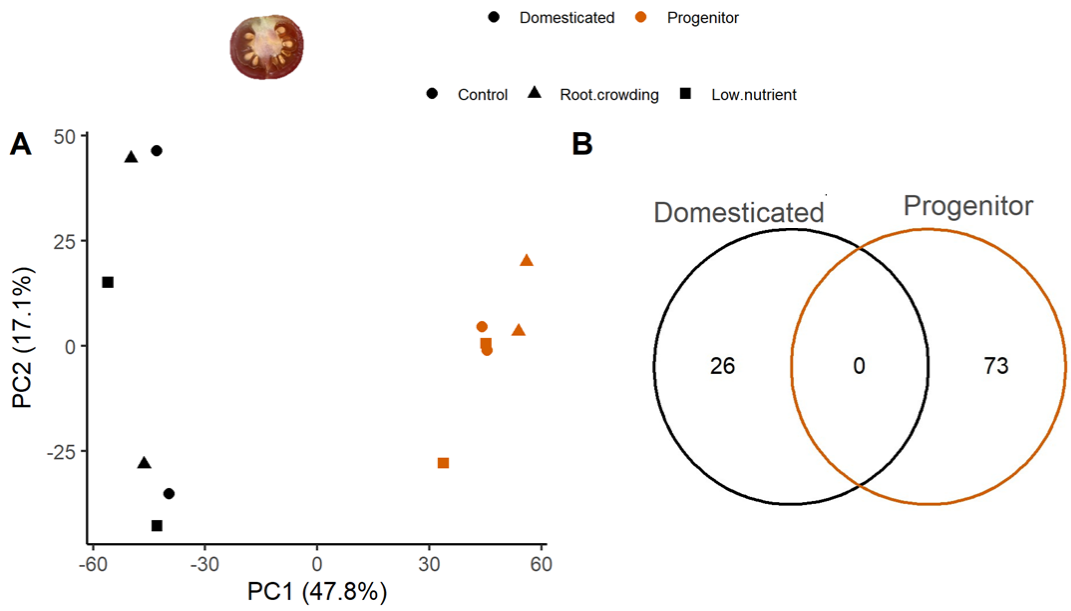


Figure 2.12: Plasticity divergence analysis in fruit.

For fruits, divergent genes between domesticated and progenitor species and plasticity in domesticated and progenitor species were revealed. (A) Principal component analysis (first two principal components) of control, root crowding and low nutrient treatments for fruit transcriptomes. (B) Venn diagram of differentially expressed genes shared.

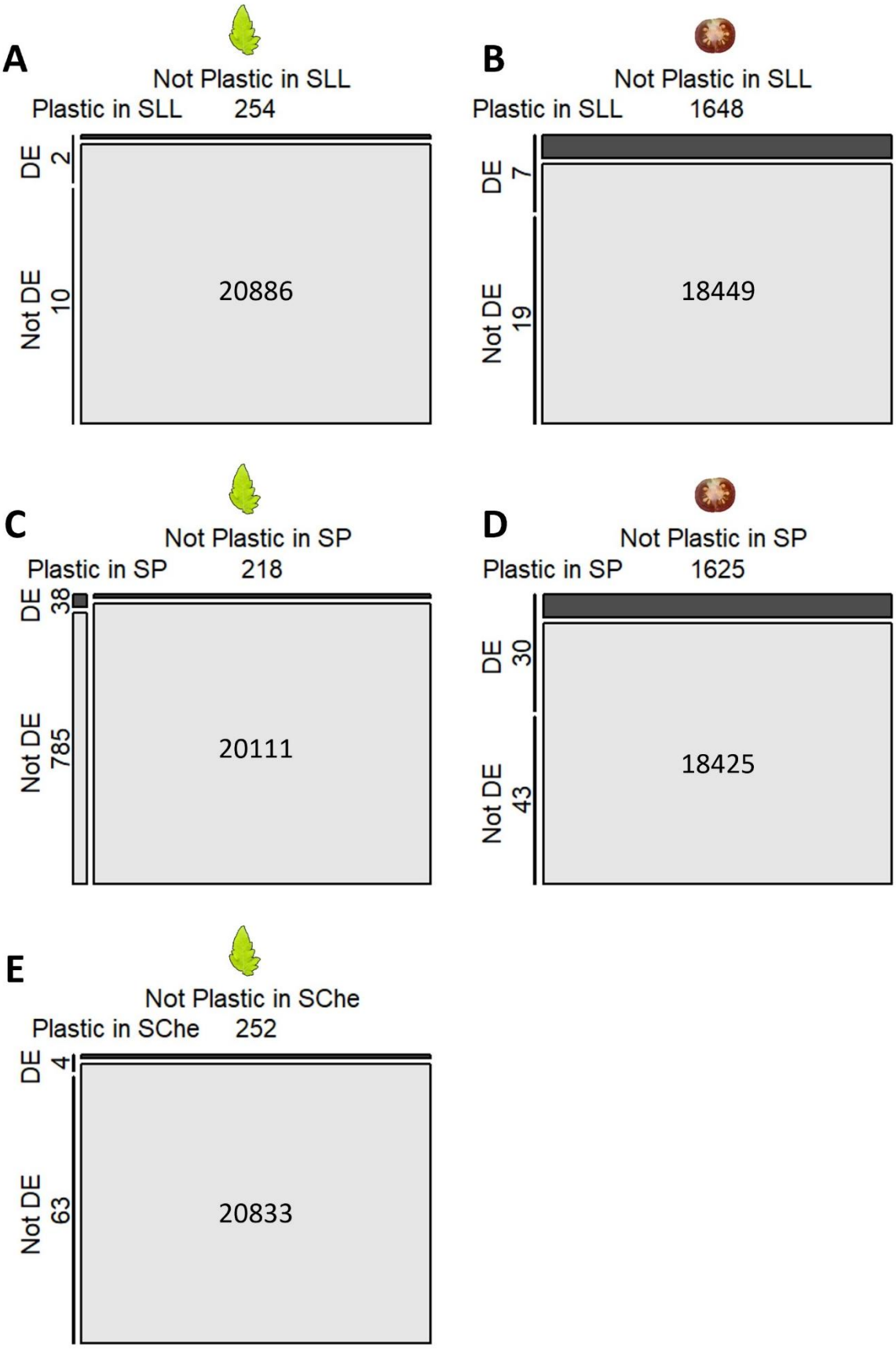


Figure 2.13: Mosaic plots in the plasticity divergence analysis.

Genes differentially expressed (DE) between the domesticated *S. lycopersicum* (SLL) and the progenitor *S. pimpinellifolium* (SP) and/or plastic in SLL (A) leaves and (B) fruits; SP (C) leaves and (D) fruits; and (E) the wild *S. cheesmaniae* (SChe) leaves were identified.

We reanalysed the gene expression data for the domestication genes (Table 2.2) to identify those differentially expressed between the domesticated and the progenitor and/or plastic in each species (i.e. the wild species was excluded because there were no fruits in two treatments). Four of the domestication genes, *TomloxC* (Figure 2.14A), *MYB12* (Figure 2.14B), *LIN5* (Figure 2.14C), and UDP-glycosyltransferase (Figure 2.14D), were plastic in the progenitor and four were differentially expressed between the domesticated and the progenitor: *TomloxC* (Figure 2.14E), *LIN5* (Figure 2.14F), UDP-glycosyltransferase (Figure 2.14G), and *LIP1* (Figure 2.14H), i.e. three overlapped. This could suggest that for these three genes, plasticity in the progenitor was selected and gave rise to a gene expression divergence during domestication, however, there was no detectable association between plastic and divergent traits (Fisher's Exact test: $p = 0.088$).

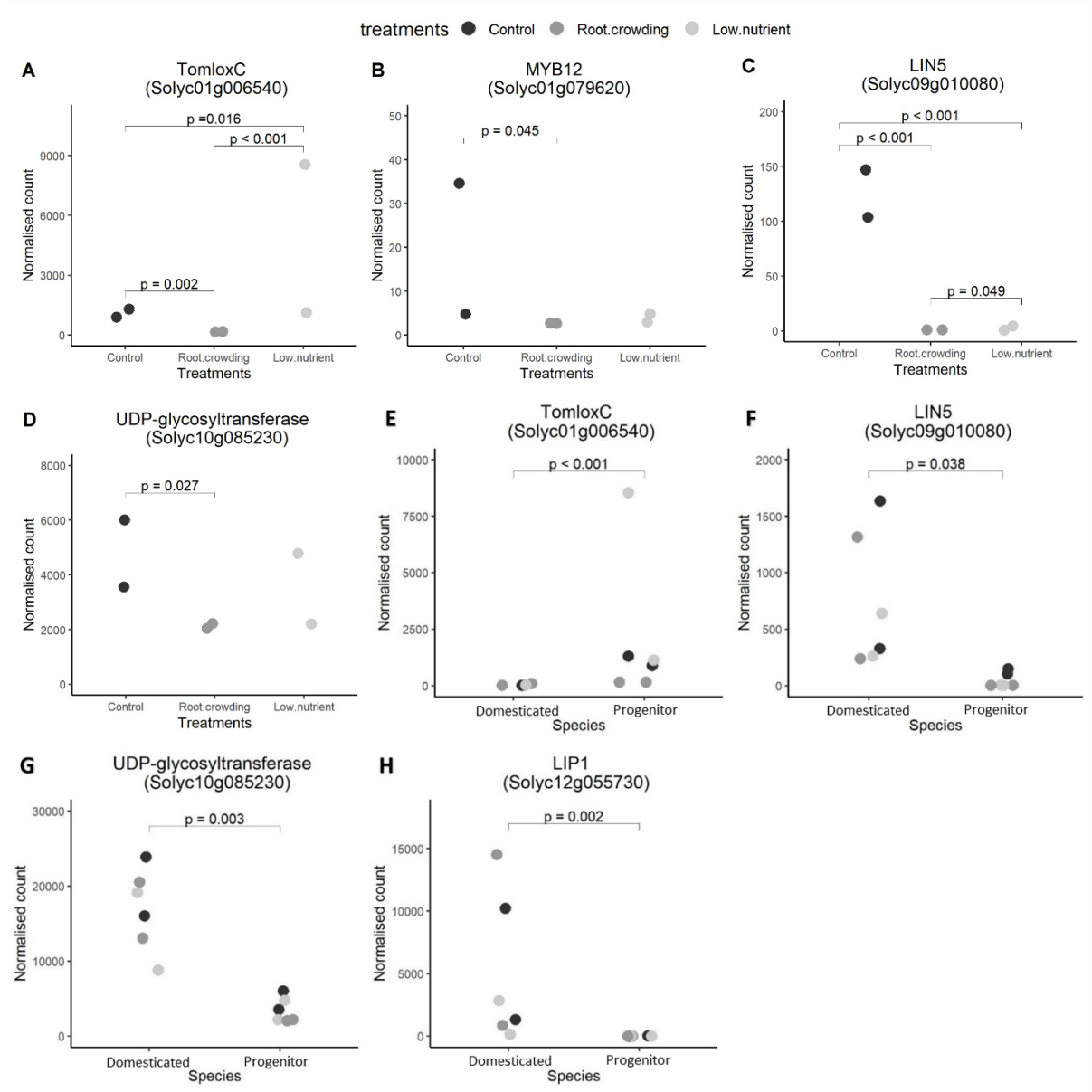


Figure 2.14: Domestication genes in the plasticity divergence analysis.

A-D, Fruit domestication genes plastic in the progenitor species: (A) *TomloxC*, (B) *MYB12,* (C) *LIN5*, and (D) UDP-glycosyltransferase. E-H, genes differentially expressed between the progenitor and the domesticated species: (E) *TomloxC*, (F) *LIN5*, (G) UDP-glycosyltransferase, and (H) *LIP1* genes.

2.5 Discussion

Our results support our hypothesis, and the following findings can be drawn from the study: (i) there are significant morphological and gene expression differences between species; (ii) there is a greater number of plastic traits in the tomato progenitor compared to the never-domesticated wild species; (iii) our data is consistent with plasticity promoting gene expression divergence during tomato domestication. This expands our understanding of why some species

were domesticated over others in early domestication as well as the role of plasticity in tomato domestication.

2.5.1 Divergence between species

2.5.1.1 Divergence between the progenitor and the never domesticated wild species.

The tomato progenitor, *S. pimpinellifolium*, had earlier flowering, greater fruit yield and larger seeds compared to the never-domesticated wilds, *S. cheesmaniae* and *S. chmielewskii* (Figure 2.3). Differences in timings of developmental stages between wild crop relatives have been observed in cereal crop progenitors with faster germination compared to other wild grasses (Cunniff et al., 2014; Preece et al., 2021). Higher fruit yield in progenitors than in never-domesticated wilds may have been more attractive and benefited humans in early domestication. However, yield comparisons in other crops such as cereal and pulse crops suggest that progenitors did not consistently have higher yields compared to other wilds (Preece *et al.*, 2018; Preece *et al.*, 2015). Larger seed sizes in progenitors may have been advantageous for seed saving in early domestication as this offers greater apparency, valuable for crops domesticated for their seeds such as cereal crops where progenitors have larger seed size compared to other wild crop relatives (Cunniff et al., 2014; Preece et al., 2018; Preece et al., 2021). These traits may have given the progenitor an advantage in early domestication.

Additionally, expression divergence in leaves between the progenitor (*S. pimpinellifolium*) and a never-domesticated wild (*S. cheesmaniae*) highlighted possible differences in stress response. Many genes related to biotic environment and defense response (Figure 2.8) were differentially expressed between the progenitor and the wild species, such as methylesterases genes (Wen *et al* 2020). Wild tomato species produce volatiles in their trichomes (Kortbeek *et al.*, 2021) and differences in these traits may contribute to the differences in defensive responses. For example, *S. pimpinellifolium* has trichome-based resistance against whiteflies with its type-IV trichomes (Escobar-Bravo *et al.*, 2016), but *S. cheesmaniae* lacks type-IV trichomes (Vosman *et al.*, 2018). Defense responses could have been advantageous in early domestication as changes in growing conditions would have been coupled with changes in herbivory and potential diseases.

In fruits, expression divergence between the progenitor and the wild included many genes involved in cell wall metabolic processes, biosynthesis of secondary metabolites (SM), and various sugar metabolism (Figure 2.8). *S. pimpinellifolium* and *S. cheesmaniae* have been

shown to differ in their SM (Iijima *et al.*, 2013; Schwahn *et al.*, 2014) and cuticle morphology and cutin chemical composition (Yeats *et al.*, 2012). Fruit-related domestication gene *LIP1* had significantly greater expression in the wild than the progenitor; this is associated with diacylglycerol and triacylglycerol degradation, resulting in the release of fatty acids that serve as precursors of flavour volatiles (Garbowicz *et al.*, 2018). This suggests that *S. cheesmaniae* may have more of these associated *LIP1* flavour volatiles than *S. pimpinellifolium*.

2.5.1.2 Divergence during domestication

Comparing the progenitor (*S. pimpinellifolium*) and the domesticated tomato (*S. lycopersicum*) revealed a reduction in height (at flowering and fruiting) and an increase in fruit weight, fruit size (fruit area, width, pericarp area and pericarp thickness) and seed size during domestication (Figure 2.3). These are aspects of the domestication syndrome that have been previously reported for tomatoes (Bai and Lindhout, 2007). Reduction in plant height during domestication has been linked to increased yield, damage prevention from wind and rain, and favourable plant architecture for harvest (Lenser and Theißen, 2013). However, studies on various crops have indicated no difference or an increase in plant height during domestication (Milla *et al.*, 2014; Preece *et al.*, 2017; Chacón-Labella *et al.*, 2019). Domestication genes in tomatoes have been reported such as *FAS*, *FW2.2* and *LC* for fruit size and *OVATE* and *SUN* for fruit shape (Bai and Lindhout, 2007; Meyer and Purugganan, 2013). There is typically an increase in seed size in domesticated plants compared to their progenitors (Doganlar *et al.*, 2000; Gómez-Fernández and Milla, 2022), this could be due to indirect selection for greater seedling vigour and germination uniformity under cultivation (Basu and Groot, 2023). Changes in these traits indicate selection for favoured traits advantageous in disturbed, fertile fields found in early domestication (Milla *et al.*, 2018).

Furthermore, domestication also influenced changes in gene expression depicted by the over-representation of genes involved in biotic stimulus was detected in differentially expressed genes and abiotic stresses in fruit samples, for DEGs between the domesticated and the progenitor species (Figure 2.8). This is consistent with Koenig *et al.* (2013)'s assessment of the divergence in gene expression of seedling tissues between *S. lycopersicum* and *S. pimpinellifolium*, which revealed overrepresentation of genes involved in both biotic and abiotic stresses such as defense response, stress response, photosynthesis and response to high light. The distinct growing conditions of these species imply different pests, herbivores and human interactions, as well as differential defensive response to biotic stresses. Paudel *et al.* (2019) illustrated a differential defense mechanism between *S. lycopersicum*, *S. lycopersicum* var.

cerasiforme and *S. pimpinellifolium*, with different volatile organic compound profiles and differential preference of herbivore moth, suggestive of domestication altering how cultivated tomato interacts with its biotic environment. Our analysis supports this with the down-regulation of genes in *S. lycopersicum* (relative to *S. pimpinellifolium*) associated with the GO terms glycosyltransferases, terpene synthases, and production of secondary metabolites, all involved in defense mechanisms against biotic stress (Campos *et al.*, 2019; Wang *et al.*, 2021b). Similarly, Sauvage *et al.* (2017) identified biological processes GO terms that were overrepresented for DEGs between *S. lycopersicum* and *S. pimpinellifolium* such as metabolic process, carbohydrate metabolic process, lipid metabolic process, secondary metabolite process, and response to stress. Some pathways were overrepresented in DEG between *S. lycopersicum* and *S. pimpinellifolium*: biosynthesis of secondary metabolites (Tohge *et al.*, 2020), carbon metabolism (Luo *et al.*, 2020), carotenoid biosynthesis (Karniel *et al.*, 2022). Domestication may have altered these processes that are linked to the strong selection on tomato fruits (Liu *et al.*, 2020a).

Our survey of domestication genes found a reduction in gene expression in *TomloxC* and an increase in *LIP1*, UDP-glycosyltransferase and *LIN5* associated with domestication (Figure 2.9). *TomloxC* has a role in apocarotenoid production contributing to desirable tomato flavour and its reduction in expression in the domesticated species compared to the progenitor has been reported previously (Gao *et al.*, 2019). *LIP1* expression correlates with levels of fatty acids as precursors of flavour volatiles (Garbowicz *et al.*, 2018). Similarly, *LIN5* is associated with glucose and fructose content, which underwent strong selection due to the negative correlation between sugar levels and fruit size (Tieman *et al.*, 2017). The gene coding for UDP-glycosyltransferase is linked to steroidal glycoalkaloid (SGA) associated with bitter chemicals; reduction in gene expression in *S. lycopersicum* relative to *S. lycopersicum* var. *cerasiforme* suggests a selection for non-bitter alleles (Zhu *et al.*, 2018). Contrary to our results, Liu *et al.* (2020a) assessed the tomato fruit transcriptome at the orange stage and found the downregulation of *LIN5* in *S. lycopersicum* relative to *S. pimpinellifolium*. The inconsistency with increased expression of *LIP1*, UDP-glycosyltransferase and *LIN5*, contrary to previous studies, may be due to multiple factors. Different sampling strategies with different accessions and replicates can affect the genetic and metabolic variability captured. Sampling of tomato fruits at different stages of ripeness also greatly affects transcriptome profiles (Shinozaki *et al.*, 2018; Pereira *et al.*, 2021). These factors would also affect comparisons of DEGs identified in different studies (e.g. Koenig *et al.*, 2013; Sauvage *et al.*, 2017).

2.5.2 Greater plasticity in the tomato progenitor over never-domesticated wild

New and beneficial phenotypes could have arisen in the progenitor through phenotypic and gene expression plasticity that led to its continued cultivation and domestication at the expense of other species. Many studies have explored plasticity divergence between domesticated crops and their progenitors (Fréville *et al.*, 2022). To the best of our knowledge, this is the first study to assess the plasticity of a crop progenitor and a never-domesticated wild relative.

There was a greater number of plastic traits and genes in the progenitor compared to the never-domesticated wild species (Figure 2.5). Notable traits included height at flowering, fruit weight and fruit size (fruit area, width, pericarp area and pericarp thickness; Figure 2.6). Plasticity in height at flowering could allow humans to select for optimal height for harvest or optimal light capture, whilst plasticity in fruit weight and size could allow humans to select attractive fruit traits during domestication. Cryptic phenotypic variation can be uncovered with plasticity, which can be favourable in early domestication (Piperno, 2017). For example, the maize progenitor, teosinte, under conditions reflective of ancestral climate during its domestication (lower temperature and atmospheric CO_2), and revealed maize-like traits such as shorter plant height, a single main stalk with fewer, shorter branches and synchronous seed maturation (Piperno *et al.*, 2015; Piperno *et al.*, 2019). Similarly, erect knotweed grown in low density shifted from a smaller, less branched phenotype to a highly branched shrub with more seeds, illustrating a small change in cultivation practice increasing yield within a single growing season (Mueller *et al.*, 2023).

In addition to measuring phenotypes, we surveyed gene expression which did not require *a priori* choice of traits that could have biased our findings. The progenitor had more genes that were plastic compared to the wild species (Figure 2.10). Notably, there was no overlap in the top 10 significant GO terms associated with genes plastic in these species. Many of the top GO terms in the progenitor relate to biosynthetic or metabolic processes compared to the wild with regulation of different biological processes or responses to various stimuli (Figure 2.11). However, many of the significant GO terms in the wild were also significant in the progenitor. This suggests that the progenitor was plastic in diverse biological processes than the wild species.

As stated in the differentially expressed genes between species we found domesticated transcriptomes had evolved during domestication to be more wild-like (Figure 2.7). This could be due to both *S. lycopersicum* and *S. cheesmaniae* experiencing separate genetic bottlenecks, with the domestication of *S. lycopersicum* (Li et al., 2023) and island colonisation of the Galapagos islands and recent adaptation in *S. cheesmaniae* (Nuez et al., 2004; Pailles et al.,

2017). The genetic bottleneck in *S. cheesmaniae* may have affected their plasticity as reversion in plasticity has been reported after the colonisation of a new environment (She et al., 2024). This highlights the importance of further investigations with additional never-domesticated species. She et al. (2024) also report evidence of greater plasticity in the initial stage of adaptation, contributing to the high-elevation colonisation of Eurasian Tree Sparrows. A similar scenario may have occurred in the early domestication of tomatoes, where greater plasticity in the progenitor tomato at the initial stage of adaptation contributed to the adaptation in domesticated landscapes. Greater plasticity in the wild tomato species, especially in the progenitor may be linked to genetic diversity in general. Reduced genetic variation has been reported to limit plastic and adaptive potential (Chevin and Lande, 2010; Murren et al., 2015). Therefore, genetic diversity is another factor that may have aided the domestication of the progenitor over never-domesticated wild species (see Chapter 3).

In maize, the phenotypic plasticity of teosinte was coupled with high gene expression plasticity (Piperno *et al.*, 2015). Genes differentially expressed between growing conditions included phytohormone genes such as auxin and gibberellins, known to interact with *teosinte branched 1* (*tb1*), a major quantitative trait locus (QTL) important in plant architecture in maize domestication (Piperno *et al.*, 2015). This highlights the greater plasticity of teosinte compared to maize which may have made them easier to cultivate. Similarly, in our study, genes only plastic in the tomato progenitor were significantly over-represented in the plant hormone signal transduction pathway. Genes plastic in this pathway include those related to auxin, related to plant growth (Yu *et al.*, 2022); gibberellin linked to stem growth (Davière and Achard, 2013); abscisic acid linked to stomatal closure (Sah *et al.*, 2016); ethylene linked to senescence (Iqbal *et al.*, 2017); brassinosteroids linked to cell elongation/division (Zhu *et al.*, 2013); jasmonic acid linked to stress response (Ali and Baek, 2020) and salicylic acid linked to disease resistance (Kumar, 2014). Many of these plant hormones also play a role in stress response signalling (Bari and Jones, 2009). Although plasticity may not always result in adaptive advantages (Palacio-López *et al.*, 2015; Acasuso-Rivero *et al.*, 2018), it can act as a buffer that promotes stability and allows the species time to adapt to changing environments (Chevin and Lande, 2010) that could have been important in early domestication.

2.5.3 Reduction in plasticity during domestication

Morphological traits that were plastic in the progenitor and divergent between the progenitor and the domesticates, including several related to height fruit weight and size (Table 2.5). Although the loss in phenotypic plasticity in many traits and taxa has been recorded (Schwander

and Leimar, 2011), we only detected loss in plasticity for fruit morphology traits (fruit area, width, pericarp area and thickness) in the domesticates compared to the progenitor. Loss in plasticity in fruit size would have benefited farmers with consistent yield each harvest. The maintenance of plasticity in other traits has also been reported by Matesanz and Milla (2018) which found no generalised loss in plasticity with varying levels of water and nutrient supplies in a diverse set of crops. This suggests that some domesticates retain their phenotypic plasticity despite strong selection pressures during domestication (Ménard *et al.*, 2013; Sadras *et al.*, 2016; Marques *et al.*, 2020).

Contrary to the phenotypic data, gene expression data for the progenitor and the domesticates, suggest a substantial reduction from 3.89% of genes plastic in the progenitor to only 0.06% in the domesticated species (Figure 2.10). Piperno *et al.* (2019) demonstrated that teosinte had greater phenotypic plasticity than maize, and the same pattern was found for gene expression data, indicative of genetic assimilation during maize domestication (Lorant *et al.*, 2017). A signature of genetic assimilation may be difficult to detect in extant populations as plasticity can evolve relatively quickly (Pigliucci and Murren, 2003). However, the evolution of many traits and their plasticity is compatible with the loss in plasticity characteristic of genetic assimilation (Diggle and Miller, 2013; Belcher *et al.*, 2023; Mueller *et al.*, 2023). Selection on cryptic phenotypes can lead to adaptive divergence. The establishment of favourable phenotypes under a stable environment can result in a loss of plasticity as plasticity under stable conditions does not increase fitness (Pfennig, 2021). We assume that stability is characteristic of cultivated fields where farmers ensure the predictability of resources and limit stresses. Selection acting strongly to decrease plasticity (Ghalambor *et al.*, 2015), can result in genetic assimilation of gene expression levels over multiple generations (West-Eberhard, 2003).

2.5.4 Plasticity can promote gene expression divergence

We hypothesised that if plasticity played a role in determining which species became domesticated, at the expense of others, the traits (Table 2.5) and genes (Table 2.8) that differed between the progenitor and the domesticated species would more likely to be plastic in the progenitor than expected by chance. There was not enough evidence to suggest that plastic traits were selected for or against during domestication, possibly due to the small number of phenotypic traits measured. On the other hand, divergent genes during domestication were more likely to be plastic in domesticated, progenitor and wild species; suggestive of the presence of plasticity across species, but there was overall a greater number of plastic genes in the progenitor. This supports our hypothesis that gene expression divergence during

domestication was more likely for genes which were plastic in the progenitor than genes that were not plastic. Similar results have been reported in various organisms (She et al., 2024; Corl et al., 2018; Campbell-Staton et al., 2021). Furthermore, we found four known domestication genes related to tomato flavour that were plastic in *S. pimpinellifolium* and divergent between the progenitor and the domesticated (Figure 2.14): *TomloxC* (Gao *et al.*, 2019), *LIN5* (Tieman *et al.*, 2017), UDP-glycosyltransferase (Zhu *et al.*, 2018) and *LIP1* (Garbowicz *et al.*, 2018). All four known domestication genes related to tomato flavour and quality, have been extensively studied (Bai and Lindhout, 2007; Kaur *et al.*, 2023).

Although our study does not explore adaptive and non-adaptive plasticity, the initial selection on non-adaptive variation is a possible source of adaptive traits, with cryptic genetic variation resulting in a phenotype that increases fitness (Ghalambor *et al.*, 2007). These plastic traits and genes that diverged during domestication have the potential to benefit the tomato progenitor in early domestication, as plasticity can allow them to thrive in rapidly changing environments (Milla *et al.*, 2018). Rapid response through plasticity, for example, the production of larger fruits, may encourage early cultivators to propagate these individuals. Therefore, plasticity could facilitate rapid expression evolution (She *et al.*, 2024). However, plastic genes may be more likely to diverge due to genetic drift (Pfennig *et al.*, 2010; Seymour *et al.*, 2019). In addition, highly expressed genes are under stronger selection pressures, as high mutation rates are observed in highly transcribed genes (Duret, 2002; Sharp *et al.*, 2010; Park *et al.*, 2012), meaning detection of their expression differences would be easier, suggesting that we might be enriching for highly expressed genes under these different selective regimes.

2.5.5 Limitations

Our gene expression analysis was limited to *S. lycopersicum*, *S. pimpinellifolium* and *S. cheesmaniae* and mapped to a single reference. We therefore are unable to identify novel transcripts in any of the species or accessions and therefore our analysis only focuses on quantification of genes present in the reference (Conesa *et al.*, 2016). The emergence of pan-genomes may aid the exploration into the genetic variation across domesticates, progenitors and other never-domesticated wild tomato species (Li *et al.*, 2023). Gene expression is dynamic and can change through time influenced by developmental stage, environment, and epigenetics (Rivera *et al.*, 2021). Therefore, transcriptomes from multiple time points could broaden our understanding of how gene expression plasticity may vary. Comparison between progenitors and never-domesticated species of other tissues such as fruit and root would also give a comprehensive view of each species' ability to be plastic. The limited number of phenotypic

traits had the power to bias our results. To avoid this, adding additional phenotypic traits would improve our assessment of plasticity, especially as previous studies have indicated pre-adaptation to cultivation in root morphology of crop progenitors (Martín-Robles *et al.*, 2018).

We acknowledge that the populations of tomato crop relatives used, *S. pimpinellifolium*, *S. cheesmaniae* and *S. chmielewskii*, are not identical to those that existed around the time of early domestication due to subsequent selection and gene flow in the wild (Flint-Garcia *et al.*, 2023), yet assessment of plasticity can only be performed with extant populations. Plasticity may have aided the domestication of *S. pimpinellifolium*, however, due to the geographical location of *S. cheesmaniae* in the Galapagos Islands, these two species would not have competed in early domestication. Therefore, exploration of other never-domesticated species would broaden our understanding of plasticity differences between these groups. Furthermore, the difference in plasticity detected in these species may be due to variation within species. The limited number of accessions within each species may fail to capture representative variation in plasticity. Greater plasticity in the progenitor may be due to the greater variation in plasticity between progenitor accessions. However, this could still indicate greater variability in traits and gene expression in the progenitor that humans could select from. This variation may have been important in early domestication, allowing humans to select for desirable traits such as larger fruits.

2.6 Conclusion

Plasticity may have played an important role in the early domestication of tomatoes. We showed differential plasticity between domesticates, progenitor and never-domesticated wild tomato species to different growing conditions broadly reflective of the changes in the environment in early domestication and that genes that have diverged in expression during domestication tended to be those which were plastic in the progenitor. Greater number of plastic traits and genes in the tomato progenitor could have given *S. pimpinellifolium* a selective advantage in the early stages of domestication. This is the first time phenotypic and expression plasticity has been characterised between a crop progenitor and never-domesticated relatives. Plasticity with adaptive value could be important in crop breeding programmes to address crop responses to current and future environmental unpredictability.

Chapter 3 The role of transposable elements (TEs) on tomato domestication

3.1 Abstract

Crop domestication is an evolutionary process transforming wild progenitors into cultivated crops. Transposable elements (TEs) are major drivers of plant evolution since they are potent mutagens; they may therefore have had a major impact on the evolution of domesticated plants. We explore the role of genetic variation in the domestication of the tomato progenitor, as this is positively correlated with mutation rates that can facilitate adaptation in early domestication. Single nucleotide polymorphisms (SNPs) and transposon insertion polymorphisms (TIPs) were annotated and characterised in domesticated, progenitor and never-domesticated wild tomato species. We found SNP and TIP diversity across the genome to be higher in the progenitor than never-domesticated wild tomatoes. This may suggest greater maintenance of genetic diversity in the progenitor species through higher mutation rates. SNP and TIP diversity near and within genic regions were also higher in the progenitor than the never-domesticated wilds, indicative of important variants that may contribute to phenotypic variation important for adaptation in early domestication. Genes in proximity to TIPs were enriched in biological processes related to response in stimulus and terpenoid metabolic process. There were also changes in TIP frequency between the progenitor and the domesticates, suggestive of the role of TEs in the domestication process. These TIPs under selection during domestication were located in or near genes involved in plant defence and stress response. TEs are a great source of genetic variation that could have been selected upon, aiding the early adaptation of tomato progenitors to a cultivated environment and their subsequent domestication.

3.2 Introduction

Crop domestication is an evolutionary process that allows humans to cultivate plants for food and other resources. This created a mutualistic relationship that started more than 12,000 years ago (Larson *et al.*, 2014). Selection during domestication occurred in two stages: selection between wild species, and then the selection that transformed the crop progenitor into domesticated crops (Jones *et al.*, 2021). A lot is known about the latter, where early domestication of crops is driven predominantly by natural selection, resulting in domestication as a protracted process (Purugganan, 2019). However, little is known about the initial selective process that underlies the selection between wild species.

Research on crop wild progenitor and other wild species that were not domesticated is few and often limited to phenotypic assessments, highlighting the progenitor's competitiveness in a cultivated environment (Cunniff et al., 2014; Preece et al., 2015; Preece et al., 2018; Martín-Robles et al., 2018). Changes in growing conditions in early domestication would have acted as a selection pressure, with wild species that can adapt to these changing conditions more likely to be domesticated. Adaptation to new environments can occur through standing genetic variation or new genetic variation. A main source of genetic diversity or genetic polymorphism are mutations, which are changes in the DNA sequence that can alter proteins and subsequently their functions (Peischl and Kirkpatrick, 2012). In eukaryotes, nucleotide diversity is found to be positively correlated with mutation rates (Wang and Obbard, 2023). The contributions of mutations to adaptation during early domestication between wild species are currently unknown.

A potent source of genetic diversity is transposable elements (TEs) which are mobile DNA sequences able to move from one location of the genome to another (Bourque *et al.*, 2018). TEs can have a variety of selective effects, but they are generally thought to be deleterious (Lee, 2022). Integration of TEs into a genome can be harmful to the hosts' genome, directly through altering gene functions such as disrupting the coding sequence leading to loss of function or indirectly by disruption in promoter or enhancer regions, changing gene expression (Hirsch and Springer, 2017). DNA methylation suppresses TE expression as a way of TE repression, but this can result in the downregulation of nearby genes (Hollister and Gaut, 2009). TEs have also been known to introduce new regulatory elements upon their insertion (Chuong *et al.*, 2017). Moreover, TEs can induce ectopic recombination which can lead to the duplication and deletion of genomic regions (Almojil *et al.*, 2021). However, these genomic disruptions have also given rise to advantages, because TEs generate significant genomic changes, which can drive the development of phenotypic diversity which can enable populations to adapt to changing environments (Oliver and Greene, 2012; Ramakrishnan *et al.*, 2021).

Deleterious mutations that reduce fitness under one condition can become advantageous in a different environment (Dwivedi *et al.*, 2023), such as the *sh1* loss of function mutation was vital in the domestication of maize, rice and sorghum (Lin *et al.*, 2012). This could mean that neutral or deleterious genetic variation in natural conditions could become beneficial in cultivated fields, providing already available allelic variation for selection. Important agronomic traits have been associated with TEs in several major crops (Table 1.3). This highlights the importance of TE in domestication genes and the identification of polymorphisms which contribute to phenotypic variation and can be utilised for crop improvement (Vitte *et al.*, 2014).

The degree to which TEs contribute to the overall level of phenotypic variation and adaptation is still unknown. Early anthropogenic environments are thought to be disturbed by events such as seasonal fires and flooding (Wood and Lenné, 2018). Evidence of increased TE activity in stress conditions has been documented for several TE families (Kimura *et al.*, 2001; Salazar *et al.*, 2007; Woodrow *et al.*, 2011; Ito *et al.*, 2013; Roquis *et al.*, 2021). Activation of TEs can trigger random genetic variation that is vital for adaptation through natural selection (Schrader and Schmitz, 2019). Therefore, an increase in TE activity could mean more novel phenotypes arise in certain species, promoting their domestication through the accumulation of beneficial mutations at the expense of others.

To detect TE activity, one method is identifying polymorphic insertions between individuals known as transposon insertion polymorphisms (TIPs). Annotation of TIPs has been done in a number of crops such as rice (Castanera *et al.*, 2021), tomato (Dominguez *et al.*, 2020), *Brassica rapa*, and *B. oleracea* crops (Sampath *et al.*, 2014). These reflect insertions occurring after divergence from a common ancestor or the deletion of TEs that were once fixed in the population (Huang *et al.*, 2012; Kelleher *et al.*, 2020). For example, the presence of TIPs among closely related individuals and its absence in other populations indicates that the TE may have been recently or is currently active (Figure 3.1A). These recent transpositions of TEs will only be found in a few individuals in the population and are characterised as low-frequency TIPs. Increase in TE activity can drive greater TE diversity (Osmanski *et al.*, 2023), with reports of various crops experiencing bursts of TE integration (Lu *et al.*, 2012; Diez *et al.*, 2014; Sampath *et al.*, 2014). Since, mutation rate determines genetic diversity (Nei and Li, 1979), if progenitors have greater nucleotide and TIP diversity, this might indicate faster exploration of mutation space within the genome, generating genetic variation important for adaptation in early domestication.

TIPs associated with major expression changes or phenotypic effects are also expected to be maintained at low frequencies (Lye *et al.*, 2022). These are more likely to be TE insertions in genic or upstream/downstream regions influencing gene functions. This can translate into

increased phenotypic variability within the population for selection to act on. TIPs in the coding regions of the gene can result in loss of gene function, however, untranslated regions such as introns, upstream and downstream may act as promoters and enhancers for genes in their proximity (Hirsch and Springer, 2017). An increase in the frequencies of certain TIPs in the domesticates (compared to the progenitor and wild groups) may suggest selection for the presence of these TIPs in domesticated tomatoes (Figure 3.1B). On the other hand, TIPs with detrimental effect will be eliminated through purifying selection, resulting in a decrease in frequency. If TEs are involved in the domestication process, there would be a significant shift in frequency between the domesticated and the progenitor.

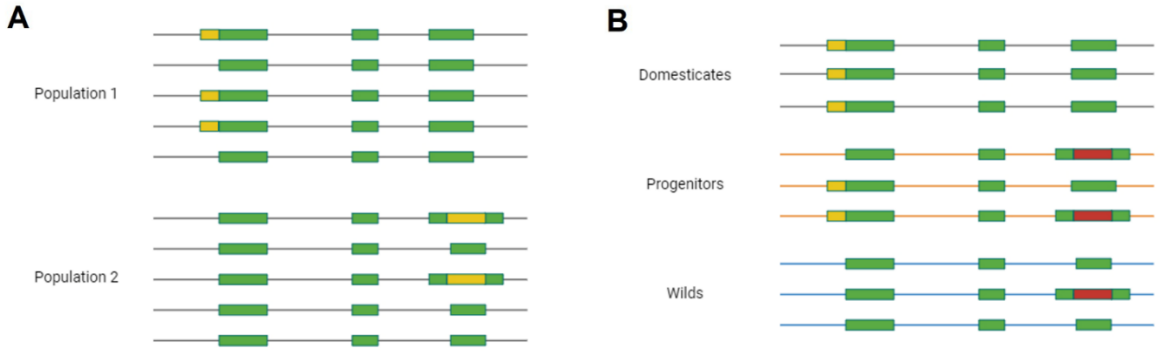


Figure 3.1: Illustration of transposon insertion polymorphisms (TIPs).
(A) Presence of polymorphic TE insertion (yellow) next to a gene (green) in Population 1 and an absence in Population 2 may suggests recent transposition. (B) TIPs can increase or decrease in frequency in the population, for example, if a TE (yellow) in the progenitor is beneficial it can increase frequency in the domesticated, whilst TE (red) with detrimental effect can be selected against during domestication.

Tomato (*Solanum lycopersicum* L.) is one of the most important crop species, with world production reaching over 180 million tons in 2019 (FAO, 2021). The domesticated tomato species is estimated to have originated ca. 7,000 years ago (Razifard *et al.*, 2020). There are 13 tomato species in the tomato clade, divided into four subgroups: Esculentum, Arcanum, Peruvianum and Hirsutum (Pease *et al.*, 2016). Wild tomatoes evolved in the Andean region of South America, including Bolivia, Chile, Colombia, Ecuador and Peru (Bergougnoux, 2014). Studies have suggested that *S. pimpinellifolium* was domesticated in South America, giving rise to *S. lycopersicum* var. *cerasiforme*, and subsequently improved to result in *S. lycopersicum*; this is described as the “two-step” process of the tomato domestication history (Lin et al., 2014; Blanca et al., 2015), accompanied by a notable increase in fruit size (Figure 3.2).

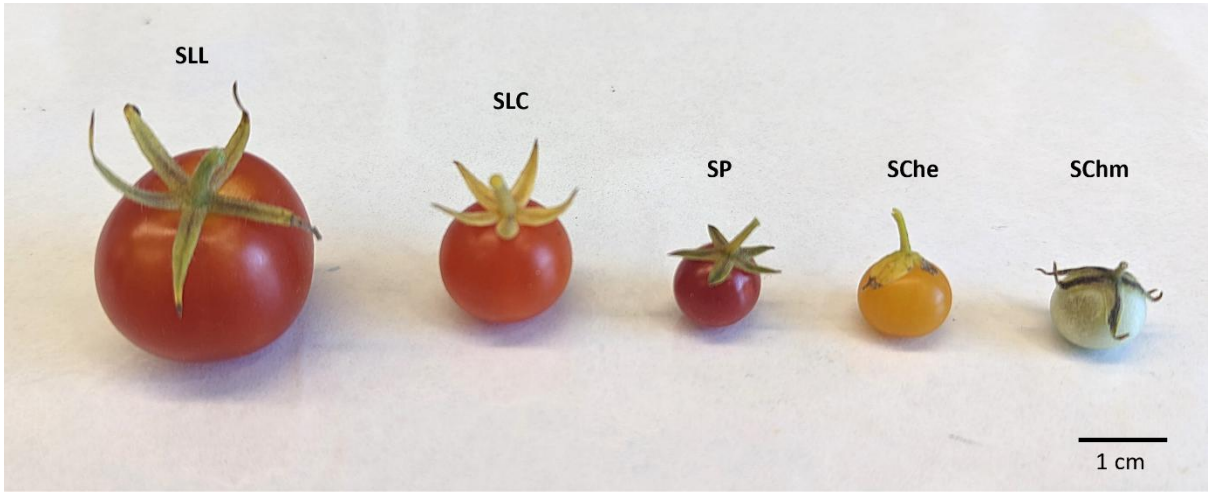


Figure 3.2: Images of different species of tomatoes.
This illustrates the size difference between the domesticated tomato *Solanum lycopersicum* (SLL; accession LA0395), the intermediate *S. lycopersicum* var. *cerasiforme* (SLC; LA1324), the progenitor *S. pimpinellifolium* (SP; LA1578) and other never-domesticated wild species, *S. cheesmaniae* (SChe; LA1039) and *S. chmielewskii* (SChm; LA2663).

Advancements in sequencing technologies have given us an abundance of resequencing data and tools available for TE insertion detection (Ou *et al.*, 2019; Vendrell-Mir *et al.*, 2019; Cho, 2021). This has supported the characterisation of TEs in various crops and their progenitors (Li *et al.*, 2017b; Gramazio *et al.*, 2019; Macko-Podgórni *et al.*, 2019; Tao *et al.*, 2021), however, focus on crop wild relatives (other than progenitors) is lacking. To examine mutation rates we used nucleotide diversity and TIP diversity because of their positive correlation (Wang and Obbard, 2023).

Here we investigate whether TEs contributed to tomato domestication. We investigate the following:

(i) whether the level of standing genetic variation in SNPs and TIPs is greater in the progenitor than in other wild relatives.
(ii) whether the progenitor has a greater variation in SNPs and TIPs near or in genic regions than never-domesticated wilds, as these might result in phenotypic effects.
(iii) whether individual TEs have changed in frequency between the domesticated and progenitor, indicating the influence of domestication of TIP frequency.

To achieve this, SNP and TIP analyses were performed on multiple domesticates, progenitors and never-domesticated wild accessions. For SNP count, distribution and diversity were

estimated for each species. Whole-genome TE annotation of the *S. lycopersicum* reference genome was performed to characterise the TE landscape. This was then utilised for TIP analysis to estimate TIP count, frequency, diversity and distribution were characterised for each species. Characterisation of TIP content will allow us to hypothesise whether genetic variation in TIP and SNPs was important for tomato domestication and can give us clues on the role of TEs in tomato domestication, with the potential to uncover adaptive alleles that may be important for future crop improvements.

3.3 Materials and Methods

3.3.1 Acquisition and processing of additional resequencing data

Raw whole genome sequencing (WGS) data were obtained from previous publications (Causse *et al.*, 2013; Aflitos *et al.*, 2014; Lin *et al.*, 2014; Alonge *et al.*, 2020; Cambiaso *et al.*, 2019; Stam *et al.*, 2019; Gramazio *et al.*, 2020; PRJNA713664) and downloaded from EBI-ENA (www.ebi.ac.uk/ena). Accessions were checked for quality using FastQC v0.11.9 (Andrews, 2010). Sequences were then trimmed and filtered using Trimmomatic v0.32 (Bolger *et al.*, 2014) to remove poor-quality bases and reads: Illumina adapters were removed; leading and trailing bases with quality below 5 were removed; reads shorter than 72bp were removed; sliding window trimming was performed for a window size of 4 with the required quality of 15. Samples were aligned to the reference *Solanum lycopersicum* SL4.0 genome (Hosmani *et al.*, 2019) using Bowtie2 v2.2.3 (Langmead and Salzberg, 2012). Subsequently, the sequencing depth of the samples for the reference genome was calculated. Samples were filtered to obtain high-quality samples using the following criteria: > 30M reads, > 80% survival read pairs, > 90% alignment rate and > 5x sequencing depth (Table 3.1).

Table 3.1: Accessions used in the study.

Sample ID	Species	Accession	Phylogroup	Domestication group	Origin	read count	filtered reads (Mbp)	average read length (bp)	alignment rate	sequencing depth
D_H1	*S. lycopersicum*	LA1090	Esculentum	Domesticates	USA	1.59E+08	137.322	100	0.9994	27.2001
D_H2	*S. lycopersicum*	EA1088	Esculentum	Domesticates	USA	1.75E+08	160.616	100	0.9989	32.7226
D_H3	*S. lycopersicum*	EA00940	Esculentum	Domesticates	USA	1.69E+08	153.330	100	0.9986	24.2054
D_H4	*S. lycopersicum*	EA01019	Esculentum	Domesticates	USA	1.57E+08	143.051	100	0.9983	27.6146
D_H5	*S. lycopersicum*	EA01037	Esculentum	Domesticates	USA	1.64E+08	136.538	100	0.9987	26.5118
D_H6	*S. lycopersicum*	TR00021	Esculentum	Domesticates	USA	1.68E+08	151.602	100	0.9987	27.1123
D_L1	*S. lycopersicum*	LA2260	Esculentum	Domesticates	Peru	3.76E+07	34.581	100	0.9961	5.63537
D_L2	*S. lycopersicum*	LA0113	Esculentum	Domesticates	Peru	1.60E+08	149.208	100	0.9905	25.5672
D_L3	*S. lycopersicum*	LA1421	Esculentum	Domesticates	Ecuador	1.69E+08	154.973	100	0.9949	26.8664
D_L4	*S. lycopersicum*	Allungato piccolo	Esculentum	Domesticates	Brazil	3.57E+07	30.201	100	0.9969	5.6700
D_L5	*S. lycopersicum*	EA03222	Esculentum	Domesticates		1.76E+08	143.410	100	0.9977	28.6837

D_L6	*S. lycopersicum*	LYC3340	Esculentum	Domesticates		1.81E+08	159.986	100	0.9930	26.9284
D_L7	*S. lycopersicum*	PI129097	Esculentum	Domesticates		1.61E+08	148.359	100	0.9960	27.1152
D_L8	*S. lycopersicum*	PI272654	Esculentum	Domesticates		1.75E+08	157.540	100	0.9912	29.6193
D_L9	*S. lycopersicum*	PI311117	Esculentum	Domesticates		1.78E+08	163.018	100	0.9985	31.8918
P_P1	*S. pimpinellifolium*	BGV006454	Esculentum	Progenitors	Peru	6.12E+07	55.434	150	0.9927	11.0483
P_P2	*S. pimpinellifolium*	BGV015382	Esculentum	Progenitors	Peru	5.27E+07	44.838	150	0.9937	8.8231
P_P3	*S. pimpinellifolium*	BGV013720	Esculentum	Progenitors	Peru	5.42E+07	47.474	150	0.9921	9.2893
P_P4	*S. pimpinellifolium*	BGV007145	Esculentum	Progenitors	Ecuador	5.25E+07	46.298	150	0.9948	9.8240
P_P5	*S. pimpinellifolium*	LA0722	Esculentum	Progenitors	Peru	6.72E_07	65.202	100	0.9875	7.6008
P_P6	*S. pimpinellifoliium*	LA1589	Esculentum	Progenitors	Peru	8.52E+07	74.937	150	0.9913	14.5598
P_P7	*S. pimpinellifoliium*	LA1584	Esculentum	Progenitors	Peru	1.52E+08	129.521	100	0.9891	18.9461
P_P8	*S. pimpinellifoliium*	LA1578	Esculentum	Progenitors	Peru	1.73E+08	154.540	100	0.9903	22.7577
P_P9	*S. pimpinellifoliium*	LA1547	Esculentum	Progenitors	Ecuador	2.36E+08	228.800	150	0.9912	26.7857
P_P10	*S. pimpinellifoliium*	LA2093	Esculentum	Progenitors	Ecuador	1.61E+08	126.170	150	0.9822	22.5896

P_P11	*S. pimpinellifoliium*		Esculentum	Progenitors		1.11E+08	102.697	150	0.9917	18.7913
P_P12	*S. pimpinellifoliium*		Esculentum	Progenitors		1.12E+08	101.109	150	0.9890	16.1079
P_P13	*S. pimpinellifoliium*	LA1578	Esculentum	Progenitors	Peru	1.04E+08	96.506	150	0.9778	15.8117
P_P14	*S. pimpinellifoliium*	LA1589	Esculentum	Progenitors	Peru	1.05E+08	97.443	150	0.9785	15.6339
W_CHE1	*S. cheesmaniae*	LA0746	Esculentum	Wilds	Ecuador	4.46E+07	41.670	100	0.9919	5.67232
W_CHE2	*S. cheesmaniae*	LA1406	Esculentum	Wilds	Ecuador	9.36E+07	89.967	150	0.9936	19.2643
W_CHE3	*S. cheesmaniae*	LA3124	Esculentum	Wilds	Ecuador	1.02E+08	97.981	150	0.9943	22.1754
W_CHE4	*S. cheesmaniae*	LA1039	Esculentum	Wilds	Ecuador	1.05E+08	97.474	150	0.9815	21.2129
W_CHE5	*S. cheesmaniae*	LA1406	Esculentum	Wilds	Ecuador	9.04E+08	82.211	150	0.9788	20.0897
W_GAL1	*S. galapagense*	LA1044	Esculentum	Wilds	Ecuador	1.55E+08	135.426	100	0.9898	20.7638
W_GAL2	*S. galapagense*	LA0483	Esculentum	Wilds	Ecuador	1.65E+08	153.518	100	0.9917	23.6611
W_GAL3	*S. galapagense*	LA1401	Esculentum	Wilds	Ecuador	1.78E+08	157.988	100	0.9925	24.3842
W_GAL4	*S. galapagense*		Esculentum	Wilds		1.14E+08	109.194	150	0.9924	23.3251
W_GAL5	*S. galapagense*	LA0436	Esculentum	Wilds	Ecuador	2.49E+08	240.023	150	0.9874	25.5510

W_ARC1	*S. arcanum*	LA2157	Arcanum	Wilds	Peru	1.6E+08	142.709	100	0.9505	11.6225
W_ARC2	*S. arcanum*	LA2172	Arcanum	Wilds	Peru	1.71E+08	143.738	100	0.9414	12.5046
W_CHM1	*S. chmielewskii*	LA2663	Arcanum	Wilds	Peru	1.68E+08	153.749	100	0.9543	13.7539
W_CHM2	*S. chmielewskii*	LA2695	Arcanum	Wilds	Peru	1.72E+08	154.289	100	0.9553	13.6897
W_NEO1	*S. neorickii*	LA0735	Arcanum	Wilds	Peru	1.75E+08	159.291	100	0.9556	13.8937
W_NEO2	*S. neorickii*	LA2133	Arcanum	Wilds	Peru	1.78E+08	152.241	100	0.9555	13.1522
W_CHI1	*S. chilense*	LA3111	Peruvianum	Wilds	Peru	1.67E+08	151.696	100	0.9410	9.34417
W_CHI2	*S. chilense*	CGN15530	Peruvianum	Wilds	Peru	1.64E+08	141.945	100	0.9405	10.1519
W_CHI3	*S. chilense*	CGN15532	Peruvianum	Wilds	Peru	1.60E+08	136.167	100	0.9340	8.7608
W_PER1	*S. peruvianum*	LA1278	Peruvianum	Wilds	Peru	1.71E+08	147.908	100	0.9335	10.7235
W_PER2	*S. peruvianum*	LA1954	Peruvianum	Wilds	Peru	1.77E+08	154.387	100	0.9349	10.5920
W_HUA1	*S. huaylasense*	LA1364	Peruvianum	Wilds	Peru	1.70E+08	148.070	100	0.9350	9.98779
W_HUA2	*S. huaylasense*	LA1365	Peruvianum	Wilds	Peru	1.58E+08	135.374	100	0.9435	9.68853
W_HUA3	*S. huaylasense*	LA1983	Peruvianum	Wilds	Peru	1.78E+08	151.624	100	0.9381	11.8373

W_PEN1	*S. pennellii*	LA0716	Hirsutum	Wilds	Peru	1.90E+08	156.346	150	0.9308	11.5727
W_PEN2	*S. pennellii*	LA1272	Hirsutum	Wilds	Peru	1.25E+08	112.186	100	0.9351	6.97782
W_PEN3	*S. pennellii*	LA1926	Hirsutum	Wilds	Peru	1.68E+08	148.052	100	0.9381	9.68308
W_HAB1	*S. habrochaites*	LYC4	Hirsutum	Wilds	Peru	1.59E+08	148.631	100	0.9246	9.81933
W_HAB2	*S. habrochaites*	LA0407	Hirsutum	Wilds	Ecuador	1.78E+08	156.803	100	0.9262	10.1320
W_HAB3	*S. habrochaites*	PI134418	Hirsutum	Wilds		1.70E+08	149.150	100	0.9255	10.0225
W_HAB4	*S. habrochaites*	CGN157592	Hirsutum	Wilds		1.67E+08	149.583	100	0.9254	9.8159

3.3.2 Whole-genome sequencing

To supplement the resequencing data, whole-genome sequencing of additional wild tomato accessions was performed. Seeds were obtained from Tomato Genetic Resource Centre (TGRC; https://tgrc.ucdavis.edu/) and Centre for Genetic Resources the Netherlands (CGN) Wageningen University (https://cgngenis.wur.nl/). The seeds were grown in a glasshouse at the University of Southampton, UK at 20°C during the day and 18°C at night. Accessions were selected based on supplementing wild tomato species with the fewest number of samples and availability of accessions. Five accessions were selected including two *S. cheesmaniae* (W_CHE), one *S. galapagense* (W_GAL), one *S. neorickii* (W_NEO), and one *S. pennellii* (W_PEN) (Table B1). We found that TE insertion identification was influenced by insert size (see below), therefore four further accessions were sequenced at large insert size including two *S. cheesmaniae* (W_CHE) and two *S. pimpinellifolium* (P_P). Leaf samples were taken from each accession and frozen prior to DNA extraction using a modified CTAB protocol (Doyle and Doyle, 1990). Sequencing was performed at Novogene Bioinformatics Institute (Cambridge, UK). Library preparation of the samples was performed, followed by sequencing using Illumina NovoSeq 6000 (Illumina, USA). These nine samples were then trimmed, filtered and aligned to the reference genome, as above (Table 3.1).

3.3.3 Single Nucleotide Polymorphism (SNP) analysis

To explore the relationship between the accessions and the contribution of SNPs to the genetic diversity in the species analysed, single nucleotide polymorphism (SNP) analysis was performed. Bam files from reference mapping were processed with Picard v2.18.14 (http://broadinstitute.github.io/picard/) and combined to make a VCF file. SNPs were called from the VCF file using samtools bcftools call (Li *et al.*, 2009). SNPs were filtered using bcftools filter with the following criteria: QUAL>20, DP>5 and MAF>0.05; and bcftools view with -m2 -M2 -v snps. Genetic distance between pairs of samples was calculated with VCF2Dis v 1.52 (https://github.com/BGI-shenzhen/VCF2Dis).

Principal Component Analysis (PCA) was produced using Plink v1.9 (Chang *et al.*, 2015). LD-pruning was performed with parameters '--indep-pairwise 50 10 0.1' based on LD decay plot produced by PopLDdecay v. 3.43 (Zhang *et al.*, 2019) with parameters '-MaxDist 500 -Het 0.1 -Miss 0.1' according to (Guo *et al.*, 2019). Next, eigenvec and eigenval values were obtained using Plink - -pca.

To compare SNP data by species, the VCF file was split according to the species and only polymorphic SNPs with no missing data were included. Another measure of genetic diversity is Watterson's estimator of theta (θ), based on the number of segregating sites or SNPs within each species (Watterson, 1975). This was estimated using Pegas in R (https://rdrr.io/cran/pegas/) and divided by genome size to obtain θ_w.

3.3.3.1 Calculation of SNP diversity

Nucleotide diversity measures the average number of pairwise differences between all possible pairs of individuals in each species, taking into account the frequency of SNPs (Nei and Li, 1979). To estimate nucleic diversity (π), allele frequency was first calculated using VCFtools v0.1.16 with - -freq option (Danecek *et al.*, 2011). To calculate SNP diversity (π) for each species, the nucleotide diversity formula by Nei and Li (1979) was used. Where *n* is the sample size of accessions, *f* is the frequency of the SNP in the species being considered and *m* is the size of the genome being assayed.

$$\text{SNP diversity } (\pi) = \left(\frac{n}{n-1}\right)\left(\frac{\sum 2f(1-f)}{m}\right)$$

To test whether the SNP diversities are significantly different to each other, we calculate the variance of each diversity estimate assuming free recombination (Tajima, 1983; Charlesworth and Charlesworth, 2010). Where *n* is the number of accessions, π is the SNP diversity and *m* is the size of the genome.

$$\text{Variance } (V) = \frac{1}{m}\left(\frac{(n+1)\pi}{3(n-1)} + \frac{2(n^2+n+3)\pi^2}{9n(n-1)}\right)$$

Assuming the nucleotide diversity is approximately normally distributed we can calculate a test statistic

$$X = \sum \frac{(\pi_i - \bar{\pi})^2}{V_i}$$

where $\bar{\pi} = \frac{\sum \pi_i / V_i}{\sum 1/V_i}$ is the weighted average nucleotide diversity, weighting each species' estimate by the reciprocal of the variance. X is approximately chi-square distributed with degrees of freedom equal to the number of species minus one.

3.3.3.2 Genome-wide distribution of SNPs

The position and annotation of SNPs was obtained using SNPdat v.1.0.5 (Doran and Creevey, 2013) with the fasta file of the SL4.0 reference genome and its GFF file with gene positions. Gene

positions were identified and 1kb upstream and 1kb downstream were annotated as up/downstream of a gene. SNPs were annotated as located in intergenic, up/downstream or genic regions.

3.3.4 Whole-genome TE annotation

For whole-genome TE annotation, the Extensive de-novo TE annotator (EDTA) pipeline (Ou *et al.*, 2019) was used. The EDTA pipeline annotates TEs in a reference genome, identifying LTR-retrotransposons, TIR transposons, MITEs and Helitrons that use structural (LTRharvest, LTR_FINDER, LTR_retriever, TIR-Learner, HelitronScanner) and homology (RepeatModeler and RepeatMasker) approach to produce a comprehensive TE library. The *Solanum lycopersicum* reference SL4.0 genome was supplied to EDTA v2.0.0, plus ITAG4.0 coding regions and gene positions of the SL4.0 genome assembly to avoid gene sequences being added to the TE library. TEs were classified into order and superfamily level using Wicker *et al.* (2007).

3.3.5 Transposon insertion polymorphism (TIP) analysis

TE detection across accessions was performed using PopoolationTE2 v1.10.03 (Kofler *et al.*, 2016). A TE-merged-reference genome was created by combining masked reference SL4.0 genome and TE library extracted from the TE annotation output of the EDTA pipeline. A TE hierarchy file was created using the entries from the TE annotation (from the EDTA pipeline) to identify the ID, order and family information of each TE sequence. Paired-end reads for each sample after Trimmomatic filtering were used for this analysis. These were mapped separately to the TE-merged-reference with bwa-bwasw v0.7.17 (Li and Durbin, 2009) and then PopoolationTE2 *se2pe* restored the paired-end information. PopoolationTE2 *ppileup* then used the bam files to produce a ppileup file with a minimum mapping quality of 15. To identify TE insertions, the following parameters were used: (i) identifySignatures (–mode joint, –min-count 2, --signature-window fix500 (other filters were explored to find the optimum setting, see 3.4.3), (ii) frequency, (iii) filterSignatures (--min-count 5 --max-otherte-count 2, --max-structvar-count 2) and (iv) pairupSignatures. Addition TE filtering was performed using zygosity score based on the portion of reads supporting the insertion: (i) TE insertions with missing data were removed to allow only for comparisons of informative sites between accessions and to account for small number of samples; (ii) TEs with a zygosity lower than 0.25 in all samples were removed as they are more likely to be absent or heterozygous (i.e. false positives due to insufficient support). The TE matrix was transformed into binary; within accessions, TEs with zygosity of > 0.05 were scored as present and zygosity of < 0.05 as absent, following (Castanera *et al.*, 2023).

3.3.5.1 Validation of TEs

For the visual validation of TEs, 50 randomly chosen TEs in each accession (1,200 randomly chosen TEs), spanning all TE superfamilies were inspected visually using IGV (Robinson *et al.*, 2011). Additionally, 50 TEs were visually validated across accessions to check their presence or absence using IGV (Robinson *et al.*, 2011).

3.3.5.2 Estimating group frequencies

TE frequencies were obtained by calculating the proportion of samples within each species that contain each TE insertion. These were then classified into low-frequency (≤ 25%), intermediate (>25%, <75%), and high-frequency (≥75%) TIPs.

3.3.5.3 Calculation of TE diversity

To calculate TE diversity (*x*) for each species, the equivalent of nucleotide diversity (Nei and Li, 1979) was adapted but on TEs. Where *n* is the sample size of accessions, *f* is the frequency of the TE in the species being considered and *m* is the size of the genome being assayed. The variance of this estimate was calculated with the same method as the SNPs.

$$\text{TE diversity } (x) = \left(\frac{\mathrm{n}}{\mathrm{n}-1}\right)\left(\frac{\sum 2\mathrm{f}(1-\mathrm{f})}{m}\right)$$

3.3.5.4 Genome-wide distribution of TIPs

The position of TEs was obtained by mapping onto SL4.0 reference genic features (i.e., intergenic, 1kb upstream, genes, 1kb downstream). Gene positions were identified and 1kb upstream and 1kb downstream were extracted. These genomic positions were intersected with TE positions to annotate location within the reference genome.

3.3.6 Gene Ontology (GO) analysis

TIPs that were selected during tomato domestication would have a drastic change in population frequency. For example, TIPs at low frequency (≤ 25%) in the progenitor and high frequency (≥

75%) in the domesticated species, or vice versa. These TIPs were identified and those found within 1kb of a gene, or in a gene, were extracted. To understand the biological processes associated with these genes, TopGO (Alexa and Rahnenfuhrer, 2022) was used to test for over-representation of gene ontology (GO) terms. SlimGO annotation of ITAG4.0 *Solanum lycopersicum* was used in a gene-to-GO format (http://systemsbiology.cau.edu.cn/agriGOv2/download.php). Fisher's exact tests were used to compare genes and the background list of all genes associated with their GO annotation, focusing on biological processes as the ontology of interest with a minimum of five genes per GO term. P-values were adjusted using the Benjamini & Hochberg (1995) method to control the false discovery rate; those with an adjusted p-value < 0.05 were considered significant.

3.3.7 Genome size estimation

To check if genome mapping using SL4.0 reference genome was influenced by genome size, we estimated genome sizes for seven species. Leaf samples were sent to Kew for C-value estimates through flow cytometry. Two accessions of seven species (*S. lycopersicum*, *S. pimpinellifolium*, *S. cheesmaniae*, *S. galapagense*, *S. chmielewskii*, *S. pennellii*, and *S. habrochaites* were estimated for their genome size following the protocol described in Hanson *et al.* (2005).

3.3.8 Statistical analysis

We also assessed whether there were differences in mapping rates among groups, as low sequencing depth may influence the number of TIPs detected. To explore variation among samples in the dataset, the following tests were conducted. To test for the differences in the number of clean reads, mapping rate, genome size and insert size between tomato groups, one-way ANOVA (analysis of variance) was implemented if homogeneity and normality assumptions were satisfied. Homogeneity was tested with Levene's Test for Homogeneity of Variance. Normality was tested with the Shapiro-Wilk normality test. ANOVA was followed by Tukey testing for multiple comparisons of means. For comparisons that did not satisfy the assumptions of ANOVA, a Kruskal-Wallis test was performed, followed by Dunn test for pairwise comparison.

We observed substantial difference between TIP PCA and SNP PCA (see 3.4.3), therefore we tested whether read insert size, read inner distance, read length and signature window setting affected the TIPs detected by PopoolationTE2. PopoolationTE2 identifies signatures of TE

insertions using a window-based approach. To test for differences in TIP count under different window settings, the Kruskal-Wallis test was performed. In addition, correlation between TIP count and inner distance, and TIP count and read length was tested with Spearman correlation.

To test for significant difference between the SNP/TIP diversity means using the variance of each estimate (V) and weighted mean diversity (x_w) calculated above. The test statistic (k) is calculated. If we just have two means, then we can just use a normal test; so calculate k_1. This was performed for each pair of species and repeated for regions near and within genes for SNPs and TIPs, separately.

$$\boldsymbol{k_1 = \frac{(\pi_1 - \pi_2)}{\sqrt{V_1 + V_2}}}$$

To test for association between tomato group, TE order, group frequency, and genomic distribution, $\chi 2$ tests were performed as a pairwise comparison. To test for the difference in TIP count between tomato group, one-way ANOVA was performed. To test for the effect of two variables on TIP count, two-way ANOVA was performed between (i) the number of reference/non-reference TIPs and tomato group, (ii) TE order and tomato group, (iii) genomic region and tomato group, and (iv) TE classification frequency and tomato group.

3.4 Results

3.4.1 Whole-genome TE annotation

The first step in identifying transposon insertion polymorphisms (TIPs) is the whole-genome annotation of transposable elements (TEs) in the SL4.0 *S. lycopersicum* reference genome. To do this the EDTA pipeline was performed revealing 913,428 TE sequences, that cover 64.14% of the domesticated tomato reference genome. Only 16,413 TEs (1.80%) were structurally intact with the rest found through the homology approach. 713,363 TE sequences (78.10%) were classified at the superfamily level (Table 3.2) and 110,320 sequences (12.08%) were classified as repetitive regions that were not classified at the order level. TEs from 4,330 families were discovered in the tomato genome, with the landscape dominated by Gypsy LTRs (Table 3.2). The full EDTA output is reported in Table B3. The proportion of TEs was similar across the 12 chromosomes (Figure 3.3A) with no significant difference in TE density between TE orders across chromosomes ($\chi 2$ = 3.853, df=55, p = 1.00). Gene density was greatest in distal chromosome regions and was inversely proportional to TE density (Figure 3.3B). MITEs follow the same density distribution of genes (Figure 3.3B), indicating an insertional bias near genes as shown in other species (Lu *et al.*, 2012).

Table 3.2: Whole-genome transposable element (TE) annotation.

Whole-genome annotation of the reference SL4.0 *Solanum lycopersicum* genome from the EDTA output, with the number of identified TEs classified into orders and superfamilies.

Class	Order	Superfamily	Family count	count	total	(%)	total bp
1	LTR	Copia	626	66,359	308,766	(33.80%)	54,341,044
		Gypsy	437	152,700			164,210,454
		unknown	283	89,707			81,651,012
	non-LTR	LINE	24	1,838	1,876	(0.21%)	939,919
		unknown	1	38			8,433
2	TIR	Tc1-Mariner	64	26,249	299,106	(32.75%)	14,898,409
		hAT	131	15,027			7,999,436
		Mutator	534	141,806			83,147,407
		PIF-Harbinger	68	25,557			15,097,610
		CACTA	262	90,467			50,596,484
	MITE	Tc1-Mariner	92	10,520	35,601	(3.90%)	2,141,502
		hAT	101	7,935			1,975,167
		Mutator	148	14,818			3,607,337
		PIF-Harbinger	21	519			89,705
		CACTA	41	1,809			356,956
	Helitron	Helitron	250	157,759	157,759	(17.27%)	76,123,661
	Unknown	repeat region	1247	110,320	110,320	(12.08%)	34,061,891

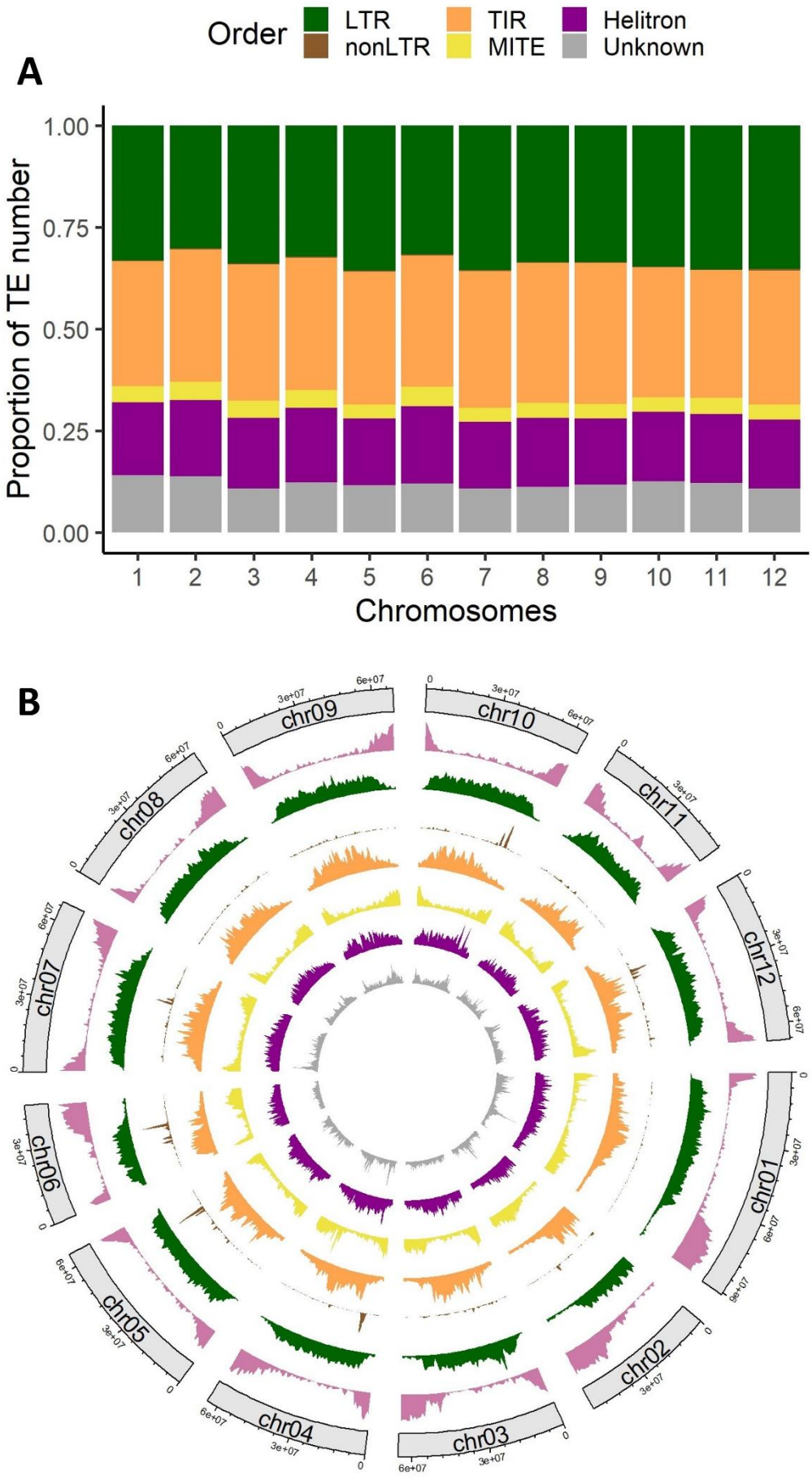


Figure 3.3: Whole-genome distribution of transposable elements (TEs).

Whole-genome TE annotation from EDTA output. (A) The proportion of LTR, non-LTR, TIR, MITE, Helitron and unknown TEs in each chromosome. (B) Relative densities of genes (pink), LTR (green), non-LTR (brown), TIR (orange), MITE (yellow), Helitron (purple) and unknown (grey) TEs.

3.4.2 Sample exploration

Accessions that met the first sequence quality criteria of '> 30M reads, > 80% survival read pairs, > 90% alignment rate and > 5x sequencing depth' were further analysed to check for differences between the three groups that may influence our analysis, the following were compared between tomato groups (i.e., domesticates [D] vs progenitor [P] vs wild [W]): (i) number of reads per sample, (ii) mapping rate, and (iii) genome size. The initial study used 60 accessions that included domesticated, progenitor and wild accessions from 10 species (Table B1). Full filtering and mapping statistics are reported in Table B2.

First, low sequencing depth may influence the number of TIPs detected, so we tested for the difference in read counts between tomato groups. There was a significant difference in the number of clean reads per individual between groups (Kruskal-Wallis: $\chi^2 = 10.252$, df=2, p = 0.01; Figure 3.4A); between domesticates and wilds (Dunn's test: p=0.006) and between progenitors and wilds (Dunn's test: p=0.003). Samples with less than 80M clean reads (ca. 10X coverage) were removed (9 samples: D_L1, D_L4, P_P1, P_P2, P_P3, P_P4, P_P5, P_P6, W_CHE1), leaving 51 samples. For the subset of 51 samples, there was no significant difference in clean reads between groups (Kruskal-Wallis: $\chi^2 = 5.614$, df=2, p = 0.060; Figure 3.4B).

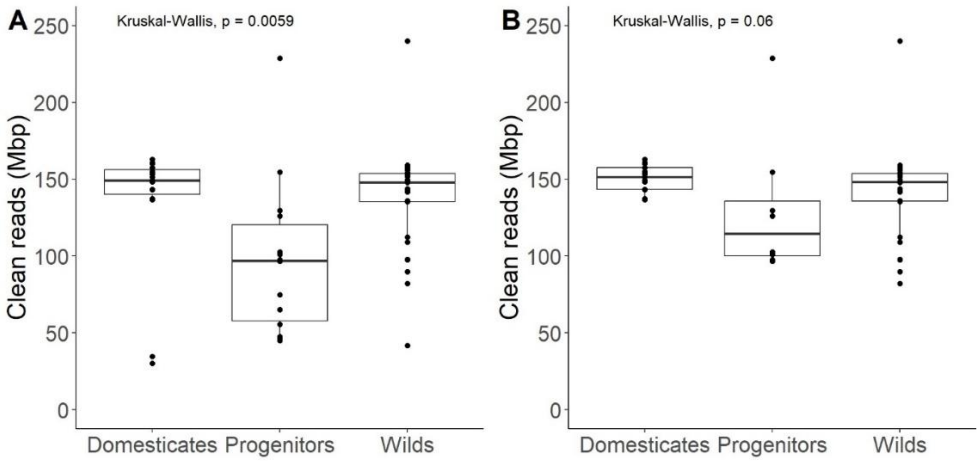


Figure 3.4: Sample filtering based on the difference in clean reads between tomato groups. Clean reads (Mbp) of (A) 60 samples and (B) 51 samples grouped by tomato group.

Secondly, reduction in mapping rate may lead to the reduction in the number or reliability of TIPs detected, so we tested for the difference in mapping rates between tomato groups. The mean mapping rate across all samples was 97.34%, but this varied between phylogroups (Pease *et al.*, 2016), i.e. separating the species into Esculentum (D [n=13], P [n=8], W_CHE and W_GAL [n=9]), Arcanum (W_ARC, W_CHM, and W_NEO; n=6), Peruvianum (W_CHI, W_PER, and W_HUA; n=8), Hirsutum (W_PEN and W_HAB; n=7). There was a significant difference in the mapping rate between Esculentum and all the other phylogroups (Kruskal-Wallis: $\chi^2 = 38.880$, df=3, $p < 0.001$; Figure 3.5A). Species more distantly related to *S. lycopersicum* (i.e., the reference genome) had lower mapping rates (ANOVA: $F(11) = 185.5$, $p < 0.001$; Figure 3.5A). Domesticates, progenitors and wilds (W_CHE and W_GAL) had significantly greater mapping rate than all other species (Tukey: $p < 0.001$).

Thirdly, to check if mapping rate was influenced by the difference in genome size, we estimated genome sizes for seven species (Table B4). There was a significant difference in genome size between species (ANOVA: $F(6) = 148.5$, $p < 0.001$; Figure 3.5B). W_PEN and W_HAB had significantly greater genome size compared to D_L, P_P, W_CHE, W_GAL and W_CHM (Tukey: $p < 0.001$). Reduction in mapping rate and difference in genome sizes may lead to the reduction in the number of TIPs detected in species in Arcanum, Peruvianum, and Hirsutum, therefore these were removed from further analysis to limit the bias from the reference genome, resulting in 30 samples.

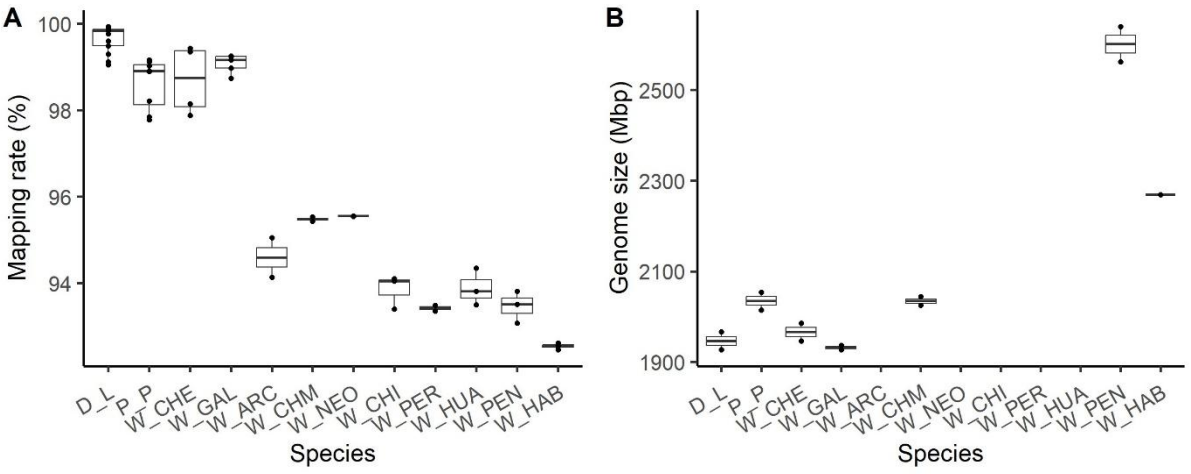


Figure 3.5: Sample filtering based on mapping rate and genome size.
(A) Mapping rate (%) of each species and (B) genome size of seven species (blank indicates no data for these species). Phylogroup Esculentum includes the domesticates *S. lycopersicum* (D_L), the progenitor *S. pimpinellifolium* (P_P), and the wilds *S. cheesmaniae* (W_CHE) and *S. galapagense* (W_GAL). Group Arcanum includes *S. arcanum* (W_ARC), *S. chmielewskii* (W_CHM), and *S. neorickii* (W_NEO), Peruvianum includes *S. chilense* (W_CHI), *S. peruvianum*

(W_PER), and *S. huaylasense* (W_HUA) and Hirsutum includes *S. pennellii* (W_PEN), and *S. habrochaites* (W_HAB).

3.4.3 TE detection

To further reduce biased comparisons between samples, Kofler *et al.* (2016) recommends subsampling reads to give an equal number per sample. Differences in read count among samples was taken into account by subsampling at 80M reads. TIPs were detected in the 30 samples using PopoolationTE2. The initial principal component analysis (PCA) of 30 samples based on SNPs (Appendix Figure B. A) differed from the TIP PCA (Appendix Figure B. 1B). Samples which should be genetically grouped together (based on prior knowledge and the SNP PCA and Appendix Figure B. A) were, for the progenitors and the wilds, split into two groups in the PCA (indicated in Appendix Figure B. 1B). This suggested that samples were differently annotated for TIPs based on insert size (length of DNA fragment) as progenitors and wilds with larger insert size clustered with the domesticates (left cluster on Appendix Figure B. B) and progenitors and wilds with shorter insert size clustering together (right cluster on Appendix Figure B. 1B). Therefore, variability in insert size could influence the accuracy of TE insertion detection. Insert sizes were computed based on median inner distance and average read length of paired ends reads for each sample (Table B5). There was a significant difference in insert size between tomato groups (Kruskal-Wallis: $\chi^2 = 8.765$, df=2, p=0.012; Figure 3.6A); this was between domesticates and wilds (Dunn's test: p=0.005). Samples with insert size below 400bp were removed from the analysis (P_11, P_12, W_CHE2, W_CHE3, W_GAL4, W_GAL5), resulting in 24 samples (D=13, P=6, W=5). There was no significant difference in insert size between tomato groups with 24 samples (Kruskal-Wallis: $\chi^2 = 3.342$, df=2, p=0.188; Figure 3.6B).

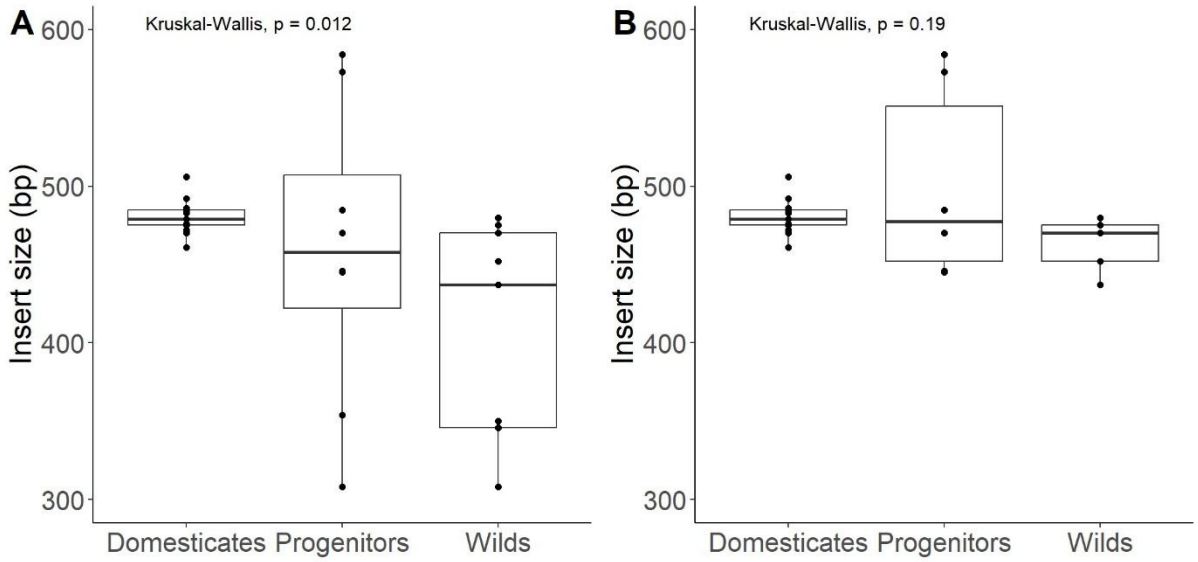


Figure 3.6: Insert size between tomato groups.

Insert size (bp) of (A) 30 samples and (B) 24 samples grouped by tomato group.

Signatures of TE insertions (paired end reads supporting a TE insertion) are identified using a window-based approach. To identify the optimal signature window setting, the pipeline was run with four different size windows for identification of TE insertions. Signature window parameters using the minimum (minimumSampleMedian), maximum (maximumSampleMedian) median of the inner distance or fixed window size of 400 bp (fix400), and 500 bp (fix500) were run for the 24 samples. Overall, detection of TE insertions under minimumSampleMedian, maximumSampleMedian, fix400, fix500 resulted in 8,982 (Table B6), 8,506 (Table B7), 8,402 (Table B8), and 8,193 (Table B9) TIPs respectively. There was no significant difference in TIP count between tomato groups for any signature window settings tested (Kruskal-Wallis: $p > 0.05$; Figure 3.7A,D,G,J). Except for fix500, TIP count negatively correlated with median inner distance and positively correlated with read length (Figure 3.7; Table B10). This suggest that under fix500 setting, median inner distance and read length does not affect TIP count.

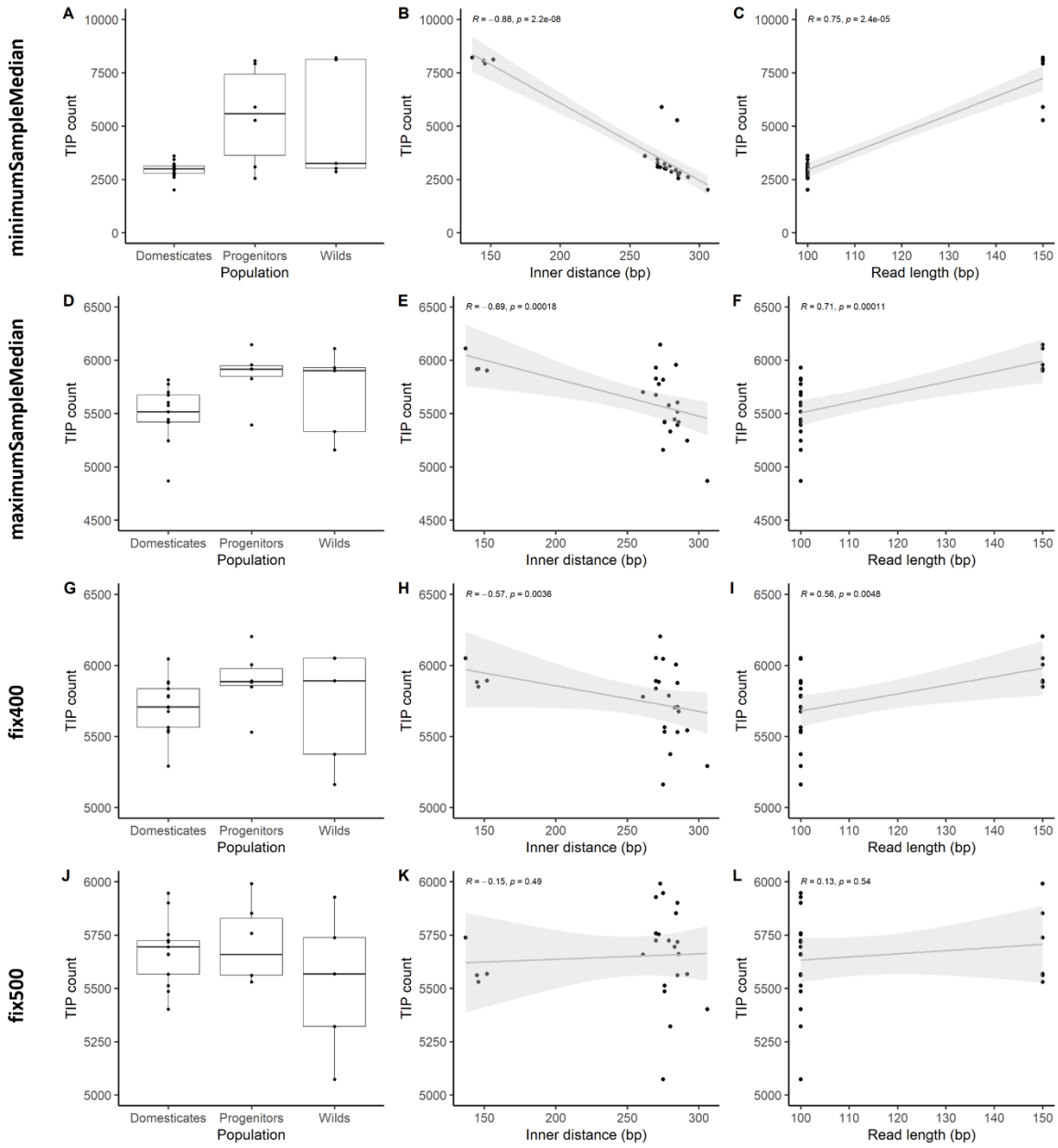


Figure 3.7: Transposon insertion polymorphism (TIP) count under different signature windows. The difference in TIP count between tomato group and its correlation with mean inner distance and read length under (A,B,C) minimumSampleMedian, (D,E,F) maximumSampleMedian, (G,H,I) fix400, and (J,K,L) fix500.

To visualise how different signature window settings affect the population structure of the 24 accessions, PCA of the first two PCs is shown in Figure 3.8. Samples clustered according to median inner distance and read length when TIPs were detected under minimumSampleMedian, with samples within the blue cluster having greater inner distance and shorter read length compared to samples in the red cluster (Figure 3.8A). Under maximumSampleMedian (Figure 3.8B), fix400 (Figure 3.8C), and fix500 (Figure 3.8D), samples are clustered more closely by species. The removal of samples with shorter insert sizes and the

use of the fix500 signature window setting eliminated insert size, inner distance and read length biases in the dataset, fixing the lack of clustering by tomato group (Appendix Figure B. C).

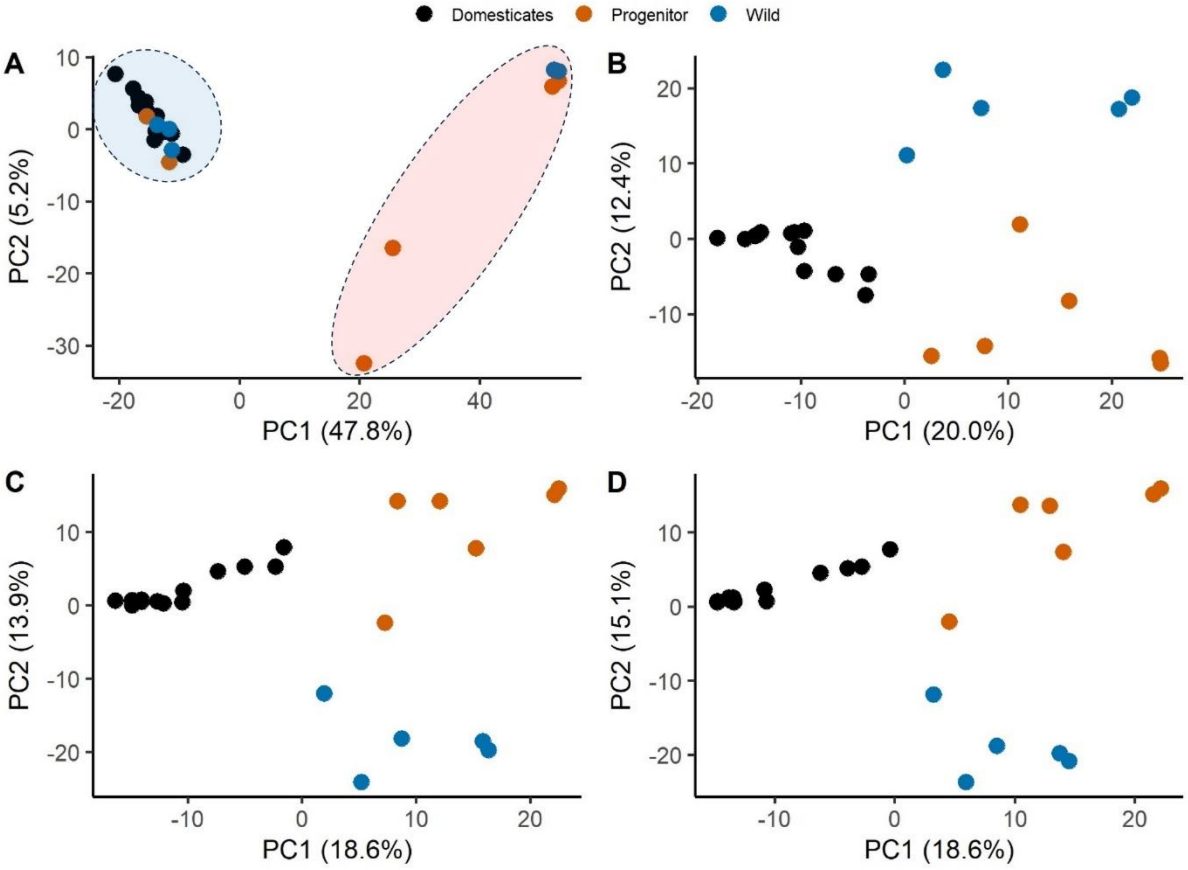


Figure 3.8: Principal Component Analysis (PCA) under different signature window settings. PCA of Transposon insertion polymorphisms (TIPs) identified under signature window settings (A) minimumSampleMedian, (B) maximumSampleMedian, (C) fix400 and (D) fix500. Clustering of samples with similar inner distance is evident under minimumSampleMedian setting, with accessions in the blue cluster having large inner distances and red cluster with short inner distances.

3.4.4 Single Nucleotide Polymorphism (SNP) analysis

To examine the contribution of SNPs on the genetic variation in domesticated (number of accessions; n=13), progenitor (n=6) and wild tomatoes (W_CHE: n=2; W_GAL: n=3), short variant calling detected 24,045,542 genetic variants comprising 1,154,595 indels and 22,890,947 single nucleotide polymorphisms (SNPs). Filtering steps resulted in 13,367,384 SNPs. Full SNP analysis statistics are reported in Table B11.

3.4.4.1 SNP count

Genetic diversity estimates were calculated using SNP diversity (π) and Watterson's estimator of theta (θ_w). Nucleotide diversities (π) in our tomato species were estimated (Table 3.3; Table B13). The progenitor had significantly greater nucleotide diversity than both wild species (two-tailed k-test: $p < 0.001$), but not domesticated species. However, Watterson's estimator of theta (θ_w) was used to estimate the number of segregating sites in each species and the progenitor had significantly different theta (θ_w) from the wild W_GAL but not W_CHE (two-tailed k-test: $p < 0.001$) or the domesticated species. This suggests greater genetic diversity in the progenitor compared to some wilds which could be due to a higher mutation rate or effective population size.

Table 3.3: Genome-wide estimates of SNP diversity.
A. Single nucleotide polymorphism (SNP) count, Nucleotide diversity (π), and Watterson's estimator of theta (θ_w) for the domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and wild species (*S. cheesmaniae*: W_CHE; *S. galapagense*: W_GAL). B. To test for significant difference between the SNP diversity means using the variance of each estimate (V) and weighted mean diversity (x_w) and test statistic (k) is calculated. C. Pairwise comparison of π means between species.

A. Species	SNP count	π	θw
Domesticates	8,335,520	2.76×10^{-3}	3.48×10^{-3}
Progenitors	9,032,417	5.66×10^{-3}	5.12×10^{-3}
Wild (CHE)	5,541,146	3.32×10^{-3}	7.17×10^{-3}
Wild (GAL)	4,675,654	2.42×10^{-3}	4.03×10^{-3}

B. Species	variance	weighted π	k
Domesticates	1.4E-12	2.0E+09	160296.8
Progenitors	3.4E-12	1.6E+09	1721610.6
Wild (CHE)	4.3E-12	7.7E+08	1731.7
Wild (GAL)	2.1E-12	1.2E+09	315998.0

C. Reference	Test	Normal deviate	P-value (2-tail)
Domesticates	Progenitors	-1321.863	2.00E+00
Domesticates	Wild (CHE)	-233.996	2.00E+00
Domesticates	Wild (GAL)	182.532	0.00E+00
Progenitors	Wild (CHE)	842.771	0.00E+00
Progenitors	Wild (GAL)	1380.235	0.00E+00

3.4.4.2 SNP distribution

The distribution of SNPs across the genome was annotated as intergenic, up/downstream of a gene or within genes. This was because SNPs near or within genes are more likely to influence genes resulting in phenotypic changes that can affect fitness. Most SNPs were found in the intergenic region compared to up/downstream and genic regions (Table 3.4). SNPs near and within genic regions were the greatest in the progenitor, followed by the domesticated and the wild species (Table 3.4). In each species, SNPs were less likely to be found within genes than expected by chance ($\chi 2$: df = 2, $p < 0.001$). We estimated SNP diversity near and within genic regions for each species (Table 3.4; Table B13) and found the progenitor to be significantly more diverse than both wild species (two-tailed k-test: $p < 0.001$). This suggests greater SNP diversity in regions where the mutations might have phenotypic effects, this can be important in adaptation in changing environments.

Table 3.4: SNP diversity estimates near and within genic regions.
A. Genomic distribution of single nucleotide polymorphisms (SNPs) in the domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and wild species (*S. cheesmaniae*: W_CHE; *S. galapagense*: W_GAL). B. SNP mcount and nucleotide diversity (π) estimates for each species based on near and within genic regions. C. To test for significant difference between the SNP diversity means using the variance of each estimate (V) and weighted mean diversity (x_w) and test statistic (k) is calculated. D. Pairwise comparison of π means between species.

A. Species	**intergenic**	**up/downstream**	**genes**
Domesticates	6,972,808	561,244	801,468
Progenitors	7,471,851	662,409	898,157
Wild (CHE)	4,639,151	399,176	502,819
Wild (GAL)	3,833,551	374,573	467,530
B. Species	**SNP count**	**π**	
Domesticates	1,362,712	1.58×10^{-3}	
Progenitors	1,560,566	3.22×10^{-3}	
Wild (CHE)	901,995	1.58×10^{-3}	
Wild (GAL)	842,103	1.58×10^{-3}	
C. Species	**variance**	**weighted π**	**k**
Domesticates	3.3E-12	4.8E+08	24371.7
Progenitors	8.0E-12	4.0E+08	229448.1
Wild (CHE)	8.5E-12	1.9E+08	9131.9
Wild (GAL)	5.6E-12	2.8E+08	14004.1
D. Reference	**Test**	**Normal deviate**	**P-value (2-tail)**
Domesticates	Progenitors	-487.729	2.00E+00
Domesticates	Wild (CHE)	-1.378	1.83E+00
Domesticates	Wild (GAL)	-0.642	1.48E+00

Progenitors	Wild (CHE)	402.779	0.00E+00
Progenitors	Wild (GAL)	443.300	0.00E+00

3.4.5 Transposon insertion polymorphism (TIPs) analysis

We are interested in whether species that were domesticated had different TE landscape and diversity than those which were not domesticated, and that these TEs may have aided the process of domestication. To investigate this, TE annotation of domesticated tomato, its wild progenitor and another closely related wild species, that was not domesticated, were compared. A final set of 24 samples (D=13, P=6, W=5) were analysed using PopoolationTE2 with fix500 signature window, resulting in 119,565 TE across all accessions. A total of 1,200 randomly chosen TEs were visually validated (50 TIPs from each accession) with an average of 91.5% of TIPs confirmed to be present (example of IGV outputs shown in Appendix Figure B. 2). A further 50 TEs were visually validated across accessions, with 88.0% being called correctly in all accessions, with the majority (94.4%) of incorrect calls due to false positives.

The majority of these TEs (93.1%) identified in our accessions were fixed, resulting in 8,193 transposon insertion polymorphisms (TIPs). This accounted for only 0.90% of the TE sequences identified by EDTA, possibly due to most TEs being fixed and/or the inability of PopoolationTE2 to detect nested TEs. 82.4% (6,752) of the TIPs were present in the reference and 17.6% (1,441) were non-reference. The majority of the TIPs (5,328; 65.0%) were found at high frequency (≥75%). DNA transposons (4,169) were more abundant than retrotransposons (2,886), but Gypsy (1,001) and Copia (1,211) LTRs were the most abundant superfamily identified (Table B9).

To identify any significant association in the dataset between the tomato groups, TIP count, TE classification, genomic distribution, and population frequencies, χ^2 tests were performed for each pair of variables. TE order was significantly associated with genomic location ($\chi^2 = 83.12$, df=15, $p < 0.001$); LTRs and Helitrons were negatively and positively associated with the upstream region, respectively (Figure 3.9A). There was a weak positive correlation between TIP frequency and distance to the nearest gene (Spearman: $R = 0.15$, $p < 0.001$; Appendix Figure B. 3).Tomato group was significantly associated with population frequency ($\chi^2 = 805.65$, df = 4, $p < 0.001$) with TIPs in the domesticates and progenitor, negatively and positively associated with intermediate frequency, respectively (Figure 3.9B). Tomato groups were not significantly associated with TE order or genomic distribution. TIP frequencies were significantly associated with TE order ($\chi^2 = 1620.4$, df=10, $p < 0.001$) with a positive association of LTR at low frequency (Figure 3.9C). TIP frequencies were also significantly associated with genomic distribution ($\chi^2 = 182.27$, df = 6, $p < 0.001$). TIPs at low frequency were positively associated with genic regions, upstream and downstream of genes (Figure 3.9D).

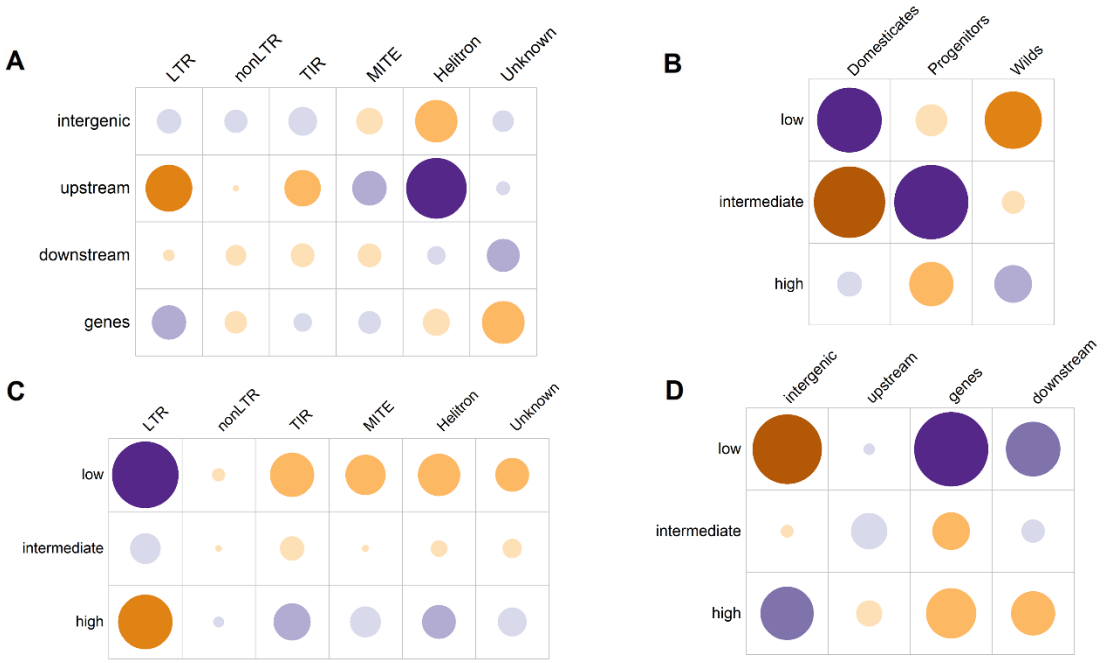


Figure 3.9: Significant association between tomato group, TE order, genomic distribution and population frequencies.

Each circle is proportional to the contribution to the $\chi 2$ test; purple depicts positive association and orange depicts negative association. There was positive association between (A) TE order x genomic, (B) tomato group x TIP population frequency, (C) TE order x TIP population frequency and (D) genomic distribution x TIP population frequency.

3.4.5.1 TIP count

The number of TIPs and their diversity in each tomato group may give an insight into the genetic variation of TEs in each species, which have been important in adaptation in early domestication. The greatest TIP count was in the progenitor than the wilds and domesticated species (Table 3.5). There was no significant difference in TIP count per accession between species (ANOVA: $F(3) = 1.222$, $p = 0.3276$; Figure 3.10A). To assess TE genetic variation between tomato groups, TE diversity was calculated for each species (Table B14). TIP diversity was significantly greater in the progenitor than in the wild species (two-tailed k-test: $p < 0.001$; Table 3.5). This may suggest high maintenance of genetic diversity in progenitors compared to the wild species.

Table 3.5: TIP counts, frequency and diversity estimates.

A. Transposon Insertion Polymorphism (TIP) count, diversity estimates and B. frequency for the domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and wild species (*S. cheesmaniae*: W_CHE; *S. galapagense*: W_GAL). C. To test for significant difference between the SNP diversity means using the variance of each estimate (V) and weighted mean diversity (x_w) and test statistic (k) is calculated. D. Pairwise comparison of π means between species.

A.	**Species**	**TIP count**	**TIP diversity**	
	Domesticates	7,090	1.29E-06	
	Progenitors	7,335	2.31E-06	
	Wild (CHE)	6,160	1.31E-06	
	Wild (GAL)	6,477	2.00E-06	
B.	**Species**	**low**	**intermediate**	**high**
	Domesticates	1,051	493	5,546
	Progenitors	662	1,507	5,166
	Wild (CHE)	0	1,014	5,146
	Wild (GAL)	0	2,323	4,154
C.	**Species**	**variance**	**weighted π**	**k**
	Domesticates	6.5E-16	2.0E+09	173.8
	Progenitors	1.4E-15	1.7E+09	331.7
	Wild (CHE)	1.7E-15	7.7E+08	59.5
	Wild (GAL)	1.7E-15	1.2E+09	80.9
D.	**Reference**	**Test**	**Normal deviate**	**P-value (2-tail)**
	Domesticates	Progenitors	-22.477	2.00E+00
	Domesticates	Wild (CHE)	-0.382	1.30E+00
	Domesticates	Wild (GAL)	-14.562	2.00E+00
	Progenitors	Wild (CHE)	17.948	0.00E+00
	Progenitors	Wild (GAL)	5.481	4.23E-08

TIPs were grouped into low (≤ 25%), intermediate (>25%, <75%), and high-frequency (≥75%) frequency TIPs to annotate TE insertions that are relatively recent. There was no significant species effect on TIP count, but TIP frequency had a significant effect on TIP count, with the majority of TIPs found at high frequency (two-way ANOVA: $F(2) = 6906.019$, $p < 0.001$; Figure 3.10B). There was a significant interaction effect on TIP count (two-way ANOVA: $F(6) = 22.806$, $p < 0.001$), with the progenitor having greater TIPs at an intermediate frequency than W_GAL (Tukey: $p < 0.001$).

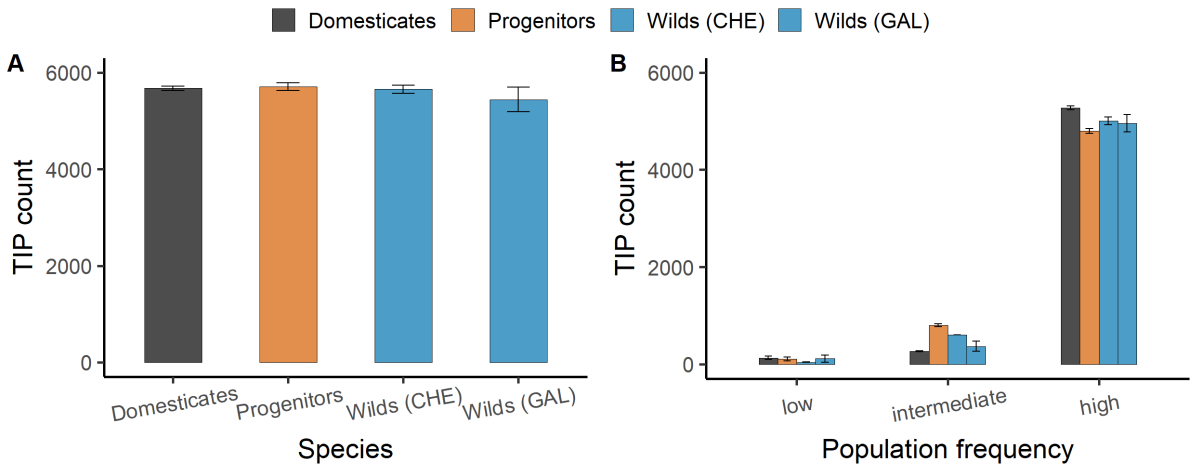


Figure 3.10: Transposon Insertion Polymorphism (TIP) count and frequency.
(A) TIPs count per species and (B) TIP frequency in each species was grouped into low (≤ 25%), intermediate (>25%, <75%), and high frequency (≥75%).

3.4.5.2 TE classification

The difference in TE groups between the progenitor and the wilds was assessed to examine if certain groups contribute to TIP diversity more than others. TIP diversity of retrotransposons was significantly greater in the progenitor than in both the wild species (two-tailed k-test: $p < 0.001$; Table 3.6; Table B14), however, TE diversity of DNA transposons was only significantly greater than W_CHE (two-tailed k-test: $p < 0.001$; Table 3.6; Table B14).

TIPs among TE order were similar between species (Figure 3.11A). There was no significant species effect (two-way ANOVA: $F(3) = 0.737$, $p = 0.532$), but there was a significant TE order effect (two-way ANOVA: $F(5) = 570.497$, $p < 0.001$) and a significant interaction between the effect of TE order and tomato group (two-way ANOVA: $F(15) = 2.022$, $p = 0.019$). TIP counts in each TE order were all significantly different from each other with the TIR as the most abundant (Tukey: $p < 0.05$). However, there was no significant difference between the progenitor and the wilds within each TE order (Tukey: $p > 0.05$). For TE families, there was no significant difference in the number of TE families between species (ANOVA: $F(3) = 2.129$, $p = 0.129$; Figure 3.11B).

Table 3.6: Retrotransposon and DNA transposon counts and diversity estimates.
A. Transposon Insertion Polymorphism (TIP) count for the domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and wild species (*S. cheesmaniae*: W_CHE; *S. galapagense*: W_GAL). B. TIP diversity estimates of retrotransposons and variance of each estimate (V) and weighted mean diversity (x_w) and test statistic (k) is calculated. C. Pairwise comparison of

retrotransposons diversity means between species. D. TIP diversity estimates of DNA transposons and variance of each estimate and weighted mean diversity and test statistic (k) is calculated. E. Pairwise comparison of DNA transposons diversity means between species.

A. Species	LTR	nonLTR	TIR	MITE	Helitron	Unknown	Total families
Domesticates	2,065	24	1,918	705	1,331	1,047	2,148
Progenitors	2,228	22	1,956	713	1,346	1,070	2,162
Wild (CHE)	1,515	21	1,762	663	1,236	963	2,087
Wild (GAL)	1,733	23	1,809	667	1,262	983	2,107

B. Species	retrotransposon	diversity	variance	weighted π	k
Domesticates	2,089	4.37 × 10−7	2.20E-16	1.99E+09	27.4
Progenitors	2,250	9.25 × 10−7	5.59E-16	1.66E+09	301.5
Wild (CHE)	1,536	2.98 × 10−7	3.85E-16	7.73E+08	122.5
Wild (GAL)	1,756	6.11 × 10−7	5.27E-16	1.16E+09	17.5

C. Reference	Test	Normal deviate	P-value (2-tail)
Domesticates	Progenitors	-17.488	2.00E+00
Domesticates	Wild (CHE)	5.677	1.37E-08
Domesticates	Wild (GAL)	-6.348	2.00E+00
Progenitors	Wild (CHE)	20.430	0.00E+00
Progenitors	Wild (GAL)	9.546	0.00E+00

D. Species	DNA transposon	diversity	variance	weighted π	k
Domesticates	3,954	6.66 × 10−7	3.35E-16	1.99E+09	105.7
Progenitors	4,015	1.10 × 10−6	6.63E-16	1.66E+09	89.9
Wild (CHE)	3,661	8.11 × 10−7	1.05E-15	7.73E+08	1.8
Wild (GAL)	3,738	1.07 × 10−6	9.24E-16	1.16E+09	50.9

E. Reference	Test	Normal deviate	P-value (2-tail)
Domesticates	Progenitors	-13.684	2.00E+00
Domesticates	Wild (CHE)	-3.897	2.00E+00
Domesticates	Wild (GAL)	-11.415	2.00E+00
Progenitors	Wild (CHE)	6.944	3.81E-12
Progenitors	Wild (GAL)	0.686	4.93E-01

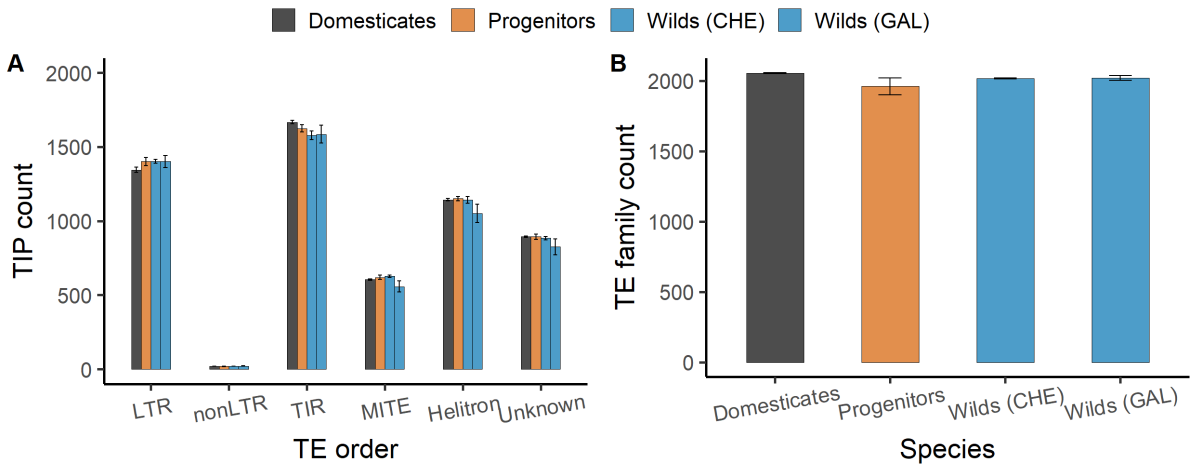


Figure 3.11: TE order and family between tomato groups.

Transposon Insertion Polymorphism (TIP) from each (A) TE order and (B) family.

3.4.5.3 TIP distribution throughout the genome.

The distribution of TIPs could indicate possible fitness effects as TEs near or in genes are more likely to affect gene function. The majority of TIPs were located in intergenic regions with 52.28% (4,283), followed by genic regions (18.64%; 1,527), upstream (16.44%; 1,347) and downstream (12.65%; 1,036), which significantly differs from what would be expected by chance ($\chi 2 = 4709.2$, $df = 3$, $p < 0.001$). This was most prominent with more TIPs found upstream of genes than expected by chance. 3,555 TIPs mapped to genes or were located in the upstream or downstream regions of genes and Gene Ontology (GO) analysis of these genes found two significantly enriched biological functions: response to stimulus and terpenoid metabolic process (Table B13).

There was no significant species effect but there was a significant genomic region effect (two-way ANOVA: $F(3) = 119.401$, $p < 0.001$), with the greatest number of TIPs found in the intergenic region (Figure 3.12A). There were also no significant genomic regions and species interaction (two-way ANOVA: $F(9) = 1.669$, $p = 0.110$). TIP diversity in near and genic regions was greater in the progenitor than in both the wild species (two-tailed k-test: $p < 0.001$; Table 3.7). This suggests greater TIP diversity in regions where phenotypic effects are greatly affected, which can be important in adaptation to changing environments.

Table 3.7: TIP counts across genomic regions and TIP diversity estimates near and within genic regions.

A. Genomic distribution of Transposon Insertion Polymorphism (TIP)in the domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and wild species (*S. cheesmaniae*: W_CHE; *S. galapagense*: W_GAL). B. TIP count and diversity estimates for each species based on near and

within genic regions. C. To test for significant difference between the TIP diversity means using the variance of each estimate and weighted mean diversity and test statistic (k) is calculated. D. Pairwise comparison of TIP diversity means between species.

A. Species	intergenic	up/downstream	genes
Domesticates	3,850	2,010	1,230
Progenitors	3,949	2,118	1,268
Wild (CHE)	3,397	1,720	1,043
Wild (GAL)	3,536	1,795	1,146

B. Species	TIP count	TIP diversity
Domesticates	3,240	2.51 × 10−6
Progenitors	3,386	4.57 × 10−6
Wild (CHE)	2,763	2.22 × 10−6
Wild (GAL)	2,941	4.03 × 10−6

C. Species	variance	weighted diversity	k
Domesticates	5.21E-15	4.82E+08	69.8
Progenitors	1.14E-14	4.01E+08	185.7
Wild (CHE)	1.18E-14	1.87E+08	68.1
Wild (GAL)	1.43E-14	2.81E+08	58.5

D. Reference	Test	Normal deviate	P-value (2-tail)
Domesticates	Progenitors	-15.969	2.00E+00
Domesticates	Wild (CHE)	2.254	2.42E-02
Domesticates	Wild (GAL)	-10.865	2.00E+00
Progenitors	Wild (CHE)	15.432	0.00E+00
Progenitors	Wild (GAL)	3.351	8.04E-04

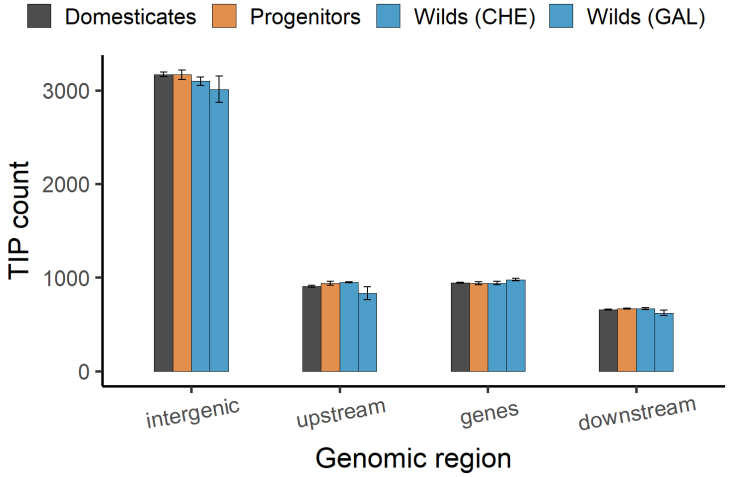


Figure 3.12: Genomic distribution of TIPs.

Transposon insertion polymorphisms (TIPs) per species in each genomic region.

3.4.5.4 Focus on TIPs potentially under selection.

Identification of TIPs under selection may give an insight into why the progenitor was selected for domestication and never-domesticated species were not. TIPs under selection would have either increased in frequency (selection for TE presence) or decreased in frequency (selection against TE presence) during domestication.

In the progenitor, there were 662 TIPs found at low frequencies (≤ 25%), 143 (1.95% of total TIPs in the progenitor) of which were found at high frequency (≥75%) in the domesticates; and over half of these (77) were found in genes or upstream/downstream of genes, this is more than expected by chance ($\chi 2$ = 68.302, df=1, p < 0.001). There are no significant GO terms associated with these 77 genes. Genes with TIPs upstream, downstream and within the genic region at low frequency in the progenitor and high frequency in the domesticates (n=77) were further explored. This included 26 TIPs upstream of genes, 23 TIPs downstream of genes and 28 TIPs within genic regions (Table B14). Many notable genes were linked to defense or stress response (Table 3.8). These TIPs may have neutral or beneficial effects on genes in their proximity resulting in their increase in frequency during domestication.

TIPs could also be selected against during domestication, resulting in high frequencies in the progenitor and low frequencies in the domesticates. In the progenitor, there were 5,166 TIPs found at high frequencies, 235 (3.20% of total TIPs in the progenitor) of which were found at low frequencies in the domesticates. Almost half of these, 110 TIPs, were found in genes or upstream/downstream of genes, this is more than expected by chance ($\chi 2$ = 65.234, df=1, p < 0.001). There were no significant GO terms associated with these 110 genes. Genes with TIPs upstream, downstream and within genic regions at high frequency in the progenitor and low frequency in the domesticates (n=110) were further explored (Table B15). This included 39 TIPs upstream of genes, 30 TIPs downstream of genes and 41 TIPs within genic regions. 14 of these were found at high frequency in P but absent in D. Notable genes include functions linked to plant and fruit development such as MADS-box transcription factor, class I heat shock gene, xyloglucan endotransglucosylase/hydrolase, SANT/Myb domain and YTH domain-containing protein (Table 3.8). These TIPs may have neutral or detrimental effect on genes in their proximity resulting in their decrease in frequency during domestication.

Table 3.8: Genes with TEs putatively selected during domestication.

Genes with transposon insertion polymorphisms (TIPs) within or upstream/downstream putatively selected for (A) and against (B) during domestication (full list on Table B14 and B15).

TE order	TE location	gene ID	gene description	role	reference
A. Genes with TEs putatively selected for during domestication					
LTR	downstream	Solyc09g008505	Protein LAZ1-like protein 1	programmed cell death and defense response against pathogens	Malinovsky *et al.* (2010)
LTR	downstream	Solyc10g054770	Small auxin up-regulated (SAUR) RNA81	auxin-responsive genes	Wu *et al.* (2012)
TIR	gene	Solyc04g072010	Carboxypeptidase D	wound response	Moura *et al.* (2001)
TIR	gene	Solyc12g009680	Heptahelical transmembrane protein 1	cold acclimation response	Wang *et al.* (2019a)
LTR	genes	Solyc02g091810	bHLH transcription factor 015	pathogen interaction, fruit development and ripening in tomato	Wang *et al.* (2015a); Sun *et al.* (2015); Zhang *et al.* (2020)
B. Genes with TEs putatively selected against during domestication					
Unknown	upstream	Solyc01g087990	MADS-box transcription factor	reproductive growth	Wang *et al.* (2019b)
TIR	upstream	Solyc06g076570	Class I heat shock protein	tomato fruit ripening	Shukla *et al.* (2017)
LTR	downstream	Solyc07g056000	Xyloglucan endotransglucosylase/hydrolase	tomato fruit ripening	Muñoz-Bertomeu *et al.* (2013)
Unknown	genes	Solyc03g119050	SANT/Myb domain	tomato fruit development	Barg *et al.* (2005)
TIR	genes	Solyc08g007730	YTH domain-containing protein	tomato development and fruit ripening	Yin *et al.*, 2021 Yin *et al.* (2021)

3.5 Discussion

The contribution of genetic diversity in early domestication, especially transposable elements (TEs), is generally unknown. In this study, comparative assessments of SNP and TIP diversity

and distribution between the tomato crop progenitor and other never-domesticated wild species gives us an insight into possible advantages of the progenitor in early domestication.

The *Solanum lycopersicum* genome was annotated using the EDTA pipeline, revealing the genome is made up of 64.14% transposable elements (TEs), which is similar to the findings of other studies (The Tomato Genome Consortium, 2012; Oliver *et al.*, 2013; Mehra *et al.*, 2015; Su *et al.*, 2021). Higher proportions of DNA transposons (TIR, MITE and Helitron) were identified compared to previous reports (The Tomato Genome Consortium, 2012; Oliver *et al.*, 2013). The EDTA pipeline uses structure-based identification of these TEs using specific annotation tools for structurally intact elements that can reveal novel TEs.

3.5.1 Greater diversity of SNPs and TIPs in the progenitor than the never-domesticated wilds

Genetic diversity estimates were measured using nucleotide diversity (π) and Watterson's estimator of theta (θ_w). Our estimates for the domesticated species resulted in 2.76×10^{-3} for SNP diversity and 3.48×10^{-3} for Watterson's theta. These are comparable with nucleotide diversity range of 3.00×10^{-5} to 3.23×10^{-3} from previous studies in tomato (Lin *et al.*, 2014; Sauvage *et al.*, 2017; Razifard *et al.*, 2020). We found the progenitor (5.66×10^{-3}) had higher nucleotide diversity than both wild species, *S. cheesmaniae* (3.32×10^{-3}) and *S. galapagense* (2.42×10^{-3}). However, with Watterson's estimator of theta, the progenitor (5.12×10^{-3}) was only greater than *S. galapagense* (4.03×10^{-3}) but not *S. cheesmaniae* (7.17×10^{-3}). This difference between nucleotide diversity and Watterson's theta has also been reported previously by Razifard *et al.* (2020). The difference in nucleotide diversity and Watterson's theta is central to the calculation of Tajima's D, which detects deviations from neutral theory of evolution (Tajima, 1983). Greater nucleotide diversity than Watterson's theta suggest a positive Tajima's D value which indicate balancing selection due to the long term maintenance of distinct haplotypes resulting in an excess intermediate frequency SNPs (Yamasaki *et al.*, 2007). On the other hand, greater Watterson's theta than nucleotide diversity suggest a negative Tajima's D value, which could imply a genetic bottleneck which has been reported in tomato (Koenig *et al.*, 2013) and other domesticated species (Ndjiondjop *et al.*, 2019). For wilds, both of these species are endemic to the Galapagos islands (Pailles *et al.*, 2017) and have experienced strong bottlenecks due to island colonisation and recent adaptation (Nuez *et al.*, 2004) resulting in the increased accumulation of potential deleterious mutations (Koenig *et al.*, 2013). This genetic bottleneck is possibly the reason for the low genetic diversity in the wild species, highlighting the importance of exploring other never-domesticated wild species.

This high nucleotide diversity in the progenitor was also supported by our TIPs analysis with 2.31 x 10^{-6} compared to the wild species 1.31 x 10^{-6} and 2.00 x 10^{-6} respectively. Greater diversity of TIPs In the tomato progenitor has also been characterise by Dominguez *et al.* (2020) with their annotation of TIPs in 602 accessions of domesticated and wild tomatoes resulting in the progenitor having the richest genetic diversity based on the number of TE families than the other wild species. Similarly, the greatest TE family count was found in the progenitor, but this was not statistically different from the wild species.

TIP diversity is much lower than SNP diversity, this may be due to fewer TIPs (8,193) detected compared to SNPs (13,367,384). Compared to Dominguez *et al.* (2020) annotation of 6,906 non-reference TIPs, our analysis managed to capture an equivalent of 20.9% (1,441 non-reference TIPs), with only 4.0% of the accessions used. Since, most of the TIPs found by Dominguez *et al.* (2020) were only present in one or a few accessions, increasing the number of accessions in our study would increase the number of TIPs detected. The majority of non-reference TIPs detected by Dominguez *et al.* (2020) were from Copia and Gypsy, similar to our findings, but MITEs were not included in their analysis. The lower number of TIPs detected could also be due to the following reasons: (i) many TEs identified as fragmented, likely due to degradation as mutations accumulate (Bourque *et al.*, 2018); (ii) strict filtering steps to avoid false positives; (iii) many TE insertions were fixed in the accessions sampled; (iv) the inability of PopoolationTE2 to detect nested TEs (Kofler *et al.*, 2016). Despite limitations in the number of accessions, we still found a consensus of greater diversity of SNPs and TEs in the progenitor compared to never-domesticated wilds.

As discussed above, genetic diversity of the progenitor suggests a balancing selection within the population. Balancing selection is the maintenance of advantageous genetic diversity within populations by natural selection (Bitarello *et al.*, 2023). In *Arabidopsis*, balancing selection has been shown to contribute to adaptation to diverse habitats, with evidence of long-term balancing selection on genes involved in response to biotic and abiotic stress (Wu *et al.*, 2017). The role of maintaining genetic variation by balancing selection has also been reported in maize and its progenitor, teosinte (Chen *et al.*, 2020).

Maintenance of genetic diversity is important for greater availability of allelic variation to facilitate adaptation, this is greatly dependent on mutation rates and effective population size (Nei and Li, 1979). Therefore, greater genetic diversity, both in SNPs and TIPs, could suggest higher mutation rates or higher effective population size. Genetic diversity is positively correlated with mutation rates in eukaryotes (Wang and Obbard, 2023), and therefore could drive the emergence of new variants in the population. Additionally, large effective population size allows new mutation to persist in the population as genetic drift is weaker and therefore

maintains genetic diversity over time (Wang *et al.*, 2016). New nucleotide or TEs polymorphisms provide the raw material for natural selection. Species with greater genetic variability are more likely to include individuals with beneficial traits important to environmental changes such as new growing conditions (Fustier *et al.*, 2019), new pathogens (Gladieux *et al.*, 2024) and climate change (Cortés and López-Hernández, 2021). These are changes that may have been selection pressure among competing progenitor and wild populations.

Estimates of mutation rates in plants are predominantly characterised with single nucleotide mutations (SNMs) in several plant species (Yang *et al.*, 2017; Sandler *et al.*, 2020; Monroe *et al.*, 2022; Ichikawa *et al.*, 2023). TE transposition rates have only been characterised for specific TE families (Alleman and Freeling, 1986; Nakazaki *et al.*, 2003; Tsukahara *et al.*, 2009; Vukich *et al.*, 2009), however, whole-genome transposition rates in plants have yet to be reported (see Chapter 4). Changes in environmental conditions can trigger an increase in mutation rates, such as bursts in TE activity linked to periods of stress increasing genetic variability for selection to act upon (Schrader and Schmitz, 2019). Benoit *et al.* (2019) demonstrated that the TE Rider is drought-inducible through MYB transcription factors involved in drought-stress response, highlighting the potential for TEs to transpose under stress. Other stress conditions such as salt and cold stresses have been reported to activate *mPing* DNA transposon in rice (Naito *et al.*, 2009; Yasuda *et al.*, 2013) and heat stress activates ONSEN retrotransposon in Arabidopsis (Cavrak *et al.*, 2014). The movement of crop progenitors from wild to novel environments could have also induced the activation of TEs, resulting in allelic variation in stress response facilitating plant adaptability through genetic diversity initiated by TEs.

3.5.2 Greater diversity of SNPs and TIPs near and within genic regions in the progenitor than never-domesticated wilds

The genomic distribution of SNPs and TIPs were explored to infer the proximal effect of these polymorphisms on nearby genes. The majority of SNPs and TIPs were found in intergenic regions, supported by previous studies (Mehra *et al.*, 2015; Alonge *et al.*, 2020; Dominguez *et al.*, 2020). This is a common strategy of TEs to have little or no effect on gene function allowing their retention in the genome from selective forces (Bourque *et al.*, 2018). This was further supported by a weak but significant positive correlation between TIP frequency and distance to the nearest gene suggestive of the location of high frequency TIPs further away from genic regions. These TEs are maintained by genetic drift or they could be eliminated through selection and by recombination (Capy, 2021). Over time, TEs can get degraded by point mutations that accumulate neutrally, this also makes it more difficult to identify older TE insertions (Lisch, 2013). We found almost half of TIPs were near or within genic regions, greater than the one-third

of the total TEs detected in previous reports (Mehra *et al.*, 2015; Alonge *et al.*, 2020; Dominguez *et al.*, 2020). This suggests that polymorphic TEs are more likely to be near or within genic regions.

The overall distribution of TIPs differed from that expected, with more TIPs than expected identified upstream of genes. Similar enrichment of TEs proximal to genes was reported in *Capsella rubella* leading to expression changes in adjacent genes (Niu *et al.*, 2019). Increase in TIPs near genic regions could also be due to genetic hitch-hiking on other genetic variation or background selection (Stephan, 2010). We found that genes associated with TE insertions were enriched in response to stimulus and terpenoid metabolic process. Previous studies have shown that genes near TE insertions were enriched in GO terms associated with defence response in diverse tomato species (Dominguez *et al.*, 2020) and *Arabidopsis* accessions (Baduel *et al.*, 2021), indicative of a potential contribution of TE polymorphisms to local adaptation. Terpenoids in tomatoes have roles in fruit aroma and flavour (Lewinsohn *et al.*, 2001), defense mechanisms (Besser *et al.*, 2009) and stress response (Reimer *et al.*, 2021). TEs near or within genes associated with these traits would have been important during domestication.

Adaptive genetic variation affects fitness and are therefore more likely to be found near or within genic regions. We found greater SNP and TIP diversity near and within genes in the progenitor than the other wild species. These could be mostly neutral genetic variation that can become adaptive in changing environments (Barrett and Schluter, 2008; Olson-Manning *et al.*, 2012). Enhanced diversity near genes in the progenitor could have a number of phenotypic effects such as generation of phenotypic variation through loss of function mutations (Monroe *et al.*, 2022), and changes in gene expression through various mechanisms (Hirsch and Springer, 2017). These provide a rich pool of genetic variation for selection to act upon, especially in early domestication where changes in environmental conditions were frequent (Wood and Lenné, 2018). TIPs are more likely to affect phenotypes, with larger effects compared to SNPs (Uzunović *et al.*, 2019; Dominguez *et al.*, 2020) and therefore are vital sources of phenotypic variation when trying to understand the evolutionary advantages of the progenitor over never-domesticated wilds.

Greater standing genetic variation provides a population with increased potential for beneficial mutation readily available in new environments, without the need for new mutations to arise, leading to faster evolution (Barrett and Schluter, 2008). In maize, TEs have been associated with drought tolerance, adaptation to long-day environments and higher latitudes (Mao *et al.*, 2015; Yang *et al.*, 2013; Huang *et al.*, 2018). TEs can also aid in reprograming stress gene networks by introducing regulatory sequences influencing transcriptional networks (Cowley and Oakey,

2013). Therefore, TEs can affect the genetic and phenotypic plasticity of the progenitors. Plasticity is another mechanism that can contribute to an increase in variability within the population that selection can act upon, which has been linked to the genomic advantage of the tomato progenitor (see Chapter 2).

SNPs and TIPs near or within genic regions are more likely to experience selective sweeps, allowing the spread of beneficial mutations quickly throughout the population (Messer and Petrov, 2013). Populations in early domestication were thought to be large and inter-connected (Allaby *et al.*, 2019; Alam and Purugganan, 2024), this favours soft selective sweep where one trait have multiple adaptive alleles within a species (Olson-Manning *et al.*, 2012). Evidence of adaptive alleles from standing variation reported in plants for plant height in wheat (Raquin *et al.*, 2008), reduction of branching in pearl millet (Remigereau *et al.*, 2011), control of growth habit in soybean (Zhong *et al.*, 2017) and seed colour in amaranth (Stetter *et al.*, 2020). Even in small populations, greater standing variation can buffer against genetic drift ensuring beneficial alleles are less likely to be lost, maintaining the ability of the population to adapt to novel conditions (Olson-Manning *et al.*, 2012).

3.5.3 TIPs under selection during tomato domestication

We identified TIPs that were possibly under selection during domestication, showing a strong frequency difference between the progenitor and the domesticates. These TEs putatively selected during domestication were more abundant near and within genic regions than expected by chance. The fate of a TE insertion is influenced by a number of factors, such as genomic alterations due to recombination and rearrangements, the impact of the insertion and the impact of epigenetic control of the TE insertion (Capy, 2021).

If a TE insertion confers adaptive value, it may be selected on and reach higher frequencies, conversely, the decrease in the frequency of TIPs during domestication may be due to negative or purifying selection (Barrón *et al.*, 2014; Stapley *et al.*, 2015). We identified many TIPs that increased and decreased in frequency during domestication, associated with gene functions related to biotic stress resistance and abiotic stress tolerance. During early domestication, changes in growing conditions and population size may have triggered selection and a bottleneck that resulted in the changes in frequency of TIPs. Under selection, if an allele with a TE confers an advantage it would increase in frequency; conversely, if an allele without a TE confers an advantage, then this allele would increase in frequency. A bottleneck would change the frequency of alleles by chance if TEs were approximately neutral. Changes in TIP frequency have been reported in rice during domestication, identifying positive selection of some TIPs

(Castanera *et al.*, 2023). Deleterious TEs, such as those with a negative impact on the host's fitness, are maintained at low frequencies or eliminated in the population. Deleterious TIPs could also be eliminated through recombination. Recombination lowers TE frequencies as a result of complete or partial loss of a TE sequence leading to its immobilisation (Capy, 2021).

More TIPs were selected against compared to TIPs that were selected for, suggestive of genetic polymorphisms that were lost during domestication. Many of these TIPs were associated with genes linked to biotic stress resistance and abiotic stress tolerance. This highlights the number of genetic variants in the progenitor with the potential to be adaptive, that can be utilised in improving genetic diversity in crops, which can be valuable for crop improvement and breeding.

3.5.4 Limitations

PopoolationTE2 tried to address the problems introduced by different sequencing library preparation such as differences in insert sizes, coverage heterogeneity and genome sizes that may present biases in TIP identification (Kofler *et al.*, 2016). This has performed well in a previous benchmarking analysis (Vendrell-Mir *et al.*, 2019) and has been effectively used in several studies (Castanera *et al.*, 2021; Castanera *et al.*, 2023). Nonetheless, we revealed limitations which restricted the number of samples in our analysis. Collection of resequencing data from various studies included samples with variation in read lengths and inner distance resulting in some samples diverging from the optimal conditions to run PopoolationTE2. This led to many samples being excluded from our analysis. Other researchers should take special care to ensure samples have similar read lengths and inner distance to avoid false negative TE identifications and increase the accuracy of TE positions (Kofler *et al.*, 2016).

Our study was limited by the number of progenitor and wild accessions included in the analysis; there are still limited genetic resources for many wild relatives of crops, hindering their genomic characterisation. The inclusion of wild genome references through the use of pan genomes would better improve TE annotation of wild genomes to guarantee a sufficient comparison in TEs (Li *et al.*, 2023), however, would limit the ability to compare across species. The use of additional TE detection tools, such as Jitterbug (Hénaff *et al.*, 2015) and TEfinder (Sohrab *et al.*, 2021), could confirm patterns identified in our studies. There are also alternative methods to analysing TE activity such as mutation accumulation (MA) experiments that estimate mutation rates (see Chapter 4). These have been shown to capture differences in the mutation rate of both SNPs and transposons (Lu *et al.*, 2021). Change in the frequency of TIPs could be further analysed with methods such as the Population Branch Statistic (PBS) which analyses strong differentiation in population frequencies (Yi *et al.*, 2010).

We acknowledge that the populations of tomato crop relatives, *S. pimpinellifolium*, *S. cheesmaniae* and *S. galapagense* used are not identical to those that existed around the time of early domestication due to subsequent selection and gene flow in the wild (Flint-Garcia *et al.*, 2023). Increased transposition rate may have aided the domestication of *S. pimpinellifolium*, however, due to *S. cheesmaniae* and *S. galapagense* geographical location in the Galapagos Islands, these species would not have competed in early domestication. Therefore, exploration of other never-domesticated species, as well as additional accessions would broaden our understanding of the role of TEs in early tomato domestication. However, as was shown, these were not comparable in our pipeline due to the reduced mapping efficiency.

3.6 Conclusion

This study explored the level of genetic variation, both in SNPs and TIPs, across domesticates, progenitors and never-domesticated tomato accessions. We found SNP and TIP diversity across the genome to be generally greater in the progenitor than the never-domesticated wilds. This was also the case near and within genic regions. Greater genetic diversity may be linked to higher mutation rates that translate into greater phenotypic variability in the progenitors. This is indicative of the role of nucleotide and TE genetic variation in early domestication, which have been important in adaptation to changing growing conditions, allowing faster or more efficient domestication of the tomato progenitor. TIP analyses in crops are important in identifying phenotypic variation and its role in evolution, revealing TE-associated genes with the potential to uncover adaptive variation that can be exploited for crop improvements to aid in aiding food security in current and future climates.

Chapter 4 The role of mutation rate on tomato domestication

4.1 Abstract

Only a small proportion of edible plant species were domesticated, we explore mechanisms that could have given crop progenitors an advantage in early domestication. Novel phenotypes are expected to arise from novel mutations such as single nucleotide variants (SNVs), indels (insertions and deletions) and transposable elements (TEs), with their ability to move from one location in the genome to another. However, there are no estimates of mutation rates in tomatoes and little is known about their influence in the domestication of certain species in early domestication. Here, we explore whether faster mutation rates could have played a role in aiding the domestication of the tomato progenitor, *Solanum pimpinellifolium* than the never-domesticated wild species, *S. cheesmaniae*. A mutation accumulation (MA) pot experiment was set up, where plants were selfed for four generations. Higher SNV and indel rates (per site per generation) were estimated in the progenitor (14.99×10^{-9}; 2.47×10^{-9}) than in the wilds (4.01×10^{-9}; 0.62×10^{-9}), but not the domesticated species (10.47×10^{-9}; 2.31×10^{-9}). There was no significant difference in TE insertion rate between domesticated (0.92×10^{-9}), progenitor (2.05×10^{-9}) or wild species (0.92×10^{-9}). This is the first report of mutation rates in tomatoes. Mutations identified in the progenitor were more likely to be found in the chromosome arm (than pericentric regions) and up/downstream of a gene than would be expected by chance. Mutations in or near genes could have greater impact on traits and therefore could result in novel phenotypes for selection to act on. This faster mutation rate in the progenitor could have aided the domestication of progenitors in early cultivation.

4.2 Introduction

Crop domestication is an important process that paved our transition from foraging in the wild into cultivating various crops for agriculture. This is an evolutionary process that transforms the wild progenitor into cultivated crops, vital for human consumption and/or use (Purugganan, 2019). Given our reliance on domesticated crops, it is curious that we rely on only a handful to provide 90% of the world's calories (FAO, 2017). Additionally, only a few hundred plant species have been domesticated (Zeven and De Wet, 1982), out of the roughly 200,000 edible plant species (Willis, 2017). Is there genomic constraint on domestication that allows the domestication of certain species over others? Uncovering genomic mechanisms that facilitate domestication could identify adaptive variation in crop wild relatives that could be important in the development of novel crops to tackle current and future food security.

In early domestication, the transition in growing conditions from natural environments to human-modified fields brought a lot of changes such as changes in soil quality, competition and other biotic and abiotic stresses. For example, early anthropogenic environments are thought to be disturbed by seasonal fires and flooding (Wood and Lenné, 2018). The emergence of beneficial traits or the presence of greater variation for selection to act on could have been an advantage in early domestication. There are many domestication traits as a result of new mutations (Meyer and Purugganan, 2013). During domestication, these mutations were selected over time (Wright *et al.*, 2005). Faster mutation in progenitors may mean faster emergence of beneficial phenotypes such as those mentioned above (Romero et al., 2025).

The benefits of a higher mutation rate to the speed of adaptation have been extensively studied in micro-organisms (Desai and Fisher, 2007; Desai *et al.*, 2007; Wiser *et al.*, 2013). Although a faster mutation rate produces more advantageous mutations it also generates more deleterious mutations (Eyre-Walker and Keightley, 2007). The effects of these might have been attenuated during the domestication process as purifying selection would be more relaxed therefore allowing the persistence of populations even with increased mutation load as reported in several domesticated species (Smith *et al.*, 2019; Renaut and Rieseberg, 2015; Wang *et al.*, 2021a).

There are a variety of different types of mutations from the change of a single nucleotide (single nucleotide variants or SNVs) to insertions, deletions (InDels) and rearrangements, all of which might be potentially advantageous. Amongst these transposable elements (TEs) might be a

major source of genetic variation. TEs are mobile DNA elements able to move from one location in the genome to another (Bourque *et al.*, 2018).TEs are also known drivers of genome plasticity (Lanciano and Cristofari, 2020), with stress-induced TE activation increasing the ability of the genome to respond flexibly to varying conditions through various mechanisms (Pimpinelli *et al.*, 2019). Rates of mutation along with TE composition can enable an increase in adaptation and diversification (Cobben *et al.*, 2017). TEs are a significant component of crop genomes, because of their abundance and their effect on genes (Gill *et al.*, 2021). TE insertions create mutations generating variation and increasing genetic diversity, valuable in understanding domestication and plant adaptation to new environments. TE insertions have been linked to the emergence of new varieties in crops such as the winter variety of rapeseed oil (Yin *et al.*, 2020), glutinous rice (Wei and Cao, 2016) and elongated tomato fruits (Xiao et al., 2008). The amounts of genetic variation but at the molecular and phenotypic value that is due to TEs is unknown.

Mutation rates of SNVs and indels have been characterised in several plant species such as *Arabidopsis* (Monroe *et al.*, 2022), maize (Dooner *et al.*, 2019), rice (Suganami *et al.*, 2024) and duckweed (Sandler *et al.*, 2020), using various techniques. Wang and Obbard (2023) performed a meta-analysis of mutation rates experiments in eukaryotes covering 134 species, from unicellular eukaryotes to plants and mammals. They found higher mutation rates in species with longer generation times, larger genomes and smaller effective sizes. Mammals and plant species have higher mutation rates than arthropods and unicellular eukaryotes (Wang and Obbard, 2023).

Transposition rates are predominantly studied in unicellular organisms (Sousa *et al.*, 2013; Hénault *et al.*, 2020) or multicellular organisms such as *Caenorhabditis elegans* (Bégin and Schoen, 2006) and *Drosophila melanogaster* (Adrion *et al.*, 2017; McCullers and Steiniger, 2017). TE insertion rates (per site per generation) can differ between different types of TEs in *Escherichia coli* from 4.0×10^{-8} to 1.15×10^{-5} (Sousa *et al.*, 2013). Transposition rates can also differ between populations in *D. melanogaster* with European populations having higher transposition rates than West African with 23.36×10^{-5} and 8.99×10^{-5} TE copies per generation respectively (Wang *et al.*, 2023b). The overall rate of TE insertion(per site per generation) was 4.93×10^{-9}, this was higher than the single nucleotide mutation rate estimate (SNVs and indel) of 3.30×10^{-9} in *D. melanogaster* (Wang *et al.*, 2023b). This shows the variability of TE insertion rates in the organisms that have previously been reported.

For direct estimates of per-generation mutation rates, mutation accumulation (MA) experiments are a common method. This constitutes of a single inbred or asexual genome accumulates spontaneous mutations over multiple generations, minimising the effectiveness of natural selection (Halligan and Keightley, 2009). In plants, MA experiments have been performed on the model organism *Arabidopsis thaliana*, for estimates of mutation rates based on SNVs and indels (Ossowski *et al.*, 2010; Weng *et al.*, 2019; Lu *et al.*, 2021). Other SNV estimates have also been reported for Chinese cabbage (Park *et al.*, 2019) and two species of duckweed (Sandler *et al.*, 2020). No estimates of transposition rates from MA experiments have been reported in plants. Combined with whole-genome sequencing (WGS), this experimental system can capture extensive mutation rates and distribution patterns. This can be employed to investigate the transposition rate in progenitors and never-domesticated species to assess mutation rates of different types of mutations among these species.

Tomato (*Solanum lycopersicum* L.) is an important crop species widely consumed, with world production reaching over 180 million tons in 2019 (FAO, 2021). It is estimated to have been domesticated around 7,000 years ago (Razifard *et al.*, 2020). TE transposition has been linked to several phenotypic changes in different tomato varieties (Gilbert and Feschotte, 2018). The LTR retrotransposon, *Rider*, has been associated with tomato phenotypes: 'Roma' variety with elongated fruits has a *Rider*-mediated *SUN* gene duplication (Xiao *et al.*, 2008); yellow flesh tomato has a *Rider* disrupted *PSY1* gene (Fray and Grierson, 1993); jointless fruit stem is linked to a *Rider* insertion in the *J2* gene (Soyk *et al.*, 2017). This showcases the diversity of TE insertion consequences on gene function.

In this study, we explore mutation rates in domesticated (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and never-domesticated wild (*S. cheesmaniae*) tomato species. Here we investigate whether the progenitor of domesticated tomato has a higher mutation rate for SNVs and indels, or a higher rate of transposition, both across the genome and in regions that might potentially have a phenotypic effect (i.e. regions of the genome close to or in genes). A mutation accumulation experiment was set up including accessions from different tomato species. These were grown and selfed for four generations and SNVs, indels and TEs were characterised to estimate mutation rates and their distributions were explored.

4.3 Materials and Methods

4.3.1 Plant material and growth conditions

Tomato accession seeds were obtained from Tomato Genetic Resource Centre (TGRC; https://tgrc.ucdavis.edu/) and Centre for Genetic Resources the Netherlands (CGN) Wageningen University (https://cgngenis.wur.nl/) and were grown in the glasshouses at the University of Southampton. Seeds were sown in large pots with a mix of 2:1 Levington compost (F2+sand) and vermiculite. Fruits from this generation were harvested; seeds were cleaned and stored. Seed cleaning was carried out as follows: seeds were scooped out from the fruits into a beaker with 3HCl and left for 30 minutes and then rinsed; seeds were then placed back into the beaker with 10% Trisodium phosphate (TSP) for 20 minutes and then rinsed; for drying, seeds were placed in a drying oven at 40°C overnight. Cleaned seeds were stored at 4°C until needed. Two accessions of each species were grown by selfing for further five generations in the same soil mixture, considered as G0 to G4 (Figure 4.1). MA lines from the same accessions shared the same G0 and G1 ancestor. Note that one accession of *S. cheesmaniae* (SC_A) did not produce any fruits after G1 and is therefore excluded from the analysis. Other never-domesticated wild species were not included in the experiment due to a lack of fruiting or lengthy periods of fruit emergence. The metadata for this experiment is provided in Table C1.

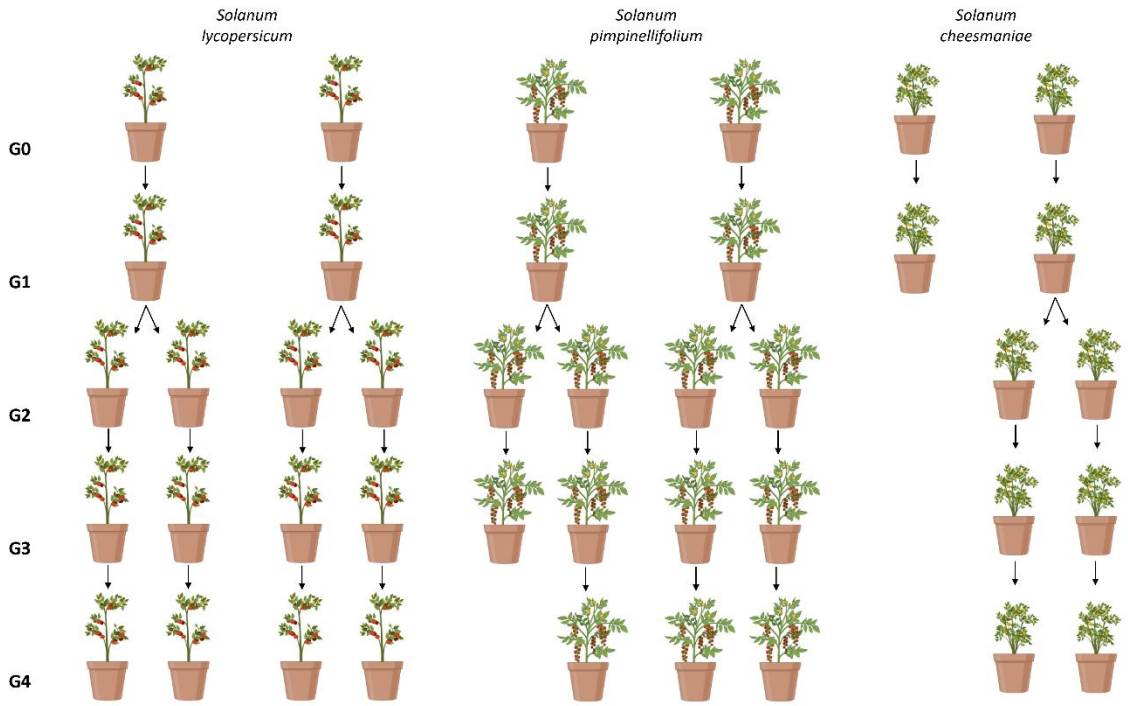


Figure 4.1: Illustration of mutation accumulation (MA) experimental design.
MA experiment to estimate mutation rate between generation zero (G0) and four (G4) in *S. lycopersicum* (domesticated), *S. pimpinellifolium* (progenitor), *S. cheesmaniae* (never-domesticated wild).

4.3.2 DNA extraction and sequencing

Leaf samples were taken from each plant (six accessions) at G0 and two replicates from each accession at G4. Due to lack of fruiting, only three and two G4 samples were taken from *S. pimpinellifolium* and *S. cheesmaniae*, respectively. DNA extraction was performed using a modified CTAB protocol (Doyle and Doyle, 1990). Extracted DNA was then sent to Novogene Bioinformatics Institute (Cambridge, UK); library preparation of the samples was performed, followed by 150 bp paired-end sequencing (with 350 base insert size) using Illumina NovoSeq 6000 (Illumina, USA).

4.3.3 Processing of sequencing data

Raw WGS resequencing data were trimmed and filtered using Trimmomatic v0.32 (Bolger *et al.*, 2014) to remove poor-quality bases and reads: Illumina adapters were removed; leading and trailing bases with quality below 5 were removed; reads shorter than 72bp were removed; sliding window trimming was performed for a window size of 4 with the required quality of 20. Accessions were aligned to the *S. lycopersicum* SL4.0 reference (Hosmani *et al.*, 2019) using Bowtie2 (--phred33) v2.2.3 (Langmead and Salzberg, 2012), to check alignment rate. To account for the difference in read number and sequencing depth, samples with clean reads >80M were subsampled to 80M reads.

4.3.4 Detection of short variants

Variant calling was performed using bcftools workflow (Danecek *et al.*, 2021). Bam files after alignment to the reference genome were processed with Picard v2.18.14 (http://broadinstitute.github.io/picard/) and combined to make a VCF file. Short variants were called from the VCF file using bcftools call (Li *et al.*, 2009). For Principal component analysis

(PCA), short variants were filtered using bcftools view with the following criteria: QUAL>30, DP>20, AD > 10, ADF > 5, ADR > 5, ", -m2 -M2, and LD-pruning was performed with plink v1.9 (Chang *et al.*, 2015) with a sliding window of 50 single nucleotide polymorphisms (SNPs), removing one in a pair of SNPs when LD > 0.1. Eigenvec and eigenval values were calculated by plink v1.9 (Chang *et al.*, 2015).

Single nucleotide variants (SNVs) between G0 and G4 pairs were filtered using bcftools view with the following criteria: QUAL>30, DP>20, AD > 10, ADF > 5, ADR > 5, GT="0/0", GT="1/1", -m2 -M2, -v snps. Indels were filtered with the following criteria: QUAL>30, DP>20, AD > 10, ADF > 5, ADR > 5, GT="0/0", GT="1/1", -v indels. Only homozygous calls were retained to avoid somatic mutations and false positive calls, as per previous MA experiments (Ossowski *et al.*, 2010; Weng *et al.*, 2019; Lu *et al.*, 2021).

4.3.5 Whole genome TE annotation

For whole-genome TE annotation, the Extensive de-novo TE annotator (EDTA) pipeline (Ou *et al.*, 2019) was used. The EDTA pipeline annotates TEs in a reference genome, identifying LTR-retrotransposons, TIR transposons, MITEs and Helitrons that use structural (LTRharvest, LTR_FINDER, LTR_retriever, TIR-Learner, HelitronScanner) and homology (RepeatModeler and RepeatMasker) approaches to produce a comprehensive TE library. The *Solanum lycopersicum* reference SL4.0 genome was supplied to EDTA v2.0.0, plus ITAG4.0 coding regions and gene positions of the SL4.0 genome assembly to avoid gene sequences being added to the TE library. TEs were classified at order, superfamily and family level using Wicker *et al.* (2007). R v4.1.2 (R Core Team, 2021) was used to illustrate TE distribution. For the full EDTA results, see Chapter 3 (3.4.1).

4.3.6 Detection of transposon events

TE detection in samples was performed using PopoolationTE2 v1.10.03 (Kofler *et al.*, 2016). A TE-merged reference genome was created by combining the masked reference genome and the TE library extracted from the TE annotation output of the EDTA pipeline. Then, a TE hierarchy file was created using the entries from the TE annotation (from the EDTA pipeline) to identify the ID, order and family information of each TE sequence. Paired-end reads for each sample after

Trimmomatic filtering were used for this analysis. These were mapped separately to the TE-merged-reference with bwa-bwasw v0.7.17 (Li and Durbin, 2009) and then PopoolationTE2 *se2pe* restored the paired-end information. PopoolationTE2 *ppileup* then used the bam files to produce a ppileup file with a minimum mapping quality of 20. Kofler *et al.* (2016) recommends subsampling by physical coverage to reduce false positives, therefore the ppileup file was subsampled to physical coverage of at least 2 to only analyse regions with sufficient physical coverage in all samples; this resulted in roughly 24.33% of the sites excluded from the analysis. To identify TE insertions, the following parameters were used: (i) identifySignatures (–mode joint, –min-count 2, --signature-window fix500), (ii) frequency (default parameters), (iii) filterSignatures (--min-count 2 --max-otherte-count 2, --max-structvar-count 2) and (iv) pairupSignatures (default parameters) as applied by Castanera *et al.* (2023). Additional TE filtering was performed to avoid false positive calls: (i) TE insertions with missing data were removed; (ii) TEs with a zygosity < 0.25 in all samples were removed; (iii) TEs with zygosity of ≤ 0.05 were scored as absent; (iv) TEs with zygosity of ≥ 0.70 were scored as present as recommended by Vendrell-Mir *et al.* (2019).

4.3.7 Validation of mutation calls

SNVs, indels and TE mutations detected in each accession were inspected visually using IGV v1.12.0 (Thorvaldsdóttir *et al.*, 2013). 92.70% of the SNVs, 65.91% of the indels and 54.84% of the TEs were correctly called (examples of IGV outputs are in Appendix Figure C. 1, Appendix Figure C. 2 and Appendix Figure C. 3). Mutations that did not pass the visual validation were removed from further analysis.

4.3.8 Mutation rate calculations

To estimate the mutation rate in each MA line, we need to consider when the mutation occurred. Somatic tissues from G0 and G4 are compared, therefore we capture mutations that occurred in the germ line once the germ cell lineage had split from the somatic lineage that led to most of the somatic tissue. There are three scenarios to consider:

(i) If the mutation occurs before this separation, then the mutation may be shared as a heterozygote between the somatic cell and the germ line. These mutations are not captured, as heterozygous sites at G0 are filtered out.

(ii) If the mutation occurred after separation in early germ-cell development, then these mutations may be shared between germ cells and hence their progeny.

(iii) If the mutation occurred after separation, late in germ-cell lineage then these mutations will be unique to their progeny.

If the rate of mutation is constant then very few mutations that occur in a germ line are shared between germ cells and hence we follow the scenario where (iii) mutations occur late in germ-cell lineage (discussed in Lanfear (2018)).

To detect new mutations, we compare G0 and G4 plants and infer a new mutation where the G4 plant is homozygous for a different allele to G0. We therefore need to calculate the probability that a new mutation in generation *x* is homozygous by G4. For mutations occurring in G0, the calculation is as follows; assuming that the mutation is present in a minority of germ cells the G1 plants, formed by selfing, are heterozygous for the mutation. In G2, the probability that a mutation is homozygous is 0.25, in G3 it is 0.375, and in G4 it is 0.4375. If we repeat this analysis for mutations occurring in G1 and G2 – note that the mutations in G3 will be heterozygous in G4 and hence not assayable - we find that the average probability of detecting a mutation occurring in G0 to G2 in G4 is 0.35. Hence, 65% of mutations are lost from our analysis. To correct for this effect, we divide the number of new mutations by 0.35 and then divide this estimate by 3 to reflect the number of generations assayed. Finally, this is the diploid mutation rate, so to get the haploid rate we divide by 2.

In other words, the number of SNVs, indels and TE mutations between G0 and G4 were counted (Figure 4.2) and the following were calculated:

$$\text{mutation frequency (per genome per generation): } m = \frac{n}{0.35 \text{ x } 3 \text{ x } 2 \text{ x } g}$$

$$\text{mutation rate (per site per generation): } \mu = \frac{m}{b}$$

With x as a multiplication sign, *n* as the total number of mutations, *g* as the number of generations and *b* as the number of bases analysed. Note that the effective number of generations in the experiment is only three because no mutations that occur in G3 can be homozygous in G4.

Shared mutations are expected from MA lines that share G1 parents. To estimate the number of homozygous mutations shared at G4 by these MA lines we calculate the probability that a new mutation heterozygous in G1 is homozygous in G4; this probability is 0.4374; hence the probability that two lines are homozygous for the same mutation is $0.4375^2 = 0.1914$; hence we expect 0.1914μ shared mutations (μ is the mutation rate per genome). The total number of mutations in G4 is 1.0625μ, therefore we expect 18.01% shared mutations (0.1914/1.0625 = 0.1801).

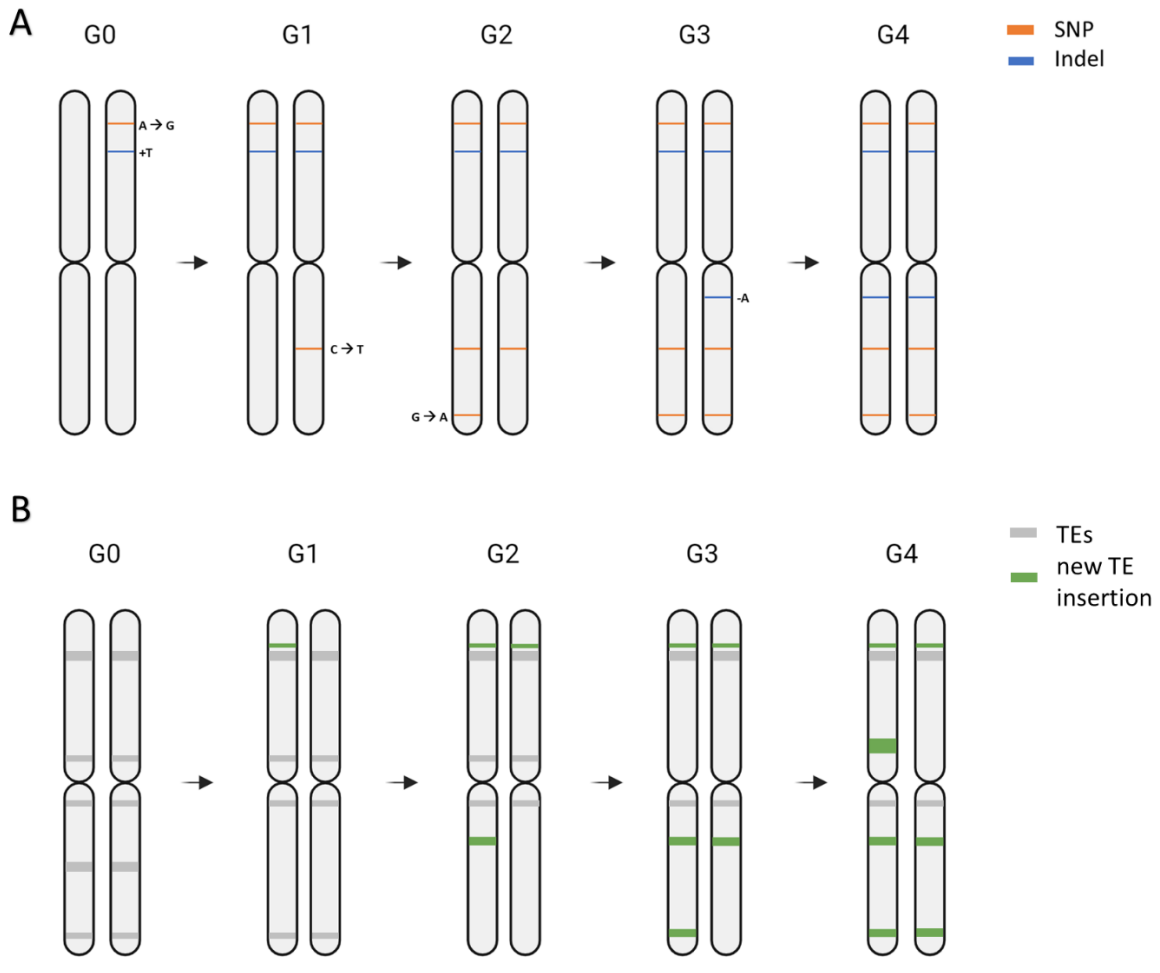


Figure 4.2: Hypothetical illustration of SNVs, indels and TE mutations.
(A) Single nucleotide variant (SNV) mutation with nucleotide changes, and indel mutations with insertion or deletions of nucleotides. (B) Transposable element (TE) mutations with a new TE insertion (TE absent at G0 and present at G4).

4.3.9 Genome-wide distribution of mutations

Positions of mutations were obtained by mapping onto SL4.0 reference genome to obtain the distribution (i) across chromosomes, (ii) between chromosome arm and pericentric regions, (iii) genic features (intergenic, 1kb upstream, genes, 1kb downstream). Regions of chromosome arms and pericentromeric were explored as these correspond to gene-rich and gene-poor regions, respectively (Wang *et al.*, 2006). Estimates of centromere and pericentric regions obtained from (The Tomato Genome Consortium, 2012). Gene positions were identified and 1kb upstream and 1kb downstream were extracted. These genomic positions were intersected SNV, indel and TE positions to annotate location within the reference genome. Mutation fitness rate was also calculated by summing the number of genic transversion SNVs, genic indels, and TEs in up/downstream and genic regions per genome per generation. Figures and statistical tests were performed using R v4.1.2 (R Core Team, 2021).

4.3.10 Statistical analysis

To explore if the total number of mutations was influenced by (i) clean reads, (ii) depth, (iii) mapping rate, and (iv) inner distance, Pearson's correlation test was also performed. We used a nested ANOVA to test whether there were significant differences between species and accessions within species. To test if the number of mutations was more than expected by chance, χ^2 test was performed for (i) A/T and C/G SNVs and (ii) insertions and deletions (separately). To test if shared mutation between genomes from the same accession were more than expected by chance, a χ^2 test was performed. Association between the number of mutations and their genomic distribution were tested using χ^2 tests comparing expected and observed counts for (i) different chromosomes, (ii) chromosome arm/ pericentromeric region, (iii) intergenic, up/downstream, or genic regions.

Nested ANOVA tests (accessions nested within species) were performed above if assumptions for homogeneity and normality were satisfied using diagnostic plots. Furthermore, homogeneity was tested with Levene's test for homogeneity of variance, and normality was tested with the Shapiro-Wilk normality test. ANOVA was followed by Tukey test for multiple comparisons of means. For comparisons that did not satisfy the assumptions of ANOVA, the data was rank transformed and diagnostic plots checked.

To understand the biological processes associated with genes with or near mutations, TopGO (Alexa and Rahnenfuhrer, 2022) was used to test the over-representation of gene ontology (GO) terms with genes of interest. SlimGO annotation of ITAG4.0 *Solanum lycopersicum* was used in a gene-to-GO format (http://systemsbiology.cau.edu.cn/agriGOv2/download.php). Fisher's exact tests compared genes bearing mutations and the background list of all genes associated with their GO annotation, focusing on biological processes as the ontology of interest with a minimum of five genes per GO term. P-values were adjusted using the Benjamini & Hochberg (1995) method to control the false discovery rate with an adjusted p-value < 0.05 considered significant.

4.4 Results

4.4.1 Mutation rates

A mutation accumulation (MA) pot experiment was set up with domesticated tomato *S. lycopersicum* (number of accessions; n=2), wild progenitor (hereafter "progenitor") *S. pimpinellifolium* (n=2) and never-domesticated wild (hereafter "wilds") *S. cheesmaniae* (n=1). Raw sequencing data resulted in an average of 87.77Mb (+/-5.02; SE) reads per sample, with an average read depth of 16x per sample. Data cleaning and filtering retained on average 87.75% of the reads per sample. No significant difference in the number of clean reads, sequencing depth, percentage of mapped reads and read inner distance was observed between progenitors and wilds (ANOVA: $p > 0.05$). Full mapping statistics are reported on Table C1.

We identified 207,243 short variants (184,966 SNVs and 22,277 indels) and 21,977 TEs; accessions clustered within their species, with tighter clustering observed in short variants compared to TEs (Appendix Figure C. 4). Single nucleotide variant (SNVs), indels (insertion-deletions ≤50bp) and transposable elements (TEs) were identified in the G0 and G4 samples of domesticates (number of genomes; n=6), progenitor (n=5) and wilds (n=3). A total of 202 homozygous *de novo* mutations were identified (Table 4.1; Full breakdown in Table C2 and species summary on Table C3). We used a nested ANOVA to test whether there were significant differences between species and accessions within species. We find both effects were significant (nested ANOVA: $p < 0.05$), with progenitor species showing significantly more mutation than the wild species (Tukey: $p = 0.005$; Figure 4.3A). The total mutation count was not significantly correlated to the number of reads ($t = -0.046$, $df = 7$, $p = 0.965$), depth ($t = -0.172$, df

= 7, p = 0.868), mapping rate (t = -0.724, df = 7, p = 0.493) or inner distance (t = 0.145, df = 7, p = 0.889; Appendix Figure C. 5). The greatest number of mutations were SNVs with 154, followed by indels with 29 and TE insertions with 19 (Table 4.1).

Table 4.1: Summary of mutation frequency and rate.
SNVs, indels and TEs mutations were identified in domesticates (*S. lycopersicum*), progenitor (*S. pimpinellifolium*) and never-domesticated wild (*S. cheesmaniae*) genomes.

species	sample ID	SNVs	SNV rate (per site per gen)	Indels	indel rate (per site per gen)	TEs	TE rate (per site per gen)	total mutation
Domesticated	SLL_A1	23	14.17×10^{-9}	4	2.47×10^{-9}	2	1.23×10^{-9}	29
Domesticated	SLL_A2	27	16.64×10^{-9}	6	3.70×10^{-9}	1	0.62×10^{-9}	34
Domesticated	SLL_B1	10	6.16×10^{-9}	3	1.85×10^{-9}	1	0.62×10^{-9}	14
Domesticated	SLL_B2	8	4.29×10^{-9}	2	1.23×10^{-9}	2	1.23×10^{-9}	12
Progenitor	SP_A1	12	7.39×10^{-9}	2	1.23×10^{-9}	3	1.85×10^{-9}	17
Progenitor	SP_B1	34	20.95×10^{-9}	6	3.70×10^{-9}	3	1.85×10^{-9}	43
Progenitor	SP_B2	27	16.64×10^{-9}	4	2.47×10^{-9}	4	2.47×10^{-9}	35
Wild	SC_B1	7	4.31×10^{-9}	1	0.62×10^{-9}	2	1.23×10^{-9}	10
Wild	SC_B2	6	3.70×10^{-9}	1	0.62×10^{-9}	1	0.62×10^{-9}	8

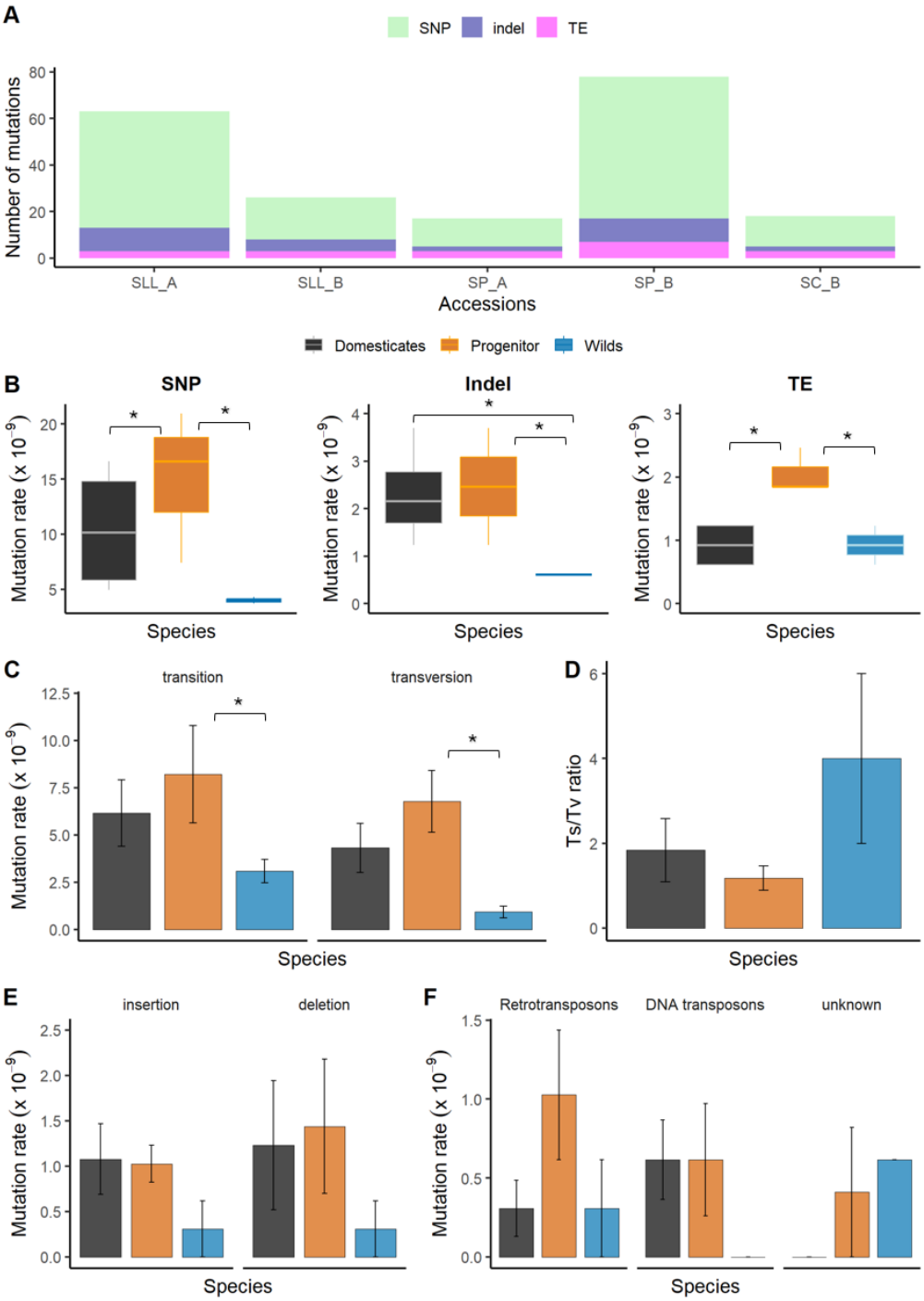


Figure 4.3: Mutation count and rate.

(A) Total number of TE, indel and SNV mutations found in each accession of domesticated (SLL: *S. lycopersicum*), progenitor (SP: *S. pimpinellifolium*) and never-domesticated wilds (SC: *S. cheesmaniae*). (B) Mutation rate (per site per generation) for each species across different mutation type including (C) SNVs mutation type of transition and transversion; (D) transition/transversion (Ts/Tv) ratio; (E) insertions and deletions; and (F) TE class (retrotransposons, DNA transposons and unknown).

SNV mutations were detected in all genomes (Table 4.1; Table C4). Using a nested ANOVA we find significant differences in the mutation rate (per site per generation) between species (nested ANOVA: $F(2) = 21.81$, $p = 0.007$) and between accessions within species (nested ANOVA: $F(2) = 27.67$, $p = 0.005$). The progenitor (14.99×10^{-9}) had a significantly higher SNV mutation rate than the wild (4.01×10^{-9}; Tukey: $p = 0.006$), but not the domesticated species (10.47×10^{-9}; Figure 4.3B). SNVs were annotated as transversions or transition as mutations can be biased towards one more than the other (Figure 4.3C). There were significantly more transitions (58.4%) than transversion (41.6%; binomial: $p = 0.018$). Overall, the largest proportion of mutations (42.2%) were C:G→T:A transitions, consisting of 54.4% in the domesticates, 27.4% in the progenitor and 61.5% in the wilds. C→T mutation in the domesticates (29.4%) and wilds (46.2%) were the most frequent, however, in the progenitor it was T→C (16.4%). There were more A/T mutations than expected by chance ($\chi^2 = 66.744$, $df = 5$, $p < 0.001$; Appendix Figure C. 6A). The transition/transversion (Ts/Tv) ratio was the greatest in the wild (mean±SE; 4.0±2.00) and the least in the progenitor (1.2±0.29), with the domesticates (1.8±0.74) in between (Figure 4.3D), but there was no significant difference in Ts/Tv ratio between species (ANOVA: $F(2) = 1.438$, $p=0.338$).

Indel mutations were detected in all genomes, with all insertions/deletions less than 30bp and the majority (69.0%) just 1bp (Table 4.1; Table C5). We found a significant difference in indel mutation rate between species (nested ANOVA: $F(2) = 20.16$, $p = 0.008$; Figure 4.3E) and between accessions within species (nested ANOVA: $F(2) = 8.28$, $p = 0.038$) effects were significant, with the progenitor (2.47×10^{-9}) having significantly higher rates of indel mutations than the wild (0.62×10^{-9};Tukey: $p = 0.025$) but not the domesticated species (2.31×10^{-9}; Figure 4.3B). There was no significant difference in the number of deletions and insertions (binomial: $p = 0.356$) and there was no significant association between species and indel type ($\chi^2 = 0.4$, $df = 5$, p-value = 0.995).

TE insertions were detected in all genomes (Table 4.1; Table C6). There was no significant difference in TE insertion rate between species (nested ANOVA: $F(2) = 3.77$, $p = 0.120$; Figure 4.3B) and between accessions within species (nested ANOVA: $F(2) = 0.04$, $p = 0.965$). Mean TE insertion rates for each species were: domesticated (0.92×10^{-9}), progenitor (2.05×10^{-9}), and wilds (0.92×10^{-9}). TE mutations were also annotated based on their TE classification (Table C6). There was no significant TE classification effect (two-way ANOVA: $F(2) = 0.96$, $p = 0.409$),

species effect (two-way ANOVA: $F(2) = 2.24$, $p = 0.149$) and interaction effect (two-way ANOVA: $F(2) = 1.80$, $p = 0.194$). This suggests no significant difference in TE insertion rate between progenitor and wild species.

Shared mutations were observed between replicates from the same accessions. There were three (SNV = 2; indel = 1) shared mutations in SLL_A genomes, 18 (SNV = 12; indel = 2; TE = 1) shared mutations in SP_B genomes and one (TE = 1) shared mutation in SC_B genomes. This results in 140 unique SNVs, 26 unique indels, and 17 unique TE mutations. This accounts for 10.00% (14/140) of SNVs, 11.54% (3/26) of indels and 11.76% (2/17) of TEs shared. This is not significantly different from the expected shared mutation of 18.01% (binomial: $p > 0.05$; see Methods).

4.4.2 Location of mutations across the genome

The genomic location of each mutation on each chromosome, taking into account chromosome size (Figure 4.4A), suggests that the number of mutations was greater in some chromosomes than others, more than expected by chance ($\chi^2 = 56.305$, $df = 11$, $p < 0.001$; Appendix Figure C. 6A). Mutations were enriched in chromosomes 4 and 6 (Figure 4.4A; Appendix Figure C. 6B). The association of mutations in chromosome arms or pericentromeric regions was also explored as these correspond to gene-rich and gene-poor regions, respectively (Wang *et al.*, 2006). The number of mutations in chromosome arms was significantly greater than those in pericentromeric regions ($\chi^2 = 62.859$, $df = 5$, $p < 0.001$), with the greatest contribution to the χ^2 score attributed to the positive association of mutations in the progenitor to the chromosome arm regions (Appendix Figure C. 6C). This highlights that mutations are not randomly distributed in the tomato genome.

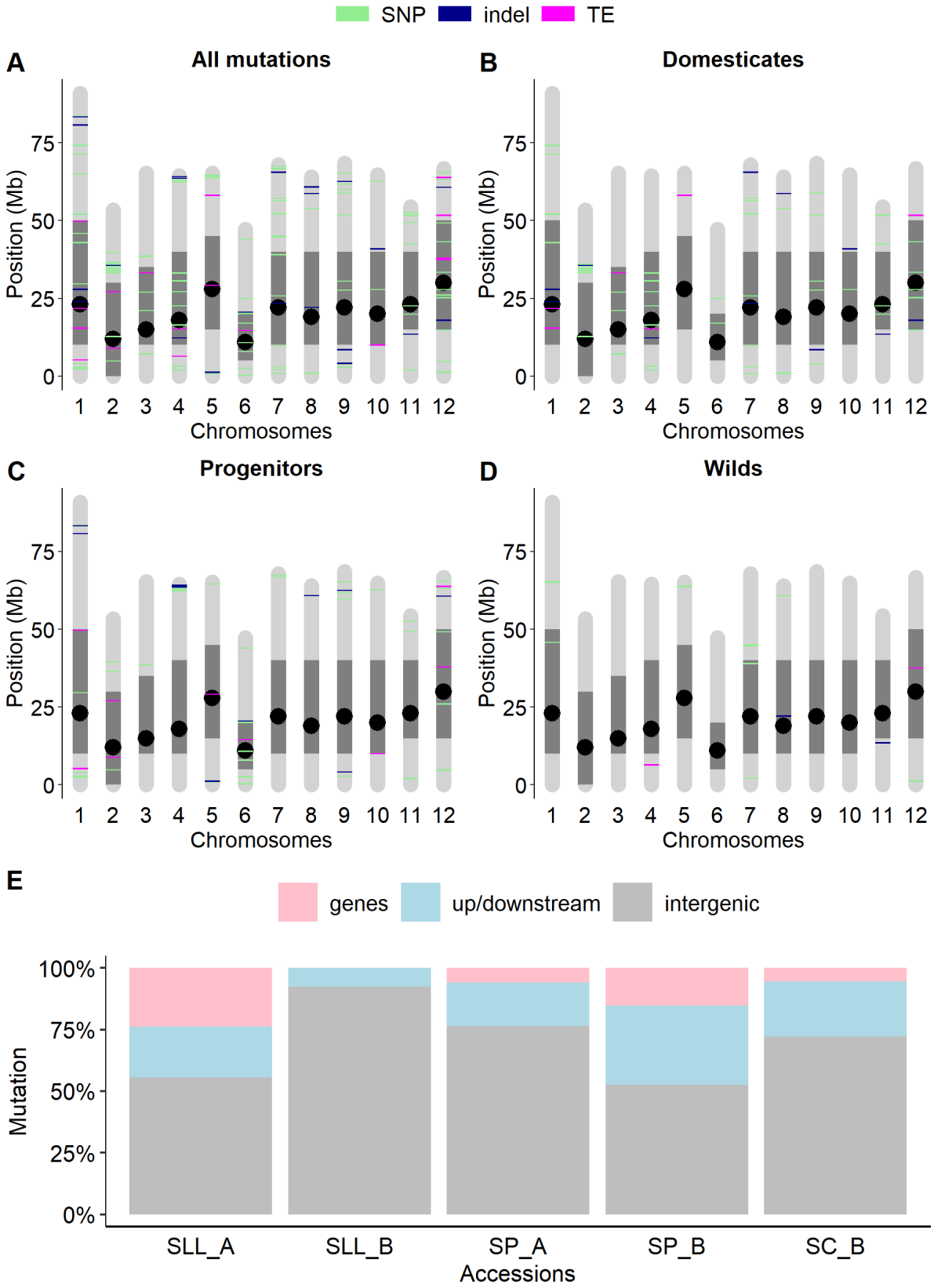


Figure 4.4: Genomic distribution of mutations across chromosomes.

(A) Positions of all mutations identified in all genomes in the analysis. Genomic positions of mutations in (B) domesticated (SLL: *S. lycopersicum*), (C) progenitor (SP: *S. pimpinellifolium*) and (D) never-domesticated wilds (SC: *S. cheesmaniae*). (E) The proportion of mutations in each genomic region. Black dots represent the estimated centromere position and dark grey regions the estimated pericentromeric region.

The majority of mutations were located in the intergenic region in all genomes (Figure 4.4E). Genomic region had a significant effect on the number of mutations (two-way ANOVA: $F(2) = 19.365$, $p < 0.001$), with the intergenic region having more mutations than up/downstream (Tukey: $p < 0.001$) and genic regions (Tukey: $p < 0.001$). There was also a significant species effect on the number of mutations (two-way ANOVA: $F(2) = 4.975$, $p < 0.010$), with greater number of mutations found in the progenitor than the wilds (Tukey: $p = 0.007$). Accounting for the relative size of each region, the number of mutations across different genomic regions was greater in up/downstream regions of genes, more than expected by chance ($\chi^2 = 75.118$, $df = 8$, $p < 0.001$), with mutations in each tomato group positively associated with up/downstream regions of genes (Appendix Figure C. 6D); the greatest contribution to the χ^2 score attributed to the progenitor. Gene Ontology (GO) analysis of genes with or near a mutation was performed, however, no enriched GO terms were detected in the domesticates, progenitors and wilds separately or together.

Assuming genic SNVs, indels and TEs mutations have the greatest effect on fitness, there were six in SLL_A (SNV = 13, indel = 2), zero in SLL_B, two in SP_A (indel = 1), eight in SP_B (SNV = 11, TE = 1), and one in SC_B (indel =1). These mutations and their associated genes show a variety of functions (Table 4.2). Noteworthy are the mutations in the progenitor heat shock protein 90 gene and UDP-N-acetylglucosamine diphosphorylase 1 gene with roles in heat acclimation and defence responses, respectively.

Table 4.2: Putative mutations affecting genome fitness.

Tomato group	Sample	Type	Mutation	Changes	Gene ID	Description	Role	References
Domesticates	SLL_A1	SNV	transversion / nonsynonymous	C→A	Solyc09g015280	Photosystem I assembly protein Ycf3	assembly of the photosystem I (PSI) complex	Naver *et al.* (2001)
Domesticates	SLL_A2	SNV	transversion / nonsynonymous	A→T; T→A; A→T	Solyc01g017120	Protein Ycf2	chloroplast protein import	Kikuchi *et al.* (2018)
Domesticates	SLL_A2	Indel	deletion	CT→C	Solyc04g039850	ATP synthase subunit beta	plant cell death regulator	Chivasa *et al.* (2011)
Domesticates	SLL_A2	Indel	deletion	$G(A)_7 \rightarrow G(A)_6$	Solyc11g021310	Protein TIC 214	Protein translocation	Kikuchi *et al.* (2013)
Progenitor	SP_A1	Indel	insertion	T→TA	Solyc04g082530	Bromodomain-containing protein	Role in epigenetic regulation	Abiraami *et al.* (2023)
Progenitor	SP_B1	SNV	transversion / synonymous	T→G	Solyc04g081630	Heat shock protein 90	role in heat acclimation	Yamada *et al.* (2007)
Progenitor	SP_B2	SNV	transversion / synonymous	G→C	Solyc02g068530	UDP-N-acetylglucosamine diphosphorylase 1	regulating leaf senescence and defence responses	Wang *et al.* (2015b)
Progenitor	SP_B2	SNV	transversion / nonsynonymous	A→T	Solyc09g065540	biotin-binding protein	bind biotin molecules	Murray *et al.* (2002)
Progenitor	SP_B1 and SP_B2	SNV	transversion / nonsynonymous	A→C; T→A; T→A	Solyc04g081690	Unknown protein	-	-
Progenitor	SP_B2	TE	MITE	TE insertion	Solyc02g032410	Unknown protein	-	-
Wilds	SC_B2	Indel	deletion	TA→T	Solyc11g021260	Protein TIC 214	Protein translocation	Kikuchi *et al.* (2013)

4.5 Discussion

Investigations of mutation rates with the use of mutation accumulation (MA) experiments have been conducted in various organisms (Lynch *et al.*, 2008; Keightley *et al.*, 2009; Denver *et al.*, 2012; Lee *et al.*, 2012; Keith *et al.*, 2016). There are limited MA experiments in plants, with the majority conducted on *Arabidopsis thaliana* (Ossowski *et al.*, 2010; Weng *et al.*, 2019; Jiang *et al.*, 2014), but none, to our knowledge, have investigated whole-genome TE insertion rates of transposable elements (TE). In our study, we employed an MA experiment in combination with whole-genome sequencing to explore if mutation rates played a role in tomato domestication. We revealed that tomato wild progenitor *S. pimpinellifolium* has a higher mutation rate than never-domesticated wild *S. cheesmaniae*, both overall and individually for SNVs, and indels, but not TE. These higher mutation rates may have facilitated adaptation through faster exploration of mutation space within the genome, generating genetic variation important for evolution in early domestication.

We found significantly more SNV mutation in the progenitor (mean per site per generation; 14.99 x 10^{-9}) than in the wilds (4.01 x 10^{-9}), but not the domesticated species (10.47 x 10^{-9}). Previous estimates of substitution rates in wild tomato species of 1.6 to 5.2 x 10^{-9} silent substitution per year is comparable with our estimate of SNV mutation rates (Roselius et al., 2005). Wang and Obbard (2023) reported estimates of SNVs in plants for duckweed species (0.23 x 10^{-9} and 0.22 x 10^{-9}), *Silene latifolia* (7.51 x 10^{-9}), peach (8.01 x 10^{-9}), rice (8.09 x 10^{-9}), *Brassica rapa* (9.30 x 10^{-9}), *Arabidopsis thaliana* (9.34 x 10^{-9}) and maize (21.26 x 10^{-9}). Our estimates of SNV mutation rates are within the range of these previously reported rates in plants, apart from maize. The high rates in maize are most likely due to greater genome size, as mutation rates are positively correlated with genome size (Wang and Obbard, 2023). On the other hand, duckweed species have relatively low mutation rates compared to most plant species, this has been attributed to the smaller number of cell divisions compared to the other larger, long-lived plant species (Sandler *et al.*, 2020). Our annotation of SNVs found more transition than transversion; a majority of SNV mutations were C:G→T:A transitions resulting in a significant A/T mutation bias. This is in accordance with patterns found in other plant studies (Ossowski *et al.*, 2010; Weng *et al.*, 2019; Xie *et al.*, 2016; Sandler *et al.*, 2020) and in eukaryotic genomes as a whole (Katju and Bergthorsson, 2019).

Our estimates of indel mutation rates for domesticates (2.31 x 10^{-9}), progenitors (2.47 x 10^{-9}) and wilds (0.62 x 10^{-9}) revealed that the progenitor had significantly more indel mutations than wilds species. The estimates for the progenitor and domesticated species are comparable to those

reported by Wang and Obbard (2023) in peach (2.49×10^{-9}), rice (2.75×10^{-9}) and *Arabidopsis* (3.05×10^{-9}), but the wild species indel mutation rates estimate is still similar to some previous estimates in *Arabidopsis* of 4.00×10^{-10} (Ossowski *et al.*, 2010). We also found no deletion bias in any tomato group, as reported in many taxa (Katju and Bergthorsson, 2019; Wang and Obbard, 2023). This might be due to fewer number of generations compared to other studies. We found SNV rates are higher than indel rates, this is a common trend that indels are reported less frequently than SNVs across eukaryotic species (Katju and Bergthorsson, 2019), as well as a positive correlation between the two (Wang and Obbard, 2023).

Even though no significant difference in TE insertion rates was detected between the domesticated (0.92×10^{-9}), progenitor (2.05×10^{-9}) and wild species (0.92×10^{-9}), to the best of our knowledge, no other studies have compared TE insertion rates between wild progenitors and other wild species in plants. However, whole genome TE insertion rates have been reported in *Drosophila* with estimates of 2.11×10^{-9} to as high as 4.93×10^{-9} (Wang *et al.*, 2023b; Adrion *et al.*, 2017). In *Arabidopsis*, previous MA experiments have not detected any novel TE insertions in up to 25 generations (Weng *et al.*, 2019; Lu *et al.*, 2021). Both studies utilised the TE detection tool Jitterbug (Hénaff *et al.*, 2015), but in a benchmarking study utilising rice, Drosophila and human genomes, PopoolationTE2 (used in this study) was picked as the overall top broad-spectrum tool compared to multiple TE detection tools that included Jitterbug (Vendrell-Mir *et al.*, 2019). It has been proposed that a combination of TE tools could improve the accuracy of TE detection (Nelson *et al.*, 2017; Vendrell-Mir *et al.*, 2019; Bajus *et al.*, 2022). We show that TE transposition can occur in tomatoes with genomes roughly seven times bigger than those of *A. thaliana* and in fewer generations. This is consistent with the TIP analysis in tomatoes by Dominguez *et al.* (2020) identifying a set of TE families with recent mobilisation activity. High TE insertion rates in some lines but not in others, have been reported previously, for example in maize, this was linked to differences in active TEs through autonomous elements with transposition functions that allow the movement of TEs in the same family (Dooner *et al.*, 2019). Therefore, to allow for better mutation rate estimations, within-species variation of mutation rates will need to be further investigated.

Various genomic features can influence the type of mutation and its frequency. We found more mutations in chromosome arms than in pericentric regions than expected by chance. In contrast to this, in previous studies, mutations were more frequent in pericentric regions but limited to AT sites and non-TE regions in *Arabidopsis* (Weng *et al.*, 2019). The difference between these two studies might be due to differences in the number of generations leading to over-estimation of mutation rates in chromosome arms than pericentric region. Even though

mutations were more frequently found in intergenic regions, we found more mutations in regions up/downstream of genes than expected by chance, most prominently in the progenitor. High mutation rates in upstream and downstream regions relative to the genic region have been reported in Arabidopsis as well (Monroe *et al.*, 2022). Mutations near genes are more likely to affect genes, generating greater genetic and phenotypic variation in the progenitor that may have aided their domestication. Another factor that influences the variability of mutation rates across the genome is recombination rates. Higher mutation rates in regions of high recombination rates have been reported in various organisms (Yang *et al.*, 2015; Roselius *et al.*, 2005). Differences in recombination rates between the progenitor and domesticated tomato indicates different mutation hotspots between these species (Fuentes *et al.*, 2022). Other factors are GC content and timing of replication on different chromosomes (Arndt *et al.*, 2005; Bracci *et al.*, 2023). All these suggest that the genome-wide distribution of mutations is non-random.

Most mutations are neutral or deleterious (Eyre-Walker and Keightley, 2007); the adaptive value of these mutations in natural environments will depend on fewer deleterious mutations and an increase in beneficial mutations. However, in early domestication, the movement of wild plants from their natural environment to human-modified fields in early domestication would have reduced effective population size. In these small populations, genetic drift can eliminate selection against deleterious mutations and can lead to the fixation of mildly harmful mutations that can reduce fitness (Whitlock, 2000). Reduction in competition and increased resources such as additional nutrients in human-modified environments, would further relax purifying selection. This permits the population to persist despite a high mutation load (Renaut and Rieseberg, 2015; Smith *et al.*, 2019; Wang *et al.*, 2021a), allowing beneficial mutations to be selected by early farmers. The influence of effective population size on TE load variance has been reported in the range expansion of *Arabidopsis*, with TE accumulation in expanded populations attributed to high transposition rate and selective sweep (Jiang *et al.*, 2024). High mutation load that may be harmful in natural environments may be beneficial in agricultural settings (Dwivedi *et al.*, 2023); for example, *sh1* loss of function mutation was vital in the domestication of maize, rice and sorghum (Lin *et al.*, 2012). Therefore, environmental changes in early domestication can create an opportunity for new beneficial mutations and adaptive evolution.

Our validation of mutations with IGV (Thorvaldsdóttir *et al.*, 2013) resulted in many false positive calls highlighting the necessity of validation tools for visual checks. This could have been influenced by the reliance on the same site in two different genomes to have accurate mapping

and enough coverage. Long-read sequencing has been proposed to improve the identifications of TEs by reducing mapping ambiguity and resolving repetitive complex regions (Amarasinghe *et al.*, 2020). Our mutation estimates could be improved by the addition of the mutations that occur between G3 and G4. These were not captured by the analysis as these were in a heterozygous state and are more difficult to identify as these would be at low frequency and are more prone to inaccurate calls. Robust statistical testing on mutation estimates is hindered by small sample sizes and biases in species representation (Katju and Bergthorsson, 2019). The number of generations over which mutations accumulate might influence estimations of mutation rates. A model predicting the rate of divergence of the population proposed underestimation (overestimation) of mutation rates with fewer generations in a homozygous (heterozygous) base population (Lynch and Hill, 1986). The mating system (i.e. selfing vs outcrossing) could also influence mutation rates, in this case, our tomato plants were selfed for four generations. Selfing increases homozygosity, which in turn decreases the recombination rate and increases the mutation rate (Wright *et al.*, 2008; Burgarella and Glémin, 2017). There is a relaxation of natural selection against mutations due to the reduction in recombination rates, resulting in the accumulation of mutations (Wright *et al.*, 2008). This is supported by higher population frequencies of TEs in selfing *A. thaliana* compared to outcrossing *A. lyrate*, attributed to reduced selection against TE insertions in selfing species (Lockton and Gaut, 2010; Bonchev and Willi, 2018).

We acknowledge that the populations of tomato crop relatives used, *S. pimpinellifolium*, and *S. cheesmaniae*, are not identical to those that existed around the time of early domestication due to subsequent selection and gene flow in the wild (Flint-Garcia *et al.*, 2023). However, assessment of mutation rate through mutation accumulation experiment can only be performed with extant populations. Higher mutation rates may have aided the domestication of *S. pimpinellifolium*, but due to *S. cheesmaniae*'s geographical location in the Galapagos Islands, these species would not have competed in early domestication. Therefore, exploration of other never-domesticated species, as well as additional accessions, would broaden our understanding of mutation rate differences between these groups. On the other hand, these were not comparable in our pipeline due to the reduced mapping efficiency of the sequencing data (Chapter 3).

4.6 Conclusion

Mutation rate is an important factor in understanding evolution that can influence the adaptation to changing environments over time. Here, we provide the first estimates of mutation rates in tomatoes. Mutation rates of SNVs and indels were significantly higher in the progenitor than in the never-domesticated wild tomatoes. High mutation rates in the progenitor, associated with the upstream/downstream region of genes may have aided its domestication, contributing to genetic and phenotypic variation. These may have been selected upon during domestication. Although no significant difference in TE insertion rates was detected between species, we provide the first estimate of TE insertion rates in plants. Our study contributes to the growing number of MA studies and improves our understanding of the role of mutation rates in early domestication.

Chapter 5 Discussion

Out of hundreds of thousands of edible plant species (Willis, 2017), only a few hundred have been domesticated (Zeven and De Wet, 1982). This begs the question, why are some species domesticated and others are not? We examined the genomic constraint on domestication that allows the domestication of certain species over others. Various environmental changes marked the period of early domestication (Wood and Lenné, 2018; Piperno *et al.*, 2019), therefore the ability of a species to adapt to changes in growing conditions, as well as produce traits favourable to early farmers could facilitate and fast-track its chance of domestication. Here we use the domestication of the tomato crop (*Solanum lycopersicum*) to explore the role of plasticity and transposable elements (TEs) in early domestication. This thesis had three main aims:

(i) **To explore the role of plasticity in early domestication**. To assess phenotypic and gene expression plasticity between the tomato progenitor (*S. pimpinellifolium*) and never-domesticated wild species (*S. cheesmaniae* and *S. chmielewskii*). To find evidence for plasticity changes between the progenitor and the domesticated species during tomato domestication.

(ii) **To explore the role of genetic diversity in early domestication as an indicator of mutation rates**. To assess single nucleotide polymorphism (SNP) and transposon insertion polymorphism (TIP) diversity within the progenitor (*S. pimpinellifolium*) and never-domesticated species (*S. cheesmaniae* and *S. galapagense*) and compare these between species. To identify genes that may have been important during domestication based on changes in TE frequency.

(iii) **To explore the role of mutation rates in early domestication**. To assess single nucleotide variant (SNV), indel and TE mutation rates between the progenitor (*S. pimpinellifolium*) and never-domesticated species (*S. cheesmaniae*). To identify putative mutations affecting genome fitness.

Overall, this thesis found evidence for the role of plasticity, genetic diversity and mutation rates in the selective advantage of the progenitor. These have not been explored in much depth individually and certainly not together in wild crop relatives. One or a combination of these could have facilitated the evolution of novel phenotypes for selection to act on in early domestication, aiding the domestication of the progenitor species.

5.1 Plasticity played a role in the domestication of tomato progenitor

In chapter one, the role of plasticity in domestication was evident by the differentially expressed genes between domesticated and progenitor species more likely to be plastic than expected by chance. Furthermore, there was a reduction in the number of plastic genes during domestication. This is suggestive of genetic assimilation during domestication that has also been reported in other crops (Lorant *et al.*, 2017; Belcher *et al.*, 2023; Mueller *et al.*, 2023). These are indicative of the importance of plasticity in early domestication. More traits and genes were plastic in the progenitor species than in the never-domesticated wild species. Noteworthy was plasticity in tomato fruit weight and size as these are important traits in tomato domestication, with tomatoes primarily grown for their fruit yield. High fruit yield in the progenitor species under cultivation could have made them more attractive to farmers. Genes that were plastic in the progenitor were enriched in a pathway involved in plant hormone signal transduction, related to several important plant processes such as germination (Harberd *et al.*, 2009), fruit ripening (Chen *et al.*, 2005) and stress response (Katsir *et al.*, 2008). Greater plasticity may have facilitated adaptation and rapid evolution of novel phenotypes for selection to act on. However, including more accession and species would strengthen these findings further. This is the first comparative assessment of both phenotypic and genetic plasticity in multiple tomato species. The field should do more to assess plasticity at different biological levels, this is necessary to give insight into underlying mechanisms of plasticity and their influence on evolvability. Genes with reduced plasticity as a result of domestication can be investigated in other crops to identify adaptive alleles that could be important for the breeding of more resilient crops (Brooker *et al.*, 2022).

5.2 High TIP diversity in tomato progenitor

In the second chapter, whole-genome resequencing datasets revealed greater nucleotide and TIP diversity in tomato progenitor species than never-domesticated wild species. Although our sample size was limited by the variability in sequencing data due to differences in library preparation that influenced the detection of TEs (which is important for other authors to consider), our findings were in accordance with previous genetic diversity reports for high nucleotide diversity in wild tomatoes (Lin *et al.*, 2014; Sauvage *et al.*, 2017; Razifard *et al.*, 2020) and the progenitor species having a diverse set of TE families (Dominguez *et al.*, 2020). TIPs

associated with genes that were putatively selected based on drastic changes in frequency between the progenitor and domesticated accessions were identified, highlighting the contribution of TEs in tomato domestication. Many of these genes had functions related to biotic stress resistance and abiotic stress tolerance, consistent with reports in tomato (Dominguez *et al.*, 2020), rice and *Arabidopsis* (Quadrana *et al.*, 2019). Greater diversity correlates with higher mutation rates (Wang and Obbard, 2023), thus these might indicate higher mutation rates in the progenitor that allowed faster exploration of mutation space within the genome, generating genetic variation important for adaptation in early domestication. This work builds on Dominguez *et al.* (2020)'s work with the addition of annotation including both reference and non-reference TIPs, that included MITEs. TIPs can be valuable molecular markers, associated with important agronomic traits that can aid future crop breeding.

5.3 First estimates of mutation rates in tomato species

In chapter three, we build on the indication in chapter two that the progenitor has high genetic diversity which could be due to high mutation rates. Estimates of SNV and indel mutation rates (per site per generation) were higher in the progenitor than in the never-domesticated wild, however, no significant differences were detected with TE insertion rates. Both retrotransposons and DNA transposons were found to be active in tomato species, consistent with the TIP analysis in tomatoes by Dominguez *et al.* (2020) identifying a set of TE families with recent mobilisation activity. These estimates are comparable with previous reports on plants (Wang and Obbard, 2023). In the progenitor, more mutations were located in up/downstream regions of a gene in tomato groups than expected by chance; these mutations can generate genetic and phenotypic variation that could have aided their domestication. Mutation accumulation (MA) experiments allow for direct estimates of mutation rates and increasing the number of generations and capturing heterozygous mutations would improve our estimates of mutation rates in the tomato species. This is the first MA experiment in tomato species to estimate mutation rates. MA experiments are still rare in plants and the field needs to do it more so we can understand fundamental biological principles underlying not just the role in the selective advantage of the tomato progenitor in early domestication but evolvability more generally.

5.4 Limitations

Our use of the tomato clade to investigate the genomic constraints in domestications provided a model organism that was easy to grow with relatively short life cycle and with an abundance of genomic resources. However, wild tomato species had great diversity in mating system, growth rates and fruit production. This limited the number of species that could be included in our experiments to those species that were self-compatible, fast-growing and able to produce fruits quickly. Accessions were also required to have high mapping efficiency onto the reference genome for the transcriptomic and genomic sequencing. Together, this constrained us to only use tomato species *S. cheesmaniae* and *S. galapagense*. The life cycle of tomato compared to *Arabidopsis*, for example, is relatively longer, restricting the number of generations in an MA experiment in a given period. The wild species in particular have longer growing periods before setting fruits.

Greater plasticity, genetic diversity and/or mutation rates may have aided the domestication of *S. pimpinellifolium*, however, due to the geographical location of the studied never-domesticated wild species, *S. cheesmaniae* and *S. galapagense* in the Galapagos Islands, these species would not have competed in early domestication. There is also evidence that these species might have experience genetic bottleneck as a result of island colonisation and recent adaptation (Nuez et al., 2004; Pailles et al., 2017; Li et al., 2023). This could affect these species plasticity, genetic diversity and mutation rates, highlighting the need for further investigation with additional never-domesticated species. However, these were not comparable in our pipeline due to the reduced mapping efficiency (Chapter 3).

The inclusion of more accessions would capture a more representative measure of plasticity, genetic diversity and mutation rates. However, many wild tomato species refused to produce fruits; one accession of *S. cheesmaniae* stopped producing fruit after a couple of generations.

We also acknowledge that the populations of progenitor and never-domesticated species used are not identical to those that existed around the time of early domestication due to subsequent selection and gene flow in the wild (Flint-Garcia *et al.*, 2023), yet assessment of plasticity and mutation rates can only be performed with extant populations.

5.5 Future directions

5.5.1 Exploration of other taxa

The selective advantage of the tomato progenitor species over never-domesticated wild species may apply to other taxa. Investigating plasticity, TIP diversity and TE insertion mutations in other crops and their wild relative would broaden our understanding of the role of plasticity and TEs in the facilitation of domestication in progenitor species. However, there are still limited genetic resources for many wild relatives of crops, hindering their genomic characterisation. Sequencing of crop wild relative accessions to increase whole-genome sequencing data availability would encourage more genomic research including these species. The inclusion of wild genome references through the use of pan genomes, for example, would also better improve TE annotation of wild genomes to guarantee a sufficient comparison of TIPs (Li *et al.*, 2023). Utilisation of this may uncover greater genetic diversity both in SNPs and TIPs.

5.5.2 Exploration of plasticity at different biological levels

Gene expression is dynamic and can change through time influenced by developmental stage, environment, and epigenetics (Rivera *et al.*, 2021). Therefore, transcriptomes from multiple time points could broaden our understanding of how gene expression plasticity may vary. Comparison between progenitors and never-domesticated species of other tissues such as fruit and root would also give a comprehensive view of each species' ability to be plastic, especially as previous studies have indicated pre-adaptation to cultivation in root morphology of crop progenitors (Martín-Robles *et al.*, 2018). This will expand our understanding of plasticity, avoiding the possible bias that could be present with only a handful of phenotypic traits.

The genetic basis of plasticity, apart from gene expression plasticity, would also give insight into other important genomic processes. Epigenetic modifications allow plants to adjust through phenotypic plasticity under different environmental conditions (Dar *et al.*, 2022). Variation in DNA methylation between progenitor and never-domesticated species may provide the genetic basis for differences in plasticity. This is important in the biotic and abiotic stress response in tomatoes (Tian *et al.*, 2021; González *et al.*, 2013; Sahu *et al.*, 2014) and thus may have been important in adaptation in early domestication. Additionally, TEs have been linked to phenotypic plasticity (Pimpinelli *et al.*, 2019) due to their interaction with genes upon their insertion. Integration of TEs into a genome can change gene expression (Hirsch and Springer, 2017), and

influence methylation as DNA methylation suppresses TE expression which can result in the downregulation of nearby genes (Hollister and Gaut, 2009). TEs are also known to introduce new regulatory elements upon their insertion (Chuong *et al.*, 2017). These changes contribute to the genetic and phenotypic plasticity of a genome. It would be interesting to explore the relationship between plasticity and TEs.

5.5.3 Effect of domestication on mutation rates

We investigated the role of plasticity, genetic diversity and mutation rates on domestication but domestication itself (and/or domestication landscapes) may have induced changes in mutation rates in early domestication. Environmental changes that occurred in early domestication, such as changes in growing conditions (Wood and Lenné, 2018) and climate (Piperno *et al.*, 2019) could have triggered changes in mutation rates. There is evidence of stress conditions increasing TE activity for several TE families in various plants (Kimura *et al.*, 2001; Salazar *et al.*, 2007; Woodrow *et al.*, 2011; Ito *et al.*, 2013; Roquis *et al.*, 2021). Activation of TEs can trigger random genetic variation, vital for adaptation through natural selection (Schrader and Schmitz, 2019). Therefore, an increase in TE activity could mean more novel phenotypes arise in certain species, promoting their domestication through the accumulation of beneficial mutations at the expense of others (Romero et al., 2025).

5.6 Conclusion

Our understanding of crop domestication has focused on the selection that transformed the crop progenitor into the domesticated crop. Little is known about the selection between wild species in early domestication. In this thesis, we explore the role of standing genetic variation (plasticity and genetic diversity) and the generation of new mutations (mutation rates) in domestication. We found that the tomato progenitor had a greater number of plastic traits and genes, greater genetic diversity and high mutation rates compared to the never-domesticated wild species we analysed. This is the first characterisation of phenotypic and genetic plasticity, TIP diversity and estimates of mutation rates in multiple tomato wild species. Our work emphasises the contributions of plasticity, genetic diversity and mutation rates in the facilitation of domestication for the progenitor species. These highlight the valuable genomic

resources we can gather from wild crop relatives that can support the development of new crop varieties that can sustainably tolerate current and future environmental challenges.

Appendix A Chapter 2

Appendix Methods A. 1:

Plant growth

Tomato accessions were grown in the glasshouses at the University of Southampton. This included the domesticated tomato (D) *Solanum lycopersicum* (SLL; n=5) and *S. lycopersicum* var *cerasiforme* (SLC; n=3), the progenitor (P) *S. pimpinellifolium* (SP; n=3) and the never-domesticated tomato wild species (W) *S. cheesmaniae* (SChe; n=3) and *chmielewskii* (SChm; n=3). Seeds were obtained from Tomato Genetic Resource Centre (TGRC; https://tgrc.ucdavis.edu/) and Centre for Genetic Resources the Netherlands (CGN) Wageningen University (https://cgngenis.wur.nl/).

In order to reduce maternal effects associated with the diverse origins of the seeds, plants were grown under the following control conditions: large pots with a soil mix of 2:1 Levington compost (F2+sand) and vermiculite. Fruits from this generation were harvested; seeds were cleaned and stored. For the subsequent generation (plasticity plant trial), two accessions from each species were selected based on shorter generation time and relatively high fruit production.

Seed cleaning

Seed cleaning was carried out as follows: seeds were scooped out from the fruits into a beaker with 3N HCl and left for 30 minutes and then rinsed; seeds were then placed back into the beaker with 10% Trisodium phosphate (TSP) for 20 minutes and then rinsed; for drying, seeds were placed in a drying oven at 40C overnight. Cleaned seeds were stored at 4°C until needed.

Relationships between phenotypic traits

To identify the relationship between traits, a correlation matrix using the Spearman rank correlation test at $p < 0.05$ was generated. To test for the overlap between differential and plastic traits in the progenitor, Fisher's exact test was performed at $p < 0.05$. All statistical tests and graphs were generated using R v.4.2.1 (R Core Team, 2022).

Gene expression analyses: Exploration of global and pairwise comparison to analyse the leaf plasticity dataset

The within-treatment variability across the species differs which suggests that the dataset should be subset prior to analysis to get a more accurate per-gene dispersion estimate for each comparison (see the DESEq2 vignette, https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html). Pairwise analysis of each pair of treatments per species was therefore performed to explore if the proportion of plastic genes identified changes among species compared to a global analysis (analysis of all the data with the three species and three treatments).

Comparisons between datasets

Three datasets were created to identify the following: (i) differentially expressed genes (DEGs) in leaves and fruits between species under control treatments [interspecific: 3 species x 1 treatment x 2 tissues], (ii) genes with plasticity in leaves (i.e. DEGs between treatments) among species [plasticity: 3 species x 3 treatments x 1 tissue], (iii) overlap between genes with plasticity in SLL, SP or SChe (in leaves and fruits) and divergent between SLL and SP ["plasticity divergence": 3 species x 3 treatments x 2 tissues]. The 'Plasticity divergence' analysis uses the 'plasticity' dataset for leaf analysis and a subset of SLL and SP for fruit analysis ('fruit' dataset: 2 species x 3 treatments x 1 tissue), because plasticity analysis for SChe cannot be performed due to lack of fruits from root crowding and low nutrient treatments. These three datasets were compared.

To assess whether plastic genes were more likely to diverge during domestication, a chi-squared test was performed for leaf and fruit at $p < 0.05$ in the plasticity divergence dataset.

Appendix Results A. 1

Relationships between phenotypic traits

Trait variability between individuals within species was larger across PC1 than PC2 (Appendix Figure A. 2); seed yield contributes the most to PC1 with 98.5%, whilst the largest contributor to PC2 was the height at fruiting with 49.2 % (Appendix Figure A. 3; Table A7). Seed yield was very variable

within species as the metrics it was derived from the number of fruit and the number of seeds, which were very variable within each species.

For the plasticity dataset, the variability is mostly explained by PC1 than PC2 (Appendix Figure A. 4), with seed yield contributing strongly to PC1 (96.5%) and height at fruiting contributing the most to PC2 (52.8%; Appendix Figure A. 5; Table A9). A correlation matrix identified traits that were highly correlated (Appendix Figure A. 1; Table A2). The height metrics were positively correlated (rho= 0.90, $p < 0.01$), as were the number of fruits and seed yield (rho= 0.80, $p < 0.01$). Fruit weight, fruit yield, fruit perimeter, fruit area, fruit width, fruit pericarp area, fruit pericarp thickness and seed size were all positively correlated (rho = 0.70, $p < 0.01$).

Gene expression analyses: Exploration of global and pairwise comparison to analyse the leaf plasticity dataset

The number of plastic genes in domesticated and progenitor increased in the pairwise analysis compared to the global analysis from 12 to 27 and from 823 to 1961 respectively (Table A1). For wild, the number of plastic genes identified decreased with the pairwise analysis compared to the global analysis from 67 to 46 (Table A14). This discrepancy was expected as subsetting the data changes the per-gene dispersion estimate for each comparison altering the power of each analysis. Compared to the global analysis, the general pattern of progenitor having a greater number of plastic genes, followed by wild and then domesticated still stands. With this exploratory analysis in mind, the subsequent analyses were performed using the global analysis outputs.

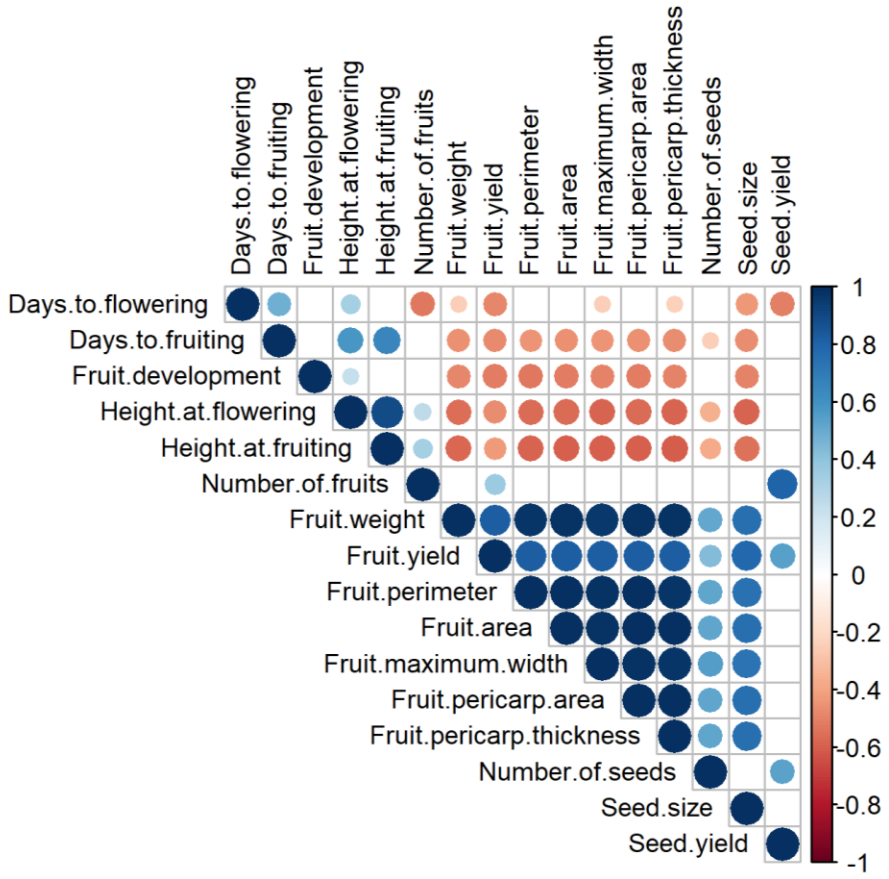
Comparisons between datasets

Species comparison in the ‘plasticity’ analysis (Table A11) revealed a similar patterns compared with the ‘interspecific’ analysis (Table A15), with the greatest number of DEGs found in progenitor vs wild, followed by domesticated vs wild and then domesticated vs progenitor. More DEGs were identified in the ‘plasticity’ dataset compared to the ‘interspecific’ dataset. This was also true for associated GO terms apart from domesticated vs progenitor with no significant GO term in the ‘plasticity’ dataset. There were 16 and 55 significant GO terms in the ‘plasticity’ dataset for domesticated vs wild and progenitor vs wild, respectively; eight of which were shared with its equivalent ‘interspecific’ comparison. Additional GO terms in both species

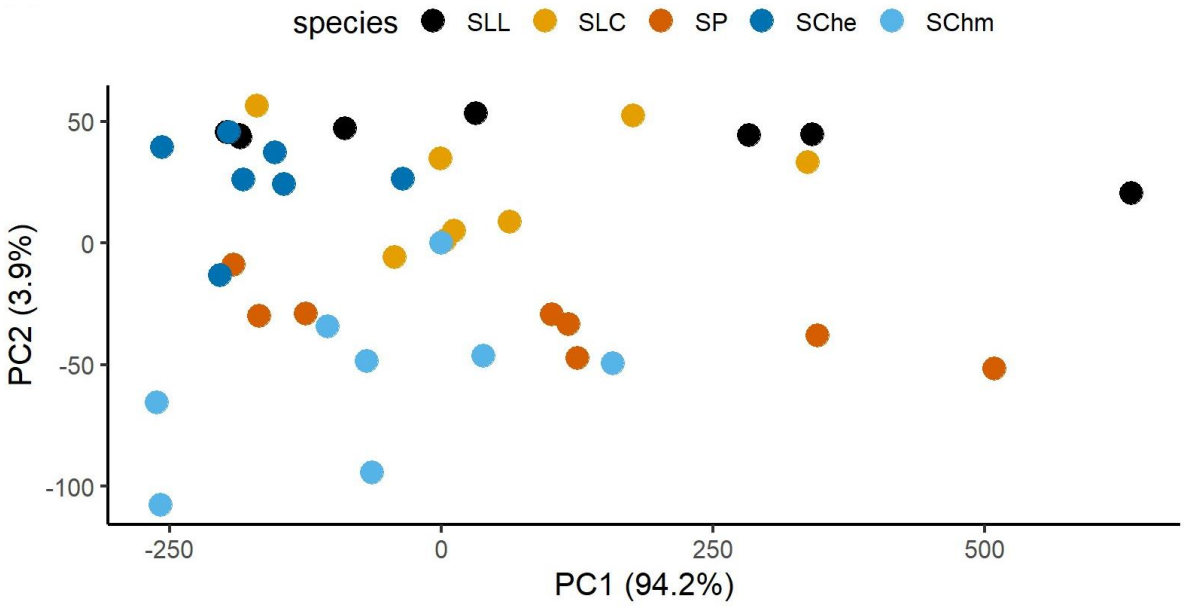
comparisons included response to stress and response to heat, most likely due to the inclusion of the stress treatments in the 'plasticity' dataset. There were no significant KEGG pathways for domesticated vs progenitor and progenitor vs wild, the same for its equivalent 'interspecific' comparison. For domesticated vs wild in the 'plasticity' analysis, there was one significant KEGG pathway, "Protein processing in endoplasmic reticulum", which was shared with its equivalent 'interspecific' comparison.

For the plasticity divergence dataset, there were 256 (Table A15) and 1655 (Table A21) DEGs in domesticated vs progenitor for leaf and fruit respectively, which are more than the number of DEGs identified in 'interspecific' dataset (Table A11). The leaf DEGs resulted in zero significant GO terms compared to nine in the 'interspecific' comparison (Table A12). The fruit DEGs resulted in 115 significant GO terms (Table A22), 55 were shared with its equivalent 'interspecific' dataset including response to external stimulus and related terms (Table A12). No significant KEGG pathways were associated with the leaf DEGs (as found in the equivalent 'interspecific' comparison), and five significant KEGG pathways in fruit (Table A23), three were shared with the equivalent 'interspecific' comparison including "Metabolic pathways", "Biosynthesis of secondary metabolites" and "Proteasome" (Table A13).

The overlap between divergent and plastic genes between species was explored to assess whether plastic genes were more likely to diverge between domesticated and progenitor. Genes divergent between progenitor and domesticated tended to be plastic within each species, with the greatest overlap found in progenitor. The greatest contribution to χ^2 for all comparisons was the positive association between divergent and plastic genes (Appendix Figure A. 8).

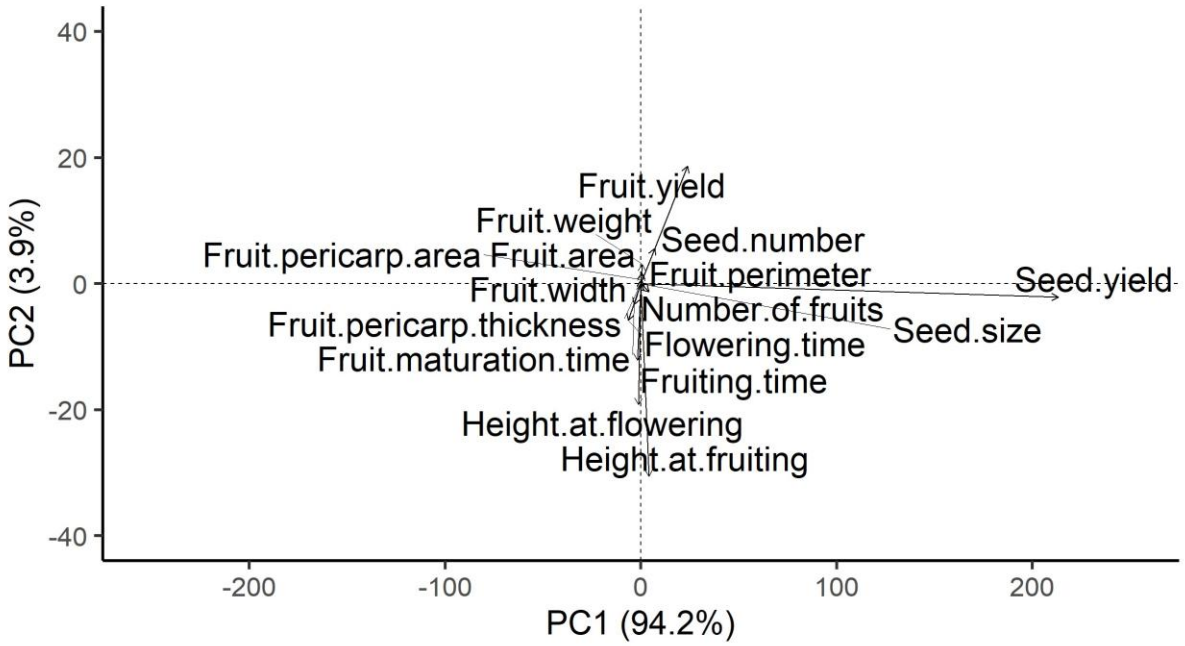


Appendix Figure A. 1: Correlation matrix with Spearman rank correlation test.

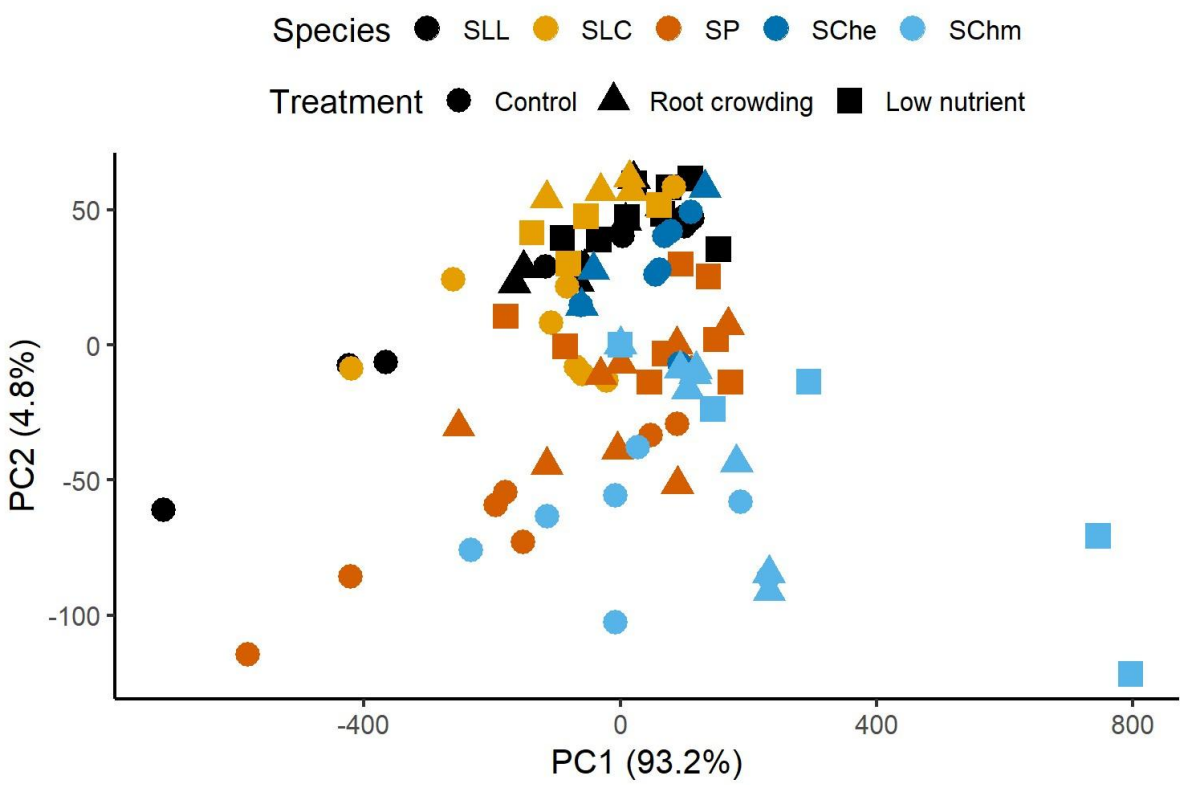


Appendix Figure A. 2: PCA of the interspecific analysis for phenotypic traits.

Principal component analysis (PCA) with the first two principal components shown for *S. lycopersicum* (SLL) and *S. lycopersicum* var *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm) under the control treatment.

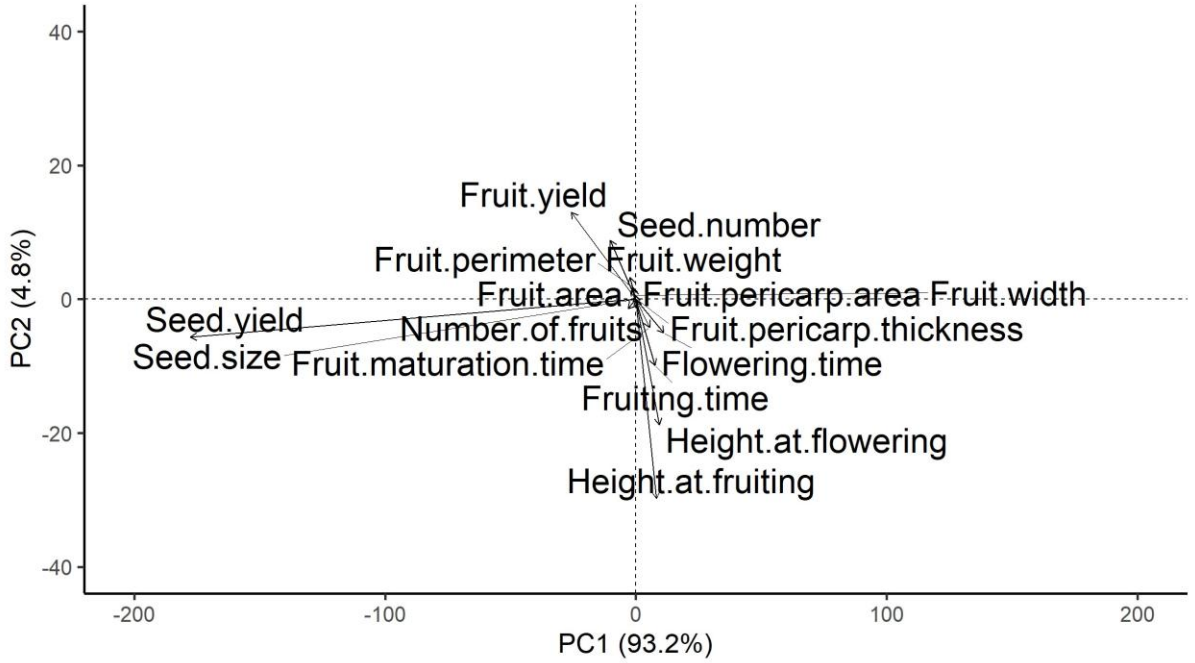


Appendix Figure A. 3: Variable correlation plot of phenotypic traits measured for the interspecific dataset.

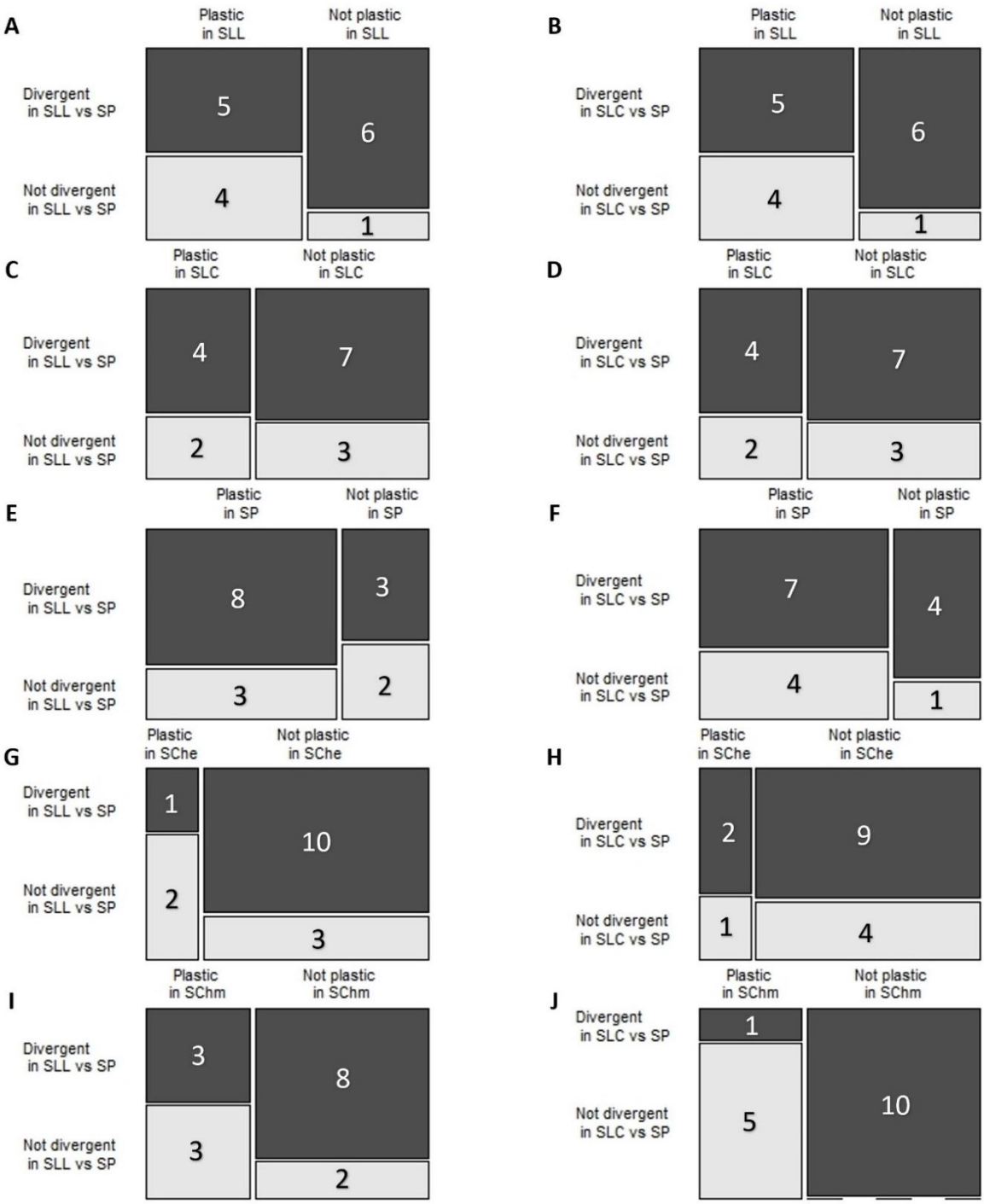


Appendix Figure A. 4: PCA for plasticity analysis of phenotypic traits.

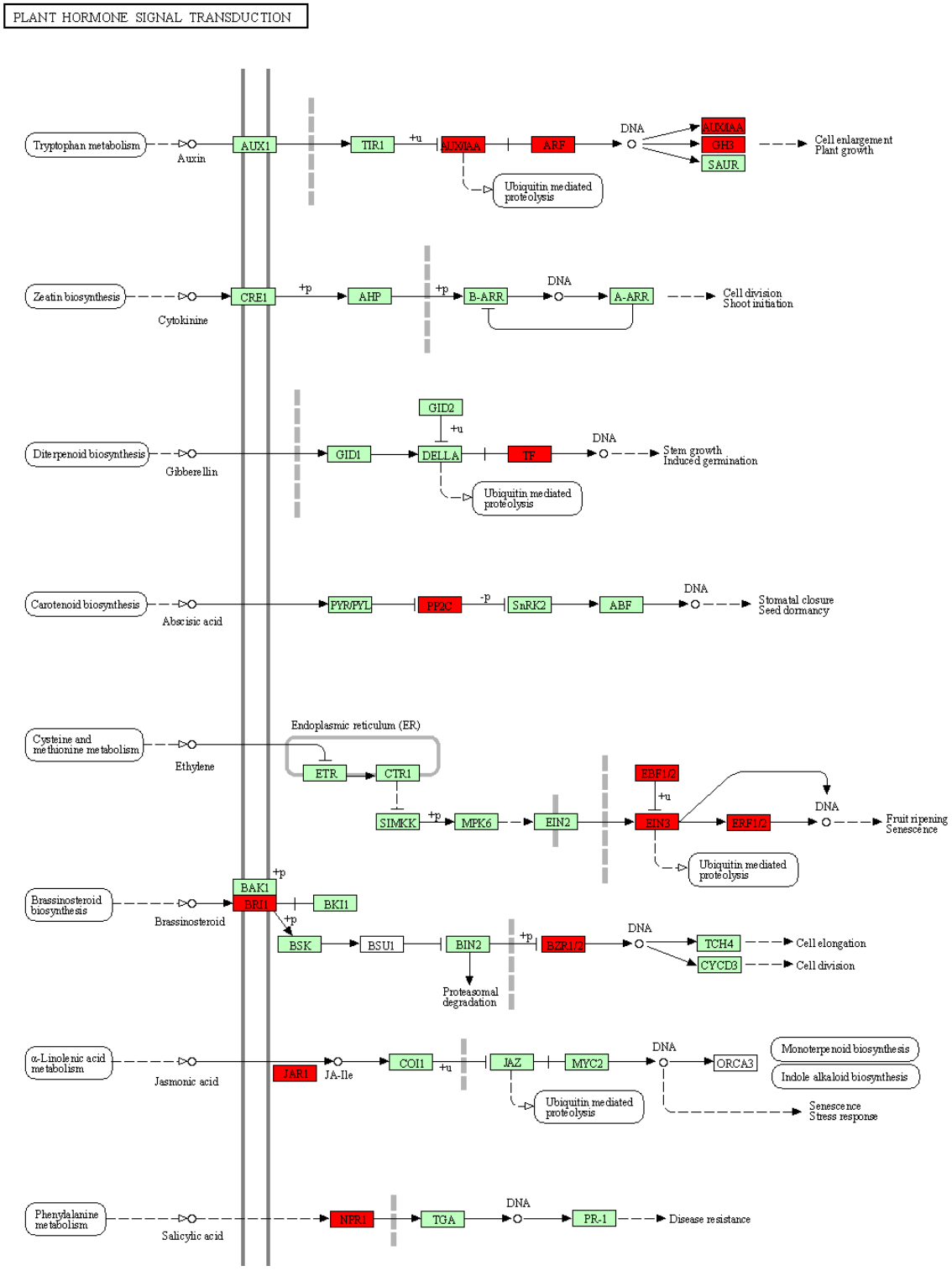
Principal component analysis (PCA) with the first two principal components shown for the morphological traits of *S. lycopersicum* (SLL) and *S. lycopersicum* var *cerasiforme* (SLC), *S. pimpinellifolium* (SP), *S. cheesmaniae* (SChe) and *S. chmielewskii* (SChm) under three conditions.



Appendix Figure A. 5: Variable correlation plot of phenotypic traits measured for the plasticity dataset.

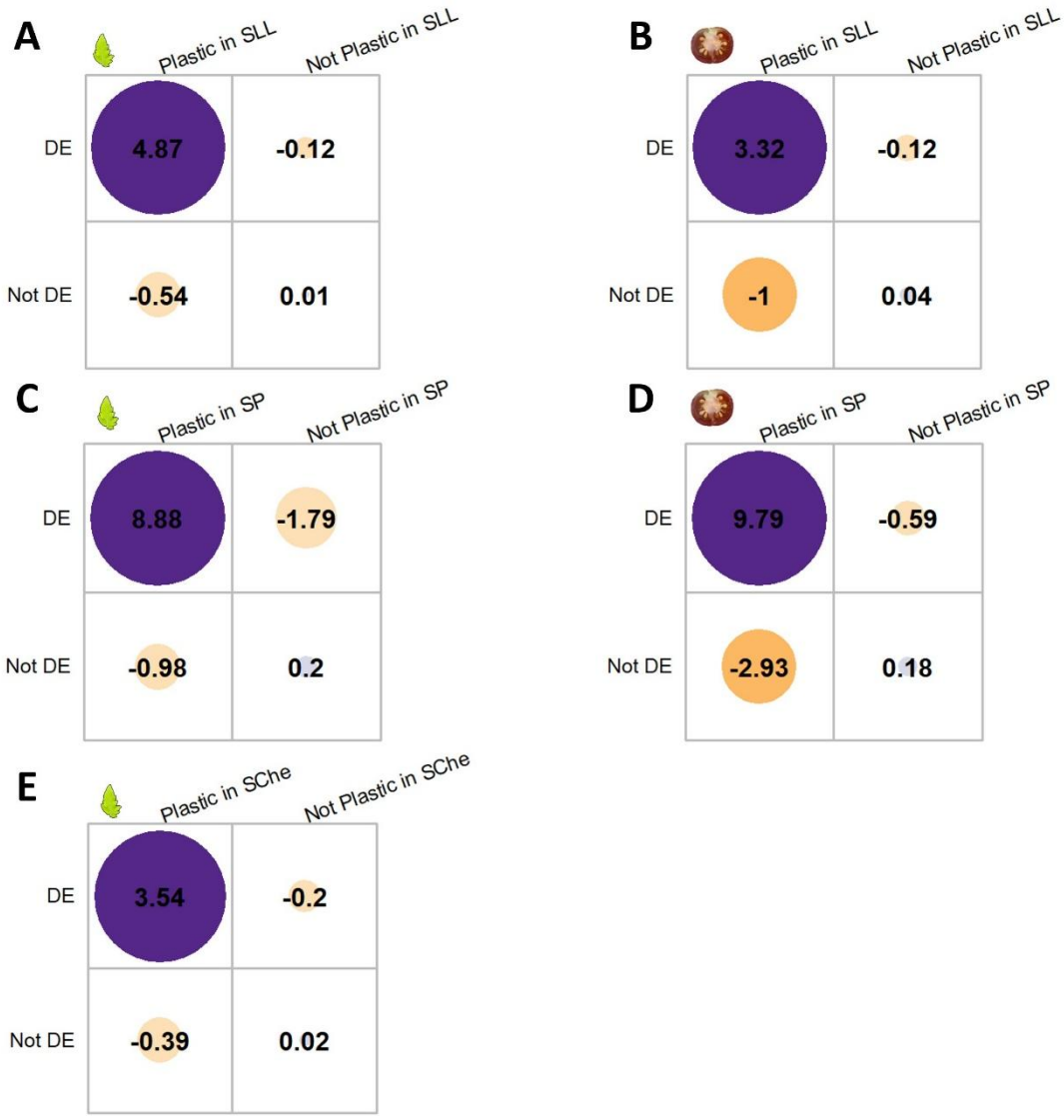


Appendix Figure A. 6: Mosaic plots of overlap between plastic and divergent traits. Plastic traits that may be selected during domestication are plastic in each species and divergent between the progenitor (P; *S. pimpinellifolium*, SP) and the domesticates (D; *S. lycopersicum*, SLL and *S. lycopersicum* var *cerasiforme*, SLC). We assessed the overlap between divergent traits between D vs P and plastic traits in (A,B) SLL, (C,D) SLC, (E,F) SP, (G,H) *S. cheesmaniae* (SChe) and (I,J) *S. chmielewskii* (SChm).



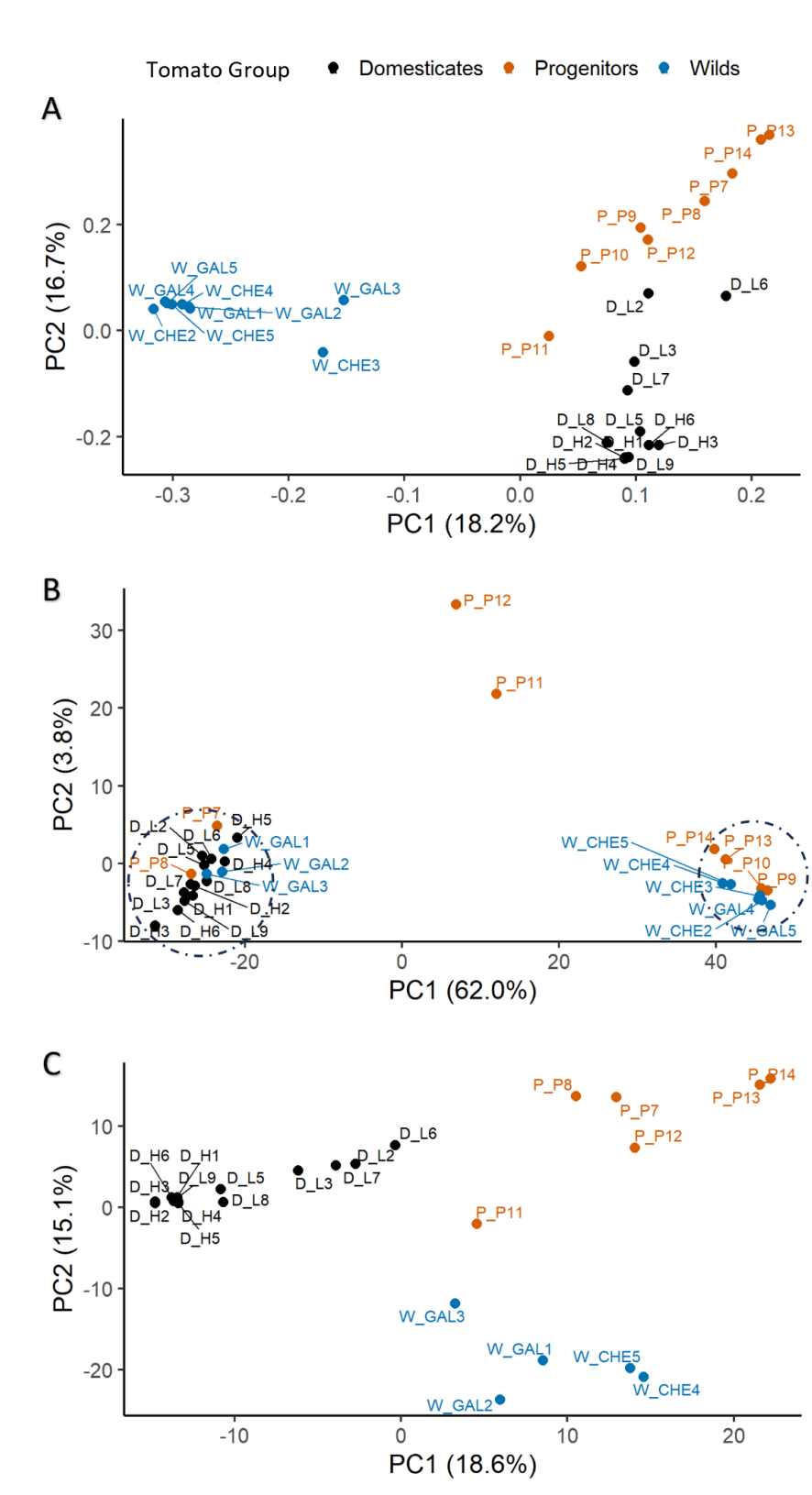
Appendix Figure A. 7: Plant hormone signal transduction pathway identified in the plasticity analysis. Plant hormone signal transduction pathway of plastic genes only found in the progenitor *Solanum pimpinellifolium*. White boxes are the enzymes and reactions in the metabolic pathways. Green boxes are genes in the reference pathway, indicating the presence

of genes in the *S. lycopersicum* (tomato). Red boxes are input plastic genes. Flowchart obtained from KEGG PATHWAY Database (https://www.genome.jp/kegg/pathway.html).



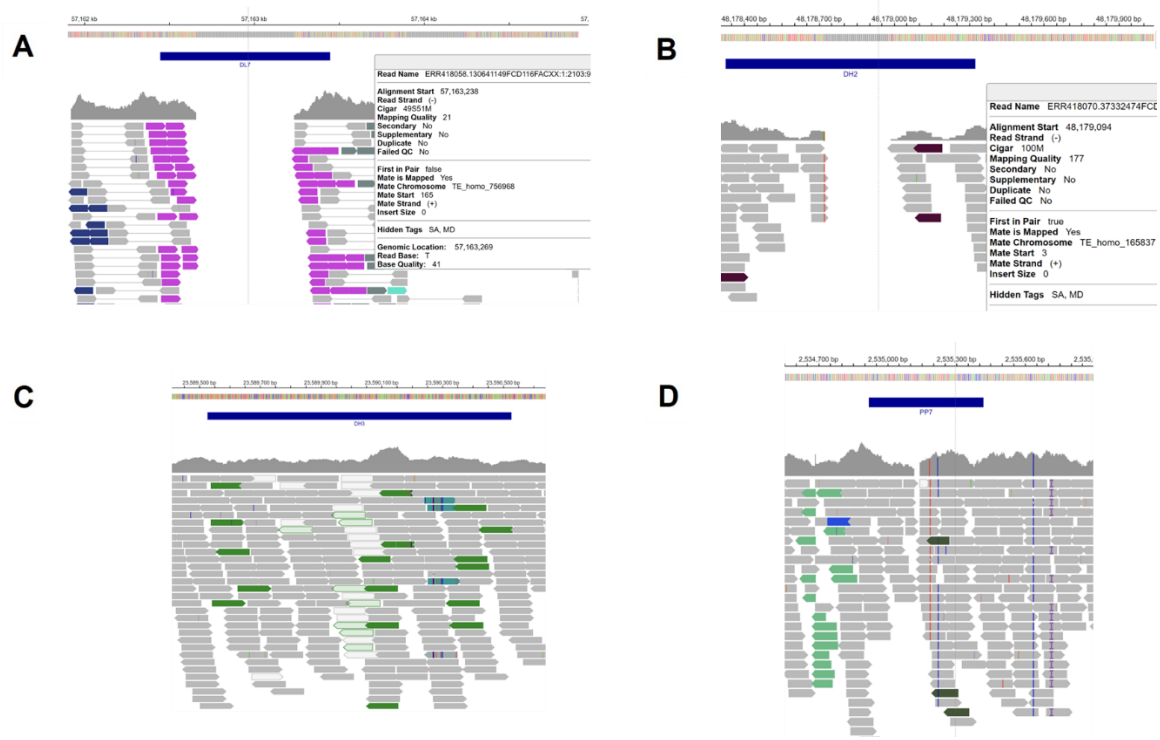
Appendix Figure A. 8: Matrix illustrating the contribution to the association between divergence in SLL vs SP and plasticity in SLL (A) leaf and (B) fruit; SP (C) leaf and (D) fruit; and (E) SChe leaf. The circle is proportional to the contribution to the χ2 test; purple depicts positive association and orange depicts negative association. Note *S. lycopersicum* (SLL), *S. pimpinellifolium* (SP), and *S. cheesmaniae* (SChe).

Appendix B Chapter 3

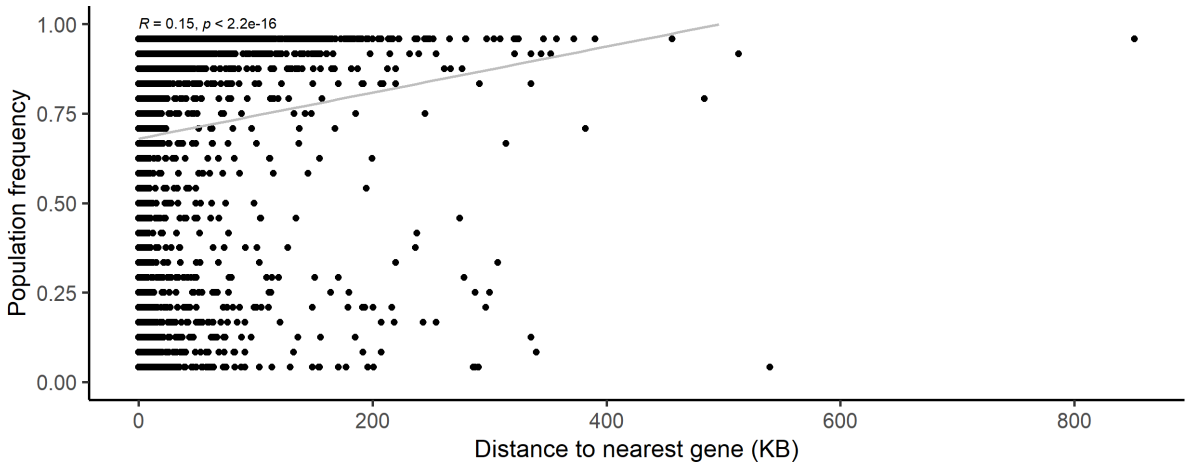


Appendix Figure B. 1: SNP and TIP PCA.

Principal component analysis (PCA) based on (A) Single Nucleotide Polymorphisms (SNPs) and (B) Transposon Insertion Polymorphisms (TIPs) from domesticates (n=13), progenitors (n=8) and never-domesticated wild (n=9) tomatoes. Samples clustered based on insert size with small insert size on the right and large insert size on the left. (C) PCA after removing samples with short insert size and change of signature window setting from minimumSampleMedian to fix500 with domesticates (n=13), progenitors (n=6) and never-domesticated wild (n=5) tomatoes.



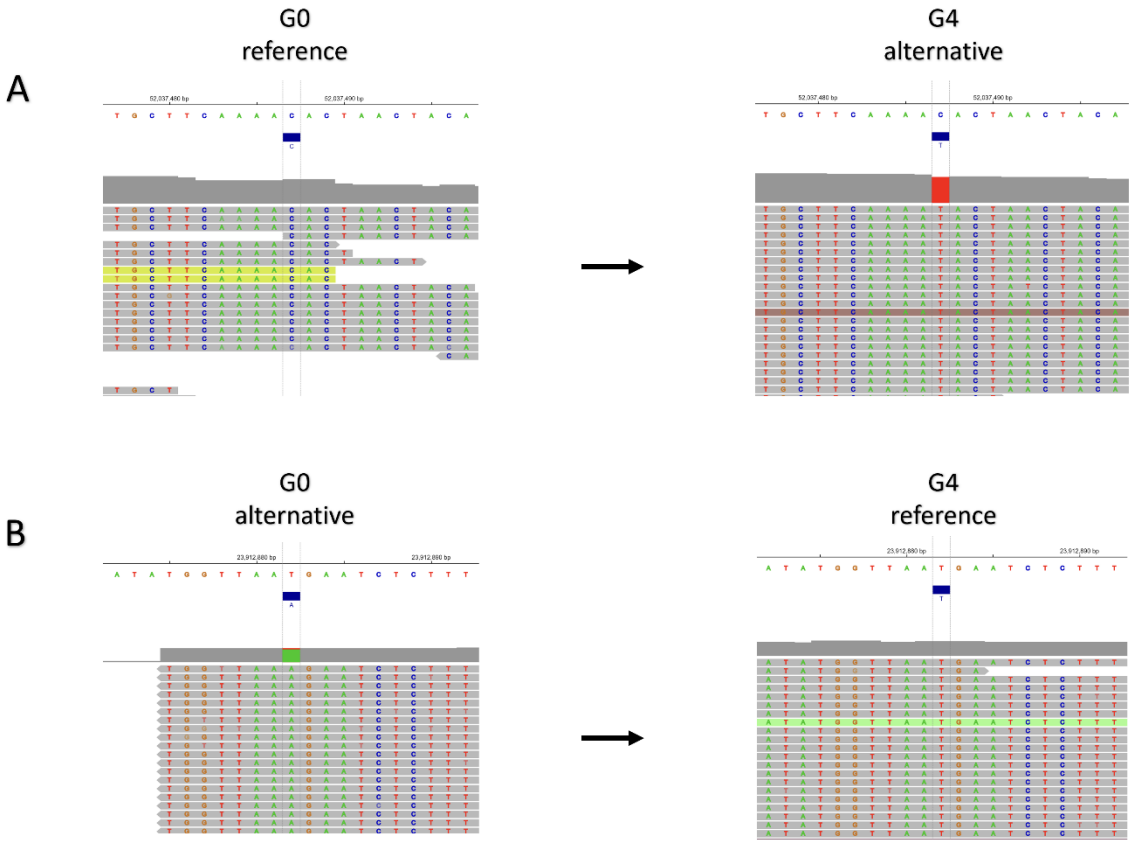
Appendix Figure B. 2: Graphical output of Integrative Genomics Viewer (IGV) depicting examples of a presence and absence of a TE insertion. (A) Presence of a reference TE insertion is illustrated as a gap in the reference genome (masked TE sequence) with one member of several pairs of reads (purple) mapping on the chromosome and the other to the TE sequence. (B) Absence of a reference TE insertion is illustrated as a gap in the reference genome (masked TE sequence) with no or few reads mapping to its corresponding TE sequence. (C) Prescence of a non-reference TE insertion is illustrated as no gap in the reference genome with one member of a pair of reads mapping on the chromosome and the other to the TE sequence (green reverse and forward reads). (D) Absence of a non-reference TE insertion is illustrated as no gap in the reference genome with no or few reads mapping to its corresponding TE sequence.



Appendix Figure B. 3: Changes in TIP frequency relative to genic regions.

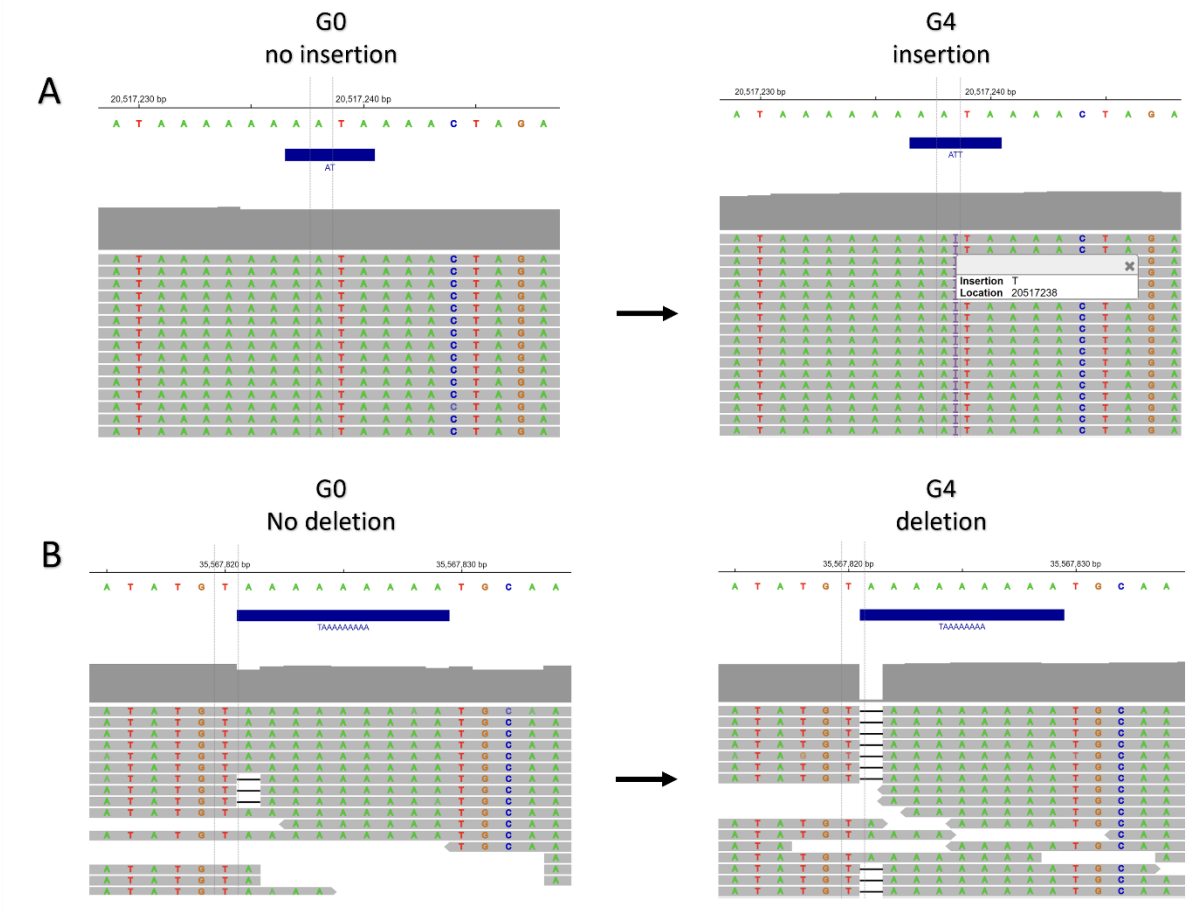
The distribution of all transposable element insertions polymorphisms (TIPs) relative to genic regions and their population frequency.

Appendix C Chapter 4



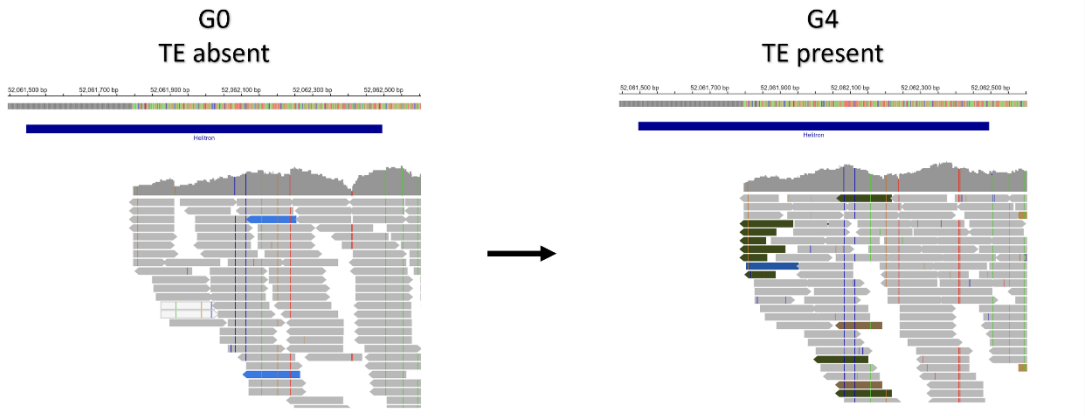
Appendix Figure C. 1: SNV mutation validation using IGV.

Graphical output of IGV depicting examples of single nucleotide variant (SNV) calls. Example of a mutation change from (A) Reference to alternative and (B) Alternative to reference.



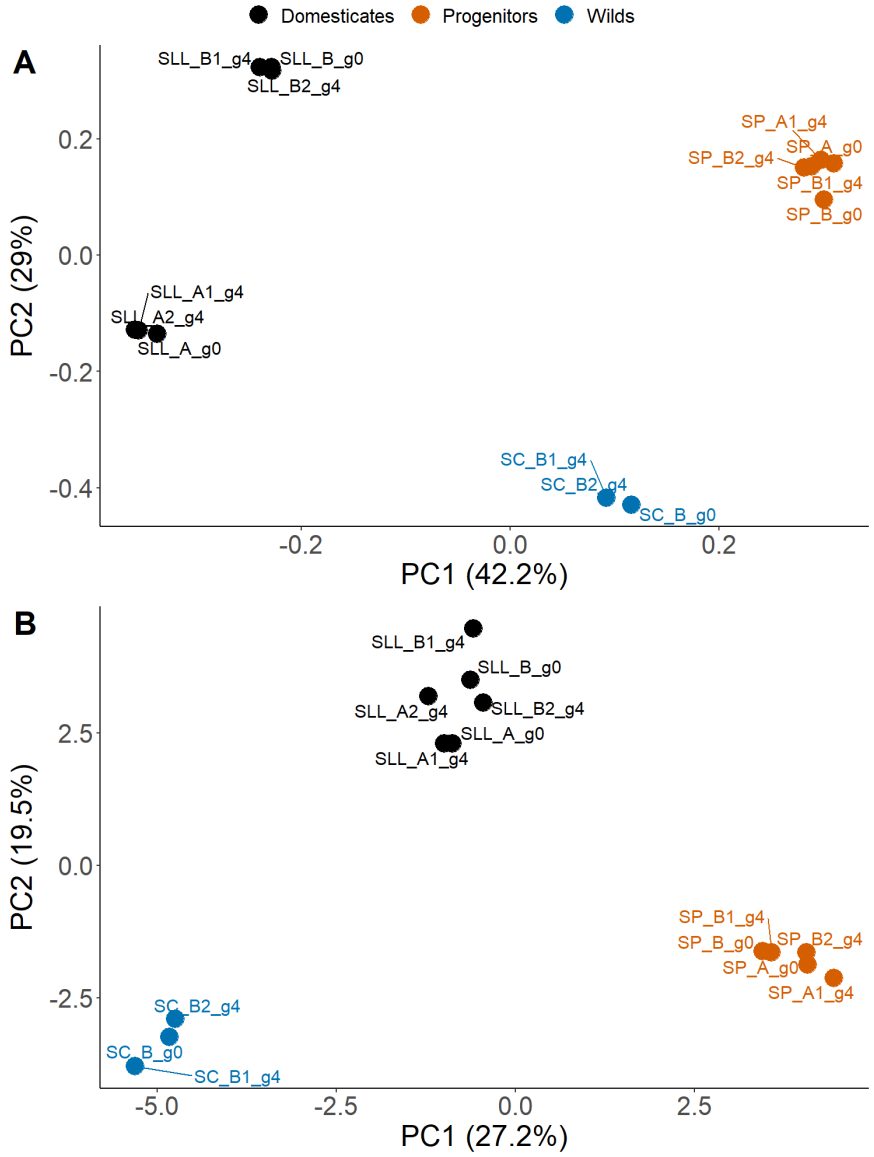
Appendix Figure C. 2: Indel mutation validation using IGV.

Graphical output of IGV depicting examples of indel calls: (A) insertion and (B) deletion.

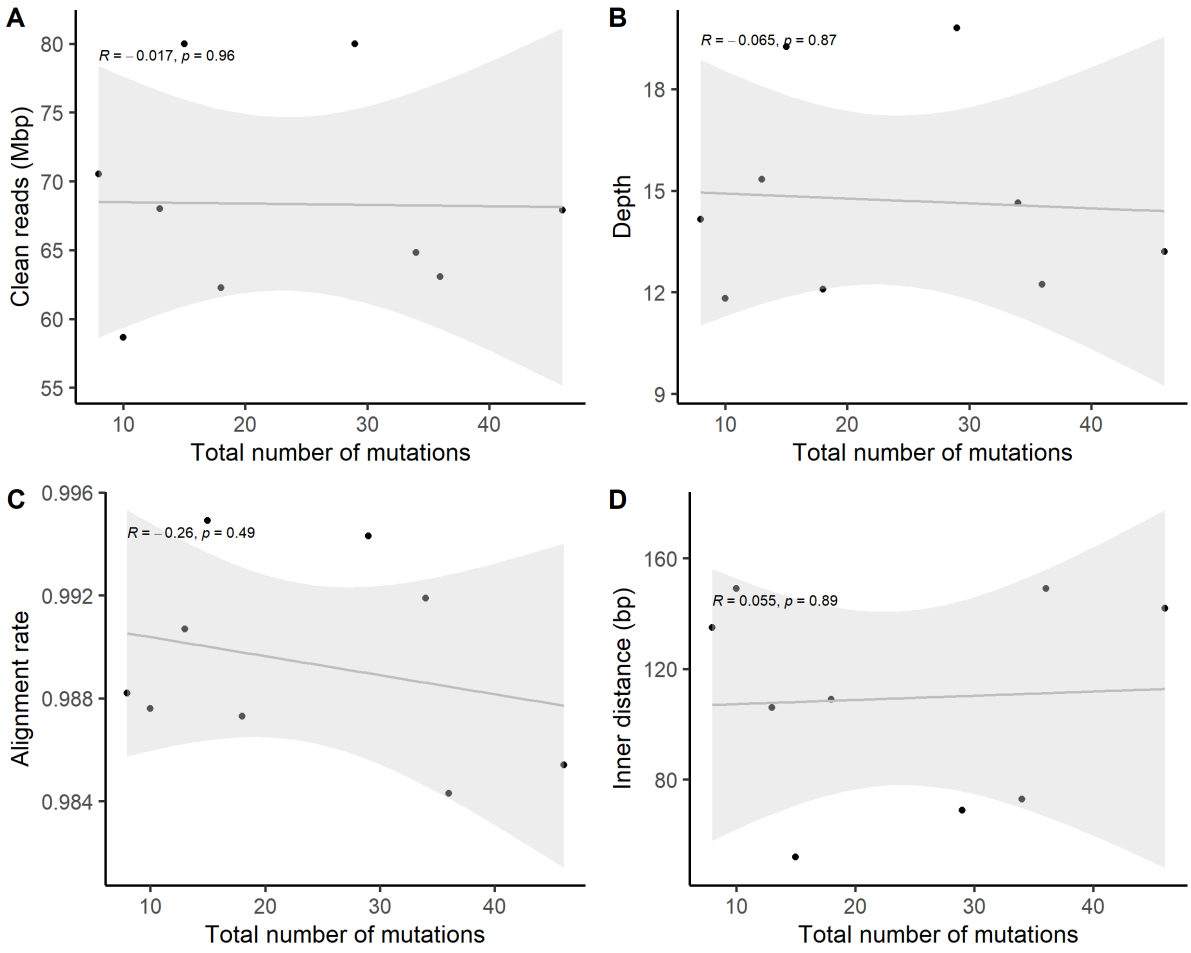


Appendix Figure C. 3: Graphical output of IGV depicting examples of a presence and absence of a TE insertion.

TE insertion depicted by a TE insertion illustrated by an absence of a TE support in generation 0 (blue reads depicts misalignment from other chromosomes) and a presence of a TE support in generation 4 (green reads mapped to this chromosome and TE sequence).



Appendix Figure C. 4: Principal component analysis (PCA) of different mutation types.
(A) short variants (indel and SNPs) and (B) transposable elements (TEs) in domesticated tomato (SLL), wild progenitor (SP) and never-domesticated wild (SC) from generation G0 and G4.



Appendix Figure C. 5: Correlation plots between total mutations and (A) Clean reads, (B) depth, (C) alignment rate and (D) inner distance.

Appendix Figure C. 6: Chi-squared association plots.

Matrix illustrating the contribution to the association between(A) mutations and chromosome 1 to 12, and between species and (B) nucleotide bias, (C) chromosome arm and pericentromeric regions (D) intergenic, up/downstream and genic regions. Each circle is proportional to the contribution to the $\chi 2$ test; purple depicts positive association and orange depicts negative association.

List of References

Abbo, S., Rachamim, E., Zehavi, Y., Zezak, I., Lev-Yadun, S. & Gopher, A. 2011. Experimental growing of wild pea in Israel and its bearing on Near Eastern plant domestication. *Annals of Botany,* 107**,** 1399-1404

Abiraami, T., Sanyal, R. P., Misra, H. S. & Saini, A. 2023. Genome-wide analysis of bromodomain gene family in arabidopsis and rice. *Frontiers in Plant Science,* 14**,** 1120012

Acasuso-Rivero, C., Murren, C. J., Schlichting, C. D. & Steiner, U. K. 2018. Adaptive phenotypic plasticity for life history and less fitness-related traits. *bioRxiv***,** 367284

Adrion, J. R., Song, M. J., Schrider, D. R., Hahn, M. W. & Schaack, S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in Drosophila melanogaster. *Genome biology and evolution,* 9**,** 1329-1340

Aflitos, S., Consortium, T. G. S., Schijlen, E., Hans De Jong, Dick De Ridder, Sandra Smit, Richard Finkers, Jun Wang, Gengyun Zhang, Ning Li, Likai Mao, Freek Bakker, Rob Dirks, Timo Breit, Barbara Gravendeel, Henk Huits, Darush Struss, Ruth Swanson-Wagner, Hans Van Leeuwen, Roeland C H J Van Ham, Laia Fito, L. G., Myrna Sevilla, Philippe Ellul, Eric Ganko, Arvind Kapur, Emannuel Reclus, Bernard De Geus, Henri Van De Geest, Bas Te Lintel Hekkert, Jan Van Haarst, Lars Smits, Andries Koops, Gabino Sanchez-Perez, Adriaan W Van Heusden, Richard Visser, Zhiwu Quan, Jiumeng Min, Li Liao, Xiaoli Wang, Guangbiao Wang, Zhen Yue, Xinhua Yang, Na Xu, E. S., Erik Smets, Rutger Vos, Johan Rauwerda, Remco Ursem, Cees Schuit, Mike Kerns, Jan Van Den Berg, Wim Vriezen, Antoine Janssen, Erwin Datema, Torben Jahrman, Frederic Moquet, Julien Bonnet & Peters, S. 2014. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.,* 80

Alam, O. & Purugganan, M. D. 2024. Domestication and the evolution of crops: variable syndromes, complex genetic architectures, and ecological entanglements. *The Plant Cell***,** koae013

Ali, M. S. & Baek, K.-H. 2020. Jasmonic acid signaling pathway in response to abiotic stresses in plants. *International Journal of Molecular Sciences,* 21**,** 621

Allaby, R. G., Kitchen, J. L. & Fuller, D. Q. 2015. Surprisingly low limits of selection in plant domestication. *Evolutionary Bioinformatics,* 11**,** EBO. S33495

Allaby, R. G., Stevens, C. J., Kistler, L. & Fuller, D. Q. 2022. Emerging evidence of plant domestication as a landscape-level process. *Trends in ecology & evolution,* 37**,** 268-279

Allaby, R. G., Ware, R. L. & Kistler, L. 2019. A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evolutionary Applications,* 12**,** 29-37

Alleman, M. & Freeling, M. 1986. The Mu transposable elements of maize: evidence for transposition and copy number regulation during development. *Genetics,* 112**,** 107-119

Almojil, D., Bourgeois, Y., Falis, M., Hariyani, I., Wilcox, J. & Boissinot, S. 2021. The structural, functional and evolutionary impact of transposable elements in eukaryotes. *Genes,* 12**,** 918

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z.,

Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., Aganezov, S., Ranallo-Benavidez, T. R., Lemmon, Z. H., Kim, J., Robitaille, G., Kramer, M., Goodwin, S., Mccombie, W. R., Hutton, S., Van Eck, J., Gillis, J., Eshed, Y., Sedlazeck, F. J., Van Der Knaap, E., Schatz, M. C. & Lippman, Z. B. 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell,* 182**,** 145-161 e23.10.1016/j.cell.2020.05.021

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol,* 21**,** 30.10.1186/s13059-020-1935-5

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I. & Pommier, C. 2019. RepetDB: a unified resource for transposable element references. *Mobile DNA,* 10**,** 1-8

Andersson, L. & Purugganan, M. 2022. Molecular genetic variation of animals and plants under domestication. *Proceedings of the National Academy of Sciences,* 119**,** e2122150119

Andrews, S. 2010. *FastQC: a quality control tool for high throughput sequence data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed].

Araus, J. L., Ferrio, J. P., Voltas, J., Aguilera, M. & Buxó, R. 2014. Agronomic conditions and crop evolution in ancient Near East agriculture. *Nature Communications,* 5**,** 3953

Arndt, P. F., Hwa, T. & Petrov, D. A. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *Journal of molecular evolution,* 60**,** 748-763

Baduel, P., Leduque, B., Ignace, A., Gy, I., Jr, J. G., Loudet, O., Colot, V. & Quadrana, L. 2021. Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana. *Genome Biology,* 22**,** 138

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**,** 1085-94.10.1093/aob/mcm150

Bajus, M., Macko-Podgórni, A., Grzebelus, D. & Baránek, M. 2022. A review of strategies used to identify transposition events in plant genomes. *Frontiers in Plant Science,* 13**,** 1080993

Baldwin, J. M. 1896. A new factor in evolution. *The American Naturalist,* 30**,** 441–451

Bao, W., Kojima, K. K. & Kohany, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna,* 6**,** 1-6

Barg, R., Sobolev, I., Eilon, T., Gur, A., Chmelnitsky, I., Shabtai, S., Grotewold, E. & Salts, Y. 2005. The tomato early fruit specific gene Lefsm1 defines a novel class of plant-specific SANT/MYB domain proteins. *Planta,* 221**,** 197-211

Bari, R. & Jones, J. D. 2009. Role of plant hormones in plant defence responses. *Plant molecular biology,* 69**,** 473-488

Barrett, R. D. & Schluter, D. 2008. Adaptation from standing genetic variation. *Trends in ecology & evolution,* 23**,** 38-44

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A. & González, J. 2014. Population genomics of transposable elements in Drosophila. *Annual review of genetics,* 48**,** 561-581

Basu, S. & Groot, S. P. 2023. Seed Vigour and Invigoration. *Malavika Dadlani***,** 67

Bégin, M. & Schoen, D. J. 2006. Low impact of germline transposition on the rate of mildly deleterious mutation in Caenorhabditis elegans. *Genetics,* 174**,** 2129-2136

Belcher, M. E., Williams, D. & Mueller, N. G. 2023. Turning Over a New Leaf: Experimental Investigations into the Role of Developmental Plasticity in the Domestication of Goosefoot (Chenopodium berlandieri) in Eastern North America. *American Antiquity***,** 1-16

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological),* 57**,** 289-300

Benoit, M., Drost, H.-G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D. & Paszkowski, J. 2019. Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLOS Genetics,* 15**,** e1008370

Bergougnoux, V. 2014. The history of tomato: from domestication to biopharming. *Biotechnol Adv,* 32**,** 170-89.10.1016/j.biotechadv.2013.11.003

Besser, K., Harper, A., Welsby, N., Schauvinhold, I., Slocombe, S., Li, Y., Dixon, R. A. & Broun, P. 2009. Divergent regulation of terpenoid metabolism in the trichomes of wild and cultivated tomato species. *Plant physiology,* 149**,** 499-514

Bishop, K. A., Betzelberger, A. M., Long, S. P. & Ainsworth, E. A. 2015. Is there potential to adapt soybean (G lycine max M err.) to future [CO2]? An analysis of the yield response of 18 genotypes in free-air CO2 enrichment. *Plant, Cell & Environment,* 38**,** 1765-1774

Bitarello, B. D., Brandt, D. Y., Meyer, D. & Andrés, A. M. 2023. Inferring balancing selection from genome-scale data. *Genome biology and evolution,* 15**,** evad032

Blanca, J., Cañizares, J., Cordero, L., Pascual, L., Diez, M. J. & Nuez, F. 2012. Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. *PloS one,* 7**,** e48198

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., Francis, D., Causse, M., Knaap, E. V. D. & Cañizares, J. 2015. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics,* 16**,** 257

Bogaard, A., Fraser, R., Heaton, T. H., Wallace, M., Vaiglova, P., Charles, M., Jones, G., Evershed, R. P., Styring, A. K. & Andersen, N. H. 2013. Crop manuring and intensive land management by Europe's first farmers. *Proceedings of the National Academy of Sciences,* 110**,** 12589-12594

Bogaerts-Márquez, M., Barron, M. G., Fiston-Lavier, A.-S., Vendrell-Mir, P., Castanera, R., Casacuberta, J. M. & González, J. 2020. T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics,* 36**,** 1191-1197

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.10.1093/bioinformatics/btu170

Bonchev, G. & Willi, Y. 2018. Accumulation of transposable elements in selfing populations of Arabidopsis lyrata supports the ectopic recombination model of transposon evolution. *New Phytologist,* 219**,** 767-778

Bonduriansky, R. 2012. Rethinking heredity, again. *Trends in ecology & evolution,* 27**,** 330-336

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvak, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L. & Feschotte, C. 2018. Ten things you should know about transposable elements. *Genome Biol,* 19**,** 199.10.1186/s13059-018-1577-z

Bracci, A. N., Dallmann, A., Ding, Q., Hubisz, M. J., Caballero, M. & Koren, A. 2023. The evolution of the human DNA replication timing program. *Proceedings of the National Academy of Sciences,* 120**,** e2213896120

Bradshaw, A. D. 1965. Evolutionary significance of phenotypic plasticity in plants. *Advances in genetics,* 13**,** 115-155

Brooker, R., Brown, L. K., George, T. S., Pakeman, R. J., Palmer, S., Ramsay, L., Schöb, C., Schurch, N. & Wilkinson, M. J. 2022. Active and adaptive plasticity in a changing climate. *Trends in Plant Science,* 27**,** 717-728

Brozynska, M., Furtado, A. & Henry, R. J. 2016. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol J,* 14**,** 1070-85.10.1111/pbi.12454

Burgarella, C. & Glémin, S. 2017. Population genetics and genome evolution of selfing species. John Wiley & Sons Ltd Chichester

Cambiaso, V., Raúlpratta, G., Costa, J. H. P. D., Zorzoli, R., Merrilfrancis, D. & Rodríguez, G. R. 2019. Whole genome re-sequencing analysis of two tomato genotypes for polymorphism insight in cloned genes and a genetic map construction. *Scientia Horticulturae,* 247**,** 58-66

Campbell-Staton, S. C., Velotta, J. P. & Winchell, K. M. 2021. Selection on adaptive and maladaptive gene expression plasticity during thermal adaptation to urban heat islands. *Nature communications,* 12**,** 6195

Campos, L., López-Gresa, M. P., Fuertes, D., Bellés, J. M., Rodrigo, I. & Lisón, P. 2019. Tomato glycosyltransferase Twi1 plays a role in flavonoid glycosylation and defence against virus. *BMC plant biology,* 19**,** 1-17

Capy, P. 2021. Taming, domestication and exaptation: trajectories of transposable elements in genomes. *Cells,* 10**,** 3590

Castanera, R., Morales-Díaz, N., Gupta, S., Purugganan, M. & Casacuberta, J. M. 2023. Transposons are important contributors to gene expression variability under selection in rice populations. *Elife,* 12**,** RP86324

Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M.-C., Panaud, O. & Casacuberta, J. M. 2021. Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *The Plant Journal,* 107**,** 118-135

Causse, M., Desplat, N., Pascual, L., Paslier, M.-C. L., Sauvage, C., Bauchet, G., Bérard, A., Bounon, R., Tchoumakov, M., Brunel, D. & Bouchet, J.-P. 2013. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics* 14

Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M. & Scheid, O. M. 2014. How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. *PLoS Genet.,* 10**,** e1004115

Chacón-Labella, J., Garcia Palacios, P., Matesanz, S., Schöb, C. & Milla, R. 2019. Plant domestication disrupts biodiversity effects across major crop types. *Ecology Letters,* 22**,** 1472-1482

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience,* 4**,** s13742-015-0047-8

Charlesworth, B. & Charlesworth, D. 2010. Elements of evolutionary genetics.

Charlesworth, B., Lande, R. & Slatkin, M. 1982. A neo-Darwinian commentary on macroevolution. *Evolution,* 36**,** 474-498

Chen, J., Basting, P. J., Han, S., Garfinkel, D. J. & Bergman, C. M. 2023. Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast. *Mobile DNA,* 14**,** 8

Chen, J., Bataillon, T., Glémin, S. & Lascoux, M. 2022. What does the distribution of fitness effects of new mutations reflect? Insights from plants. *New Phytologist,* 233**,** 1613-1619

Chen, Q., Samayoa, L. F., Yang, C. J., Bradbury, P. J., Olukolu, B. A., Neumeyer, M. A., Romay, M. C., Sun, Q., Lorant, A. & Buckler, E. S. 2020. The genetic architecture of the maize progenitor, teosinte, and how it was altered during maize domestication. *PLoS genetics,* 16**,** e1008791

Chen, Y.-F., Etheridge, N. & Schaller, G. E. 2005. Ethylene signal transduction. *Annals of botany,* 95**,** 901-915

Chevin, L.-M. & Lande, R. 2010. When do adaptive plasticity and genetic evolution prevent extinction of a density-regulated population? *Evolution,* 64**,** 1143-1150

Chivasa, S., Tome, D. F., Hamilton, J. M. & Slabas, A. R. 2011. Proteomic analysis of extracellular ATP-regulated proteins identifies ATP synthase β-subunit as a novel plant cell death regulator. *Molecular & Cellular Proteomics,* 10

Cho, J. 2021. *Plant Transposable Elements: Methods and Protocols.*, Springer US.https://doi.org/10.1007/978-1-0716-1134-0

Chu, Y. H., Jang, J. C., Huang, Z. & Van Der Knaap, E. 2019. Tomato locule number and fruit size controlled by natural alleles of lc and fas. *Plant direct,* 3**,** e00142

Chuong, E. B., Elde, N. C. & Feschotte, C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet,* 18**,** 71-86.10.1038/nrg.2016.139

Clement, C. R., Casas, A., Parra-Rondinel, F. A., Levis, C., Peroni, N., Hanazaki, N., Cortés-Zárraga, L., Rangel-Landa, S., Alves, R. P. & Ferreira, M. J. 2021. Disentangling domestication from food production systems in the Neotropics. *Quaternary,* 4**,** 4

Cobben, M. M., Mitesser, O. & Kubisch, A. 2017. Evolving mutation rate advances the invasion speed of a sexual species. *BMC Evolutionary Biology,* 17**,** 1-10

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., Mcpherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L. & Zhang, X. 2016. A survey of best practices for RNA-seq data analysis. *Genome biology,* 17**,** 1-19

Contreras-Soto, R. I., Mora, F., De Oliveira, M. a. R., Higashi, W., Scapim, C. A. & Schuster, I. 2017. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. *PloS one,* 12**,** e0171105

Corl, A., Bi, K., Luke, C., Challa, A. S., Stern, A. J., Sinervo, B. & Nielsen, R. 2018. The genetic basis of adaptation following plastic changes in coloration in a novel environment. *Current Biology,* 28**,** 2970-2977. e7

Cortés, A. J. & López-Hernández, F. 2021. Harnessing crop wild diversity for climate change adaptation. *Genes,* 12**,** 783

Cowley, M. & Oakey, R. J. 2013. Transposable elements re-wire and fine-tune the transcriptome. *PLoS genetics,* 9**,** e1003234

Crow, J. F. 2017. *An introduction to population genetics theory*, Scientific Publishers

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N. & Rhoné, B. 2018. The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. *Current Biology,* 28**,** 2274-2282. e6

Cui, H. & Fedoroff, N. V. 2002. Inducible DNA demethylation mediated by the maize Suppressor-mutator transposon-encoded TnpA protein. *The Plant Cell,* 14**,** 2883-2899

Cunniff, J., Jones, G., Charles, M. & Osborne, C. P. 2017. Yield responses of wild C3 and C4 crop progenitors to subambient CO2: a test for the role of CO2 limitation in the origin of agriculture. *Global Change Biology,* 23**,** 380-393

Cunniff, J., Wilkinson, S., Charles, M., Jones, G., Rees, M. & Osborne, C. P. 2014. Functional traits differ between cereal crop progenitors and other wild grasses gathered in the Neolithic Fertile Crescent. *PLoS One,* 9**,** e87586

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T. & Sherry, S. T. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., Mccarthy, S. A. & Davies, R. M. 2021. Twelve years of SAMtools and BCFtools. *Gigascience,* 10**,** giab008

Dar, F. A., Mushtaq, N. U., Saleem, S., Rehman, R. U., Dar, T. U. H. & Hakeem, K. R. 2022. Role of epigenetics in modulating phenotypic plasticity against abiotic stresses in plants. *International journal of genomics,* 2022**,** 1092894

Darwin, C. 1868. *The variation of animals and plants under domestication*, John murray

Davière, J.-M. & Achard, P. 2013. Gibberellin signaling in plants. *Development,* 140**,** 1147-1151.10.1242/dev.087650

Dempewolf, H., Eastwood, R. J., Guarino, L., Khoury, C. K., Müller, J. V. & Toll, J. 2014. Adapting agriculture to climate change: a global initiative to collect, conserve, and use crop wild relatives. *Agroecology and Sustainable Food Systems,* 38**,** 369-377

Denver, D. R., Wilhelm, L. J., Howe, D. K., Gafner, K., Dolan, P. C. & Baer, C. F. 2012. Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. *Genome Biology and Evolution,* 4**,** 513-522

Desai, M. M. & Fisher, D. S. 2007. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics,* 176**,** 1759-1798

Desai, M. M., Fisher, D. S. & Murray, A. W. 2007. The speed of evolution and maintenance of variation in asexual populations. *Current biology,* 17**,** 385-394

Diamond, J. 2002. Evolution, consequences and future of plant and animal domestication. *Nature,* 418**,** 700-707

Diez, C. M., Meca, E., Tenaillon, M. I. & Gaut, B. S. 2014. Three groups of transposable elements with contrasting copy number dynamics and host responses in the maize (Zea mays ssp. mays) genome. *PLoS genetics,* 10**,** e1004298

Diggle, P. K. & Miller, J. S. 2013. Developmental plasticity, genetic assimilation, and the evolutionary diversification of sexual expression in Solanum. *American journal of botany,* 100**,** 1050-1060

Dingkuhn, M., Luquet, D., Fabre, D., Muller, B., Yin, X. & Paul, M. J. 2020. The case for improving crop carbon sink strength or plasticity for a CO2-rich future. *Current Opinion in Plant Biology,* 56**,** 259-272

Doebley, J. F., Gaut, B. S. & Smith, B. D. 2006. The molecular genetics of crop domestication. *Cell,* 127**,** 1309-1321

Doganlar, S., Frary, A. & Tanksley, S. 2000. The genetic basis of seed-weight variation: tomato as a model system. *Theoretical and Applied Genetics,* 100**,** 1267-1273

Dominguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jimenez-Gomez, J. M., Colot, V. & Quadrana, L. 2020. The impact of transposable elements on tomato diversity. *Nat Commun,* 11**,** 4058.10.1038/s41467-020-17874-2

Dooner, H. K., Wang, Q., Huang, J. T., Li, Y., He, L., Xiong, W. & Du, C. 2019. Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. *Proceedings of the National Academy of Sciences,* 116**,** 10734-10743

Doran, A. G. & Creevey, C. J. 2013. Snpdat: easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC bioinformatics,* 14**,** 1-6

Doyle, J. J. & Doyle, J. L. 1990. Isolation of plant DNA from fresh tissue. *Focus,* 12**,** 13 - 15

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development,* 12**,** 640-649

Dwivedi, S. L., Heslop-Harrison, P., Spillane, C., Mckeown, P. C., Edwards, D., Goldman, I. & Ortiz, R. 2023. Evolutionary dynamics and adaptive benefits of deleterious mutations in crop gene pools. *Trends in Plant Science,* 28**,** 685-697

Ellinghaus, D., Kurtz, S. & Willhoeft, U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics,* 9**,** 1-14

Escobar-Bravo, R., Alba, J. M., Pons, C., Granell, A., Kant, M. R., Moriones, E. & Fernández-Muñoz, R. 2016. A jasmonate-inducible defense trait transferred from wild into cultivated tomato establishes increased whitefly resistance and reduced viral disease incidence. *Frontiers in plant science,* 7**,** 1732

Eyre-Walker, A. & Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics,* 8**,** 610-618

Fao 2017. *The future of food and agriculture- trends and challenges.,* Rome

Fao 2019. Voluntary guidelines for the conservation and sustainable use of farmers' varieties/landraces. FAO Rome

Fao. 2021. *FAOSTAT* [Online]. Available: http://www.fao.org/faostat/en/#data [Accessed 14 July 2021].

Flint-Garcia, S., Feldmann, M. J., Dempewolf, H., Morrell, P. L. & Ross-Ibarra, J. 2023. Diamonds in the not-so-rough: Wild relative diversity hidden in crop genomes. *PLoS biology,* 21**,** e3002235

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. & Smit, A. F. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences,* 117**,** 9451-9457

Fox, R. J., Donelson, J. M., Schunter, C., Ravasi, T. & Gaitán-Espitia, J. D. 2019. Beyond buying time: the role of plasticity in phenotypic adaptation to rapid environmental change. The Royal Society

Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., Knaap, E. V. D., Cong, B., Liu, J., Meller, J., Elber, R. & Alpert, K. B. 2000. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. *Science,* 289**,** 85-88

Fray, R. G. & Grierson, D. 1993. Identification and Genetic-Analysis of Normal and Mutant Phytoene Synthase Genes of Tomato by Sequencing, Complementation and Co-Suppression. *Plant Molecular Biology,* 22**,** 589-602.Doi 10.1007/Bf00047400

Fréville, H., Montazeaud, G., Forst, E., David, J., Papa, R. & Tenaillon, M. I. 2022. Shift in beneficial interactions during crop evolution. *Evolutionary Applications,* 15**,** 905-918

Fuentes, R. R., De Ridder, D., Van Dijk, A. D. & Peters, S. A. 2022. Domestication shapes recombination patterns in tomato. *Molecular biology and evolution,* 39**,** msab287

Fustier, M.-A., Martínez-Ainsworth, N. E., Aguirre-Liguori, J. A., Venon, A., Corti, H., Rousselet, A., Dumas, F., Dittberner, H., Camarena, M. G. & Grimanelli, D. 2019. Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude. *PLoS genetics,* 15**,** e1008512

Gao, D. & Fox-Fogle, E. 2023. Identification of transcriptionally active transposons in Barley. *BMC Genomic Data,* 24**,** 64

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A. & Sacks, G. L. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics,* 51**,** 1044-1051

Garbowicz, K., Liu, Z., Alseekh, S., Tieman, D., Taylor, M., Kuhalskaya, A., Ofner, I., Zamir, D., Klee, H. J. & Fernie, A. R. 2018. Quantitative trait loci analysis identifies a prominent gene involved in the production of fatty acid-derived flavor volatiles in tomato. *Molecular Plant,* 11**,** 1147-1165

Gaut, Seymour D. K., Liu Q., A. & Y., Z. 2018. Demography and its effects on genomic variation in crop domestication. *Nature Plants,* 4**,** 512 - 520

Gepts, P. & Papa, R. 2001. Evolution during domestication. *e LS,*

Gerszberg, A. & Hnatuszko-Konka, K. 2017. Tomato tolerance to abiotic stress: a review of most often engineered target sequences. *Plant growth regulation,* 83**,** 175-198

Ghalambor, C. K., Hoke, K. L., Ruell, E. W., Fischer, E. K., Reznick, D. N. & Hughes, K. A. 2015. Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature,* 525**,** 372-375

Ghalambor, C. K., Mckay, J. K., Carroll, S. P. & Reznick, D. N. 2007. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional ecology,* 21**,** 394-407

Gilbert, C. & Feschotte, C. 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current opinion in genetics & development,* 49**,** 15-24

Gilbert, S. F., Bosch, T. C. & Ledón-Rettig, C. 2015. Eco-Evo-Devo: developmental symbiosis and developmental plasticity as evolutionary agents. *Nature Reviews Genetics,* 16**,** 611-622

Gill, R. A., Scossa, F., King, G. J., Golicz, A. A., Tong, C., Snowdon, R. J., Fernie, A. R. & Liu, S. 2021. On the role of transposable elements in the regulation of gene expression and subgenomic interactions in crop genomes. *Critical Reviews in Plant Sciences,* 40**,** 157-189.https://doi.org/10.1080/07352689.2021.1920731

Gladieux, P., Van Oosterhout, C., Fairhead, S., Jouet, A., Ortiz, D., Ravel, S., Shrestha, R.-K., Frouin, J., He, X. & Zhu, Y. 2024. Extensive immune receptor repertoire diversity in disease-resistant rice landraces. *Current Biology,* 34**,** 3983-3995.e6.https://doi.org/10.1016/j.cub.2024.07.061

Gómez-Fernández, A. & Milla, R. 2022. How seeds and growth dynamics influence plant size and yield: Integrating trait relationships into ontogeny. *Journal of Ecology,* 110**,** 2684-2700

González, R. M., Ricardi, M. M. & Iusem, N. D. 2013. Epigenetic marks in an adaptive water stress-responsive gene in tomato roots under normal and drought conditions. *Epigenetics,* 8**,** 864-872

Gonzali, S. & Perata, P. 2021. Fruit Colour and Novel Mechanisms of Genetic Regulation of Pigment Production in Tomato Fruits. *Horticulturae,* 7**,** 259

Goubert, C., Modolo, L., Vieira, C., Valientemoro, C., Mavingui, P. & Boulesteix, M. 2015. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome biology and evolution,* 7**,** 1192-1205

Gramazio, P., Pereira-Dias, L., Vilanova, S., Prohens, J., Soler, S., Esteras, J., Garmendia, A. & Díez, M. J. 2020. Morphoagronomic characterization and whole genome resequencing of eight highly diverse wild and weedy S. pimpinellifolium and S. lycopersicum var. cerasiforme accessions used for the first interspecific tomato MAGIC population. *Horticulture Research,* 7**,** 174

Gramazio, P., Yan, H., Hasing, T., Vilanova, S., Prohens, J. & Bombarely, A. 2019. Whole-genome resequencing of seven eggplant (Solanum melongena) and one wild relative (S.

incanum) accessions provides new insights and breeding tools for eggplant enhancement. *Frontiers in plant science,* 10**,** 480262

Grandillo, S. & Tanksley, S. 1996. QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species Lycopersicon pimpinellifolium. *Theoretical and Applied Genetics,* 92**,** 935-951

Gregory, T. R. 2009. Artificial selection and domestication: modern lessons from Darwin's enduring analogy. *Evolution: Education and Outreach,* 2**,** 5-27

Guo, S., Zhao, S., Sun, H., Wang, X., Wu, S., Lin, T., Ren, Y., Gao, L., Deng, Y. & Zhang, J. 2019. Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nature genetics,* 51**,** 1616-1623

Habu, Y., Hisatomi, Y. & Iida, S. 1998. Molecular characterization of the mutable flaked allele for flower variegation in the common morning glory. *The Plant Journal,* 16**,** 371-376

Halligan, D. L. & Keightley, P. D. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics,* 40**,** 151-172

Hanson, L., Boyd, A., Johnson, M. A. & Bennett, M. D. 2005. First nuclear DNA C-values for 18 eudicot families. *Annals of Botany,* 96**,** 1315-1320

Harberd, N. P., Belfield, E. & Yasumura, Y. 2009. The angiosperm gibberellin-GID1-DELLA growth regulatory mechanism: how an "inhibitor of an inhibitor" enables flexible response to fluctuating environments. *The Plant Cell,* 21**,** 1328-1339

Hayashi, K. & Yoshida, H. 2009. Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. *Plant Journal,* 57**,** 413-425.10.1111/j.1365-313X.2008.03694.x

Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S. 2015. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC genomics,* 16**,** 1-16

Hénault, M., Marsit, S., Charron, G. & Landry, C. R. 2020. The effect of hybridization on transposable element accumulation in an undomesticated fungal species. *Elife,* 9**,** e60474

Hirsch, C. D. & Springer, N. M. 2017. Transposable element influences on gene expression in plants. *BBA,* 1860**,** 157-165

Hobson, G. & Grierson, D. 1993. Tomato. In: Seymour G.B., Taylor J.E., Tucker G.A. (eds) Biochemistry of Fruit Ripening. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-1584-1_14.

Hollister, J. D. & Gaut, B. S. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research,* 19**,** 1419-1428

Hori, Y., Fujimoto, R., Sato, Y. & Nishio, T. 2007. A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (Oryza sativa). *Theoretical and Applied Genetics,* 115**,** 217-224.10.1007/s00122-007-0557-6

Hosmani, P. S., Flores-Gonzalez, M., Geest, H. V. D., Maumus, F., Bakker, L. V., Schijlen, E., Haarst, J. V., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J.,

Mueller, L. A. & Saha, S. 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv,* 10.1101/767764.10.1101/767764

Hou, J., Long Yan, Raman H, Zou Xiaoxiao, Wang Jing, Dai Shutao, Xiao Qinqin, Li Cong, Fan Longjiang & Liu Bin 2012. A Tourist-like MITE insertion in the upstream region of the BnFLC. A10 gene is associated with vernalization requirement in rapeseed (Brassica napus L.).

Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X., Xu, G., Tian, J., Wang, C. & Li, D. 2018. ZmCCT9 enhances maize adaptation to higher latitudes. *Proceedings of the National Academy of Sciences,* 115**,** E334-E341.https://doi.org/10.1073/pnas.1718058115

Huang, C. R. L., Burns, K. H. & Boeke, J. D. 2012. Active transposition in genomes. *Annual review of genetics,* 46**,** 651-675

Huang, Z., Van Houten, J., Gonzalez, G., Xiao, H. & Van Der Knaap, E. 2013. Genome-wide identification, phylogeny and expression analysis of SUN, OFP and YABBY gene family in tomato. *Molecular genetics and genomics,* 288**,** 111-129

Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E. & Kaeppler, S. M. 2012. Comparative population genomics of maize domestication and improvement. *Nature genetics,* 44**,** 808-811

Huq, M. A., Akter, S., Nou, I. S., Kim, H. T., Jung, Y. J. & Kang, K. K. 2016. Identification of functional SNPs in genes and their effects on plant phenotypes. *Journal of Plant Biotechnology,* 43**,** 1-11

Ichikawa, M., Kato, N., Toda, E., Kashihara, M., Ishida, Y., Hiei, Y., Isobe, S. N., Shirasawa, K., Hirakawa, H. & Okamoto, T. 2023. Whole-genome sequence analysis of mutations in rice plants regenerated from zygotes, mature embryos, and immature embryos. *Breeding Science,* 73**,** 349-353

Iijima, Y., Watanabe, B., Sasaki, R., Takenaka, M., Ono, H., Sakurai, N., Umemoto, N., Suzuki, H., Shibata, D. & Aoki, K. 2013. Steroidal glycoalkaloid profiling and structures of glycoalkaloids in wild tomato fruit. *Phytochemistry,* 95**,** 145-157

Ikeda, H., Hiraga, M., Shirasawa, K., Nishiyama, M., Kanahama, K. & Kanayama, Y. 2013. Analysis of a tomato introgression line, IL8-3, with increased Brix content. *Scientia Horticulturae,* 153**,** 103-108

International Rice Genome Sequencing Project 2005. The map-based sequence of the rice genome. *Nature,* 436**,** 793-800

Iqbal, N., Khan, N. A., Ferrante, A., Trivellini, A., Francini, A. & Khan, M. 2017. Ethylene role in plant growth, development and senescence: interaction with other phytohormones. *Frontiers in plant science,* 8**,** 475

Ishikawa, R., Castillo, C. C., Htun, T. M., Numaguchi, K., Inoue, K., Oka, Y., Ogasawara, M., Sugiyama, S., Takama, N. & Orn, C. 2022. A stepwise route to domesticate rice by controlling seed shattering and panicle shape. *Proceedings of the National Academy of Sciences,* 119**,** e2121692119

Ito, H., Yoshida, T., Tsukahara, S. & Kawabe, A. 2013. Evolution of the ONSEN retrotransposon family activated upon heat stress in Brassicaceae. *Gene,* 518**,** 256-261

Jiang, C., Mithani, A., Belfield, E. J., Mott, R., Hurst, L. D. & Harberd, N. P. 2014. Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome research,* 24**,** 1821-1829

Jiang, J., Xu, Y.-C., Zhang, Z.-Q., Chen, J.-F., Niu, X.-M., Hou, X.-H., Li, X.-T., Wang, L., Zhang, Y. E. & Ge, S. 2024. Forces driving transposable element load variation during Arabidopsis range expansion. *The Plant Cell,* 36**,** 840-862

Jiang, N., Visa, S., Wu, S. & Kaap, E. V. D. 2012. Rider Transposon Insertion and Phenotypic Change in Tomato. *Springer Berlin Heidelberg, Berlin, Heidelberg.,*

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B. & Al., E. 2017. Improved maize reference genome with single molecule technologies. *Nature,* 546**,** 524-527

Johannsen, W. 1911. The genotype conception of heredity. *The American Naturalist,* 45**,** 129-159

Jones, G., Kluyver, T., Preece, C., Swarbrick, J., Forster, E., Wallace, M., Charles, M., Rees, M. & Osborne, C. P. 2021. The origins of agriculture: Intentions and consequences. *Journal of Archaeological Science,* 125.10.1016/j.jas.2020.105290

Kapazoglou, A., Gerakari, M., Lazaridi, E., Kleftogianni, K., Sarri, E., Tani, E. & Bebeli, P. J. 2023. Crop wild relatives: A valuable source of tolerance to various abiotic stresses. *Plants,* 12**,** 328

Karniel, U., Adler Berke, N., Mann, V. & Hirschberg, J. 2022. Perturbations in the carotenoid biosynthesis pathway in tomato fruit reactivate the leaf-specific phytoene synthase 2. *Frontiers in Plant Science,* 13**,** 844748

Katju, V. & Bergthorsson, U. 2019. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biology and Evolution,* 11**,** 136-165

Katsir, L., Chung, H. S., Koo, A. J. & Howe, G. A. 2008. Jasmonate signaling: a conserved mechanism of hormone sensing. *Current opinion in plant biology,* 11**,** 428-435

Kaur, G., Abugu, M. & Tieman, D. 2023. The dissection of tomato flavor: biochemistry, genetics, and omics. *Frontiers in Plant Science,* 14**,** 1144113

Kawase, M., Fukunaga, K. & Kato, K. 2005. Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions. *Molecular Genetics and Genomics,* 274**,** 131-140.10.1007/s00438-005-0013-8

Keightley, P. D. & Eyre-Walker, A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 1187-1193

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M. L. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome research,* 19**,** 1195-1201

Keith, N., Tucker, A. E., Jackson, C. E., Sung, W., Lledó, J. I. L., Schrider, D. R., Schaack, S., Dudycha, J. L., Ackerman, M. & Younge, A. J. 2016. High mutational rates of large-scale duplication and deletion in Daphnia pulex. *Genome research,* 26**,** 60-69

Kelleher, E. S., Barbash, D. A. & Blumenstiel, J. P. 2020. Taming the turmoil within: new insights on the containment of transposable elements. *Trends in Genetics,* 36**,** 474-489

Kenkel, C. D. & Matz, M. V. 2016. Gene expression plasticity as a mechanism of coral adaptation to a variable environment. *Nature Ecology & Evolution,* 1**,** 0014

Kikuchi, S., Asakura, Y., Imai, M., Nakahira, Y., Kotani, Y., Hashiguchi, Y., Nakai, Y., Takafuji, K., Bédard, J. & Hirabayashi-Ishioka, Y. 2018. A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. *The Plant Cell,* 30**,** 2677-2703

Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., Takase, M., Ide, T. & Nakai, M. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science,* 339**,** 571-574

Kim, J.-E., Oh, S.-K., Lee, J.-H., Lee, B.-M. & Jo, S.-H. 2014. Genome-wide SNP calling using next generation sequencing data in tomato. *Molecules and cells,* 37**,** 36-42

Kimura, M. 1985. *The neutral theory of molecular evolution*, Cambridge University Press

Kimura, Y., Tosa, Y., Shimada, S., Sogo, R., Kusaba, M., Sunaga, T., Betsuyaku, S., Eto, Y., Nakayashiki, H. & Mayama, S. 2001. OARE-1, a Ty1-copia retrotransposon in oat activated by abiotic and biotic stresses. *Plant and cell Physiology,* 42**,** 1345-1354

Kistler, L., Maezumi, S. Y., Gregorio De Souza, J., Przelomska, N. A., Malaquias Costa, F., Smith, O., Loiselle, H., Ramos-Madrigal, J., Wales, N. & Ribeiro, E. R. 2018. Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science,* 362**,** 1309-1313

Knapp, S. 2002. Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *J Exp Bot,* 53**,** 2001-22.10.1093/jxb/erf068

Koenig, D., Jiménez-Gómez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., Kumar, R., Covington, M. F., Devisetty, U. K., Tat, A. V., Tohge, T., Bolger, A., Schneeberger, K., Ossowski, S., Christa Lanz, G. X., Taylor-Teeples, M., Brady, S. M., Pauly, M., Weigel, D., Usadel, B., Fernie, A. R., Peng, J., Sinha, N. R. & Maloof, J. N. 2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *PNAS,* 110**,** E2655-E2662

Kofler, R., Gómez-Sánchez, D. & Schlötterer, C. 2016. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol Biol Evol.,* 33**,** 2759–2764

Kortbeek, R. W., Galland, M. D., Muras, A., Van Der Kloet, F. M., André, B., Heilijgers, M., Van Hijum, S. A., Haring, M. A., Schuurink, R. C. & Bleeker, P. M. 2021. Natural variation in wild tomato trichomes; selecting metabolites that contribute to insect resistance using a random forest approach. *BMC plant biology,* 21**,** 1-19

Kumar, D. 2014. Salicylic acid signaling in disease resistance. *Plant Science,* 228**,** 127-134

Laland, K., Uller, T., Feldman, M., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, E., Odling-Smee, J., Wray, G. A. & Hoekstra, H. E. 2014. Does evolutionary theory need a rethink? *Nature,* 514**,** 161-164

Lanciano, S. & Cristofari, G. 2020. Measuring and interpreting transposable element expression. *Nature Reviews Genetics,* 21**,** 721-736.https://doi.org/10.1038/s41576-020-0251-y

Lanfear, R. 2018. Do plants have a segregated germline? *PLoS biology,* 16**,** e2005439

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9**,** 357-359

Larson, G., Piperno, D. R., Allaby, R. G., Purugganan, M. D., Andersson, L., Arroyo-Kalin, M., Barton, L., Climer Vigueira, C., Denham, T. & Dobney, K. 2014. Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences,* 111**,** 6139-6146

Lee, H., Popodi, E., Tang, H. & Foster, P. L. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences,* 109**,** E2774-E2783

Lee, Y. C. G. 2022. Synergistic epistasis of the deleterious effects of transposable elements. *Genetics,* 220**,** iyab211

Lei, G. J., Fujii-Kashino, M., Wu, D. Z., Hisano, H., Saisho, D., Deng, F. L., Yamaji, N., Sato, K., Zhao, F. J. & Ma, J. F. 2020. Breeding for low cadmium barley by introgression of a Sukkula-like transposable element. *Nature Food,* 1**,** 489-U17.10.1038/s43016-020-0130-x

Lenser, T. & Theißen, G. 2013. Molecular mechanisms involved in convergent crop domestication. *Trends in plant science,* 18**,** 704-714

Levis, N. A. & Pfennig, D. W. 2016. Evaluating 'plasticity-first'evolution in nature: key criteria and empirical approaches. *Trends in ecology & evolution,* 31**,** 563-574

Lewinsohn, E., Schalechet, F., Wilkinson, J., Matsui, K., Tadmor, Y., Nam, K.-H., Amar, O., Lastochkin, E., Larkov, O. & Ravid, U. 2001. Enhanced levels of the aroma and flavor compound S-linalool by metabolic engineering of the terpenoid pathway in tomato fruits. *Plant Physiology,* 127**,** 1256-1265

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754–1760

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9

Li, H., Qi, M., Sun, M., Liu, Y., Liu, Y., Xu, T., Li, Y. & Li, T. 2017a. Tomato transcription factor SlWUS plays an important role in tomato flower and locule development. *Frontiers in plant science,* 8**,** 457

Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., Yang, T., Tang, Y., Yang, S. & Aisimutuola, P. 2023. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nature Genetics,* 55**,** 852-860

Li, X., Guo, K., Zhu, X., Chen, P., Li, Y., Xie, G., Wang, L., Wang, Y., Persson, S. & Peng, L. 2017b. Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. *BMC Genomics,* 18**,** 55.10.1186/s12864-016-3454-z

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S. & Wang, X. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature genetics,* 46**,** 1220-1226

Lin, Z., Li, X., Shannon, L. M., Yeh, C.-T., Wang, M. L., Bai, G., Peng, Z., Li, J., Trick, H. N. & Clemente, T. E. 2012. Parallel domestication of the Shattering1 genes in cereals. *Nature genetics,* 44**,** 720-724

Lisch, D. 2013. How important are transposons for plant evolution? *Nat Rev Genet,* 14**,** 49-61.10.1038/nrg3374

Liu, D., Yang, L., Zhang, J.-Z., Zhu, G.-T., Lü, H.-J., Lü, Y.-Q., Wang, Y.-L., Cao, X., Sun, T.-S. & Huang, S.-W. 2020a. Domestication and breeding changed tomato fruit transcriptome. *Journal of Integrative Agriculture,* 19**,** 120-132

Liu, H., Fang, X., Zhou, L., Li, Y., Zhu, C., Liu, J., Song, Y., Jian, X., Xu, M. & Dong, L. 2022a. Transposon insertion drove the loss of natural seed shattering during foxtail millet domestication. *Molecular Biology and Evolution,* 39**,** msac078

Liu, J., Zhou, R. J., Wang, W. X., Wang, H., Qiu, Y., Raman, R., Mei, D. S., Raman, H. & Hu, Q. 2020b. A copia-like retrotransposon insertion in the upstream region of the SHATTERPROOF1 gene, BnSHP1.A9, is associated with quantitative variation in pod shattering resistance in oilseed rape. *Journal of Experimental Botany,* 71**,** 5402-5413.10.1093/jxb/eraa281

Liu, W., Liu, K., Chen, D., Zhang, Z., Li, B., El-Mogy, M. M., Tian, S. & Chen, T. 2022b. Solanum lycopersicum, a model plant for the studies in developmental biology, stress biology and food science. *Foods,* 11**,** 2402

Liu, Z., Fan, M., Yue, E.-K., Li, Y., Tao, R.-F., Xu, H.-M., Duan, M.-H. & Xu, J.-H. 2020c. Natural variation and evolutionary dynamics of transposable elements in Brassica oleracea based on next-generation sequencing data. *Horticulture research,* 7

Lockton, S. & Gaut, B. S. 2010. The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. *BMC Evolutionary Biology,* 10**,** 1-11

Loewe, L. & Hill, W. G. 2010. The population genetics of mutations: good, bad and indifferent. The Royal Society

Lorant, A., Pedersen, S., Holst, I., Hufford, M. B., Winter, K., Piperno, D. & Ross-Ibarra, J. 2017. The potential role of genetic assimilation during maize domestication. *PloS one,* 12**,** e0184202

Lorenzetti, A. P. R., Antonio, G. Y. a. D., Paschoal, A. R. & Domingues, D. S. 2016. PlanTE-MIR DB: a database for transposable element-related microRNAs in plant genomes. *Functional & integrative genomics,* 16**,** 235-242

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W. & Kuang, H. 2012. Miniature inverted–repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in Oryza sativa. *Molecular biology and evolution,* 29**,** 1005-1017

Lu, Z., Cui, J., Wang, L., Teng, N., Zhang, S., Lam, H.-M., Zhu, Y., Xiao, S., Ke, W. & Lin, J. 2021. Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. *Genome biology,* 22**,** 160

Luo, A., Kang, S. & Chen, J. 2020. SUGAR model-assisted analysis of carbon allocation and transformation in tomato fruit under different water along with potassium conditions. *Frontiers in Plant Science,* 11**,** 712

Lupo, Y. & Moshelion, M. 2024. The balance of survival: comparative drought response in wild and domesticated tomatoes. *Plant Science,* 339**,** 111928

Lye, Z., Choi, J. Y. & Purugganan, M. D. 2022. Deleterious mutations and the rare allele burden on rice gene expression. *Molecular Biology and Evolution,* 39**,** msac193

Lynch, M. & Hill, W. G. 1986. Phenotypic evolution by neutral mutation. *Evolution,* 40**,** 915-935

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S. & Hartl, D. L. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences,* 105**,** 9272-9277

Macko-Podgórni, A., Stelmach, K., Kwolek, K. & Grzebelus, D. 2019. Stowaway miniature inverted repeat transposable elements are important agents driving recent genomic diversity in wild and cultivated carrot. *Mobile DNA,* 10**,** 1-17

Malinovsky, F. G., Brodersen, P., Fiil, B. K., Mckinney, L. V., Thorgrimsen, S., Beck, M., Nielsen, H. B., Pietra, S., Zipfel, C. & Robatzek, S. 2010. Lazarus1, a DUF300 protein, contributes to programmed cell death associated with Arabidopsis acd11 and the hypersensitive response. *PLoS One,* 5**,** e12586

Mammadov, J., Buyyarapu, R., Guttikonda, S. K., Parliament, K., Abdurakhmonov, I. Y. & Kumpatla, S. P. 2018. Wild relatives of maize, rice, cotton, and soybean: treasure troves for tolerance to biotic and abiotic stresses. *Frontiers in plant science,* 9**,** 886

Mao, H. D., Wang, H. W., Liu, S. X., Li, Z., Yang, X. H., Yan, J. B., Li, J. S., Tran, L. S. P. & Qin, F. 2015. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nature Communications,* 6.ARTN 8326

10.1038/ncomms9326

Marques, E., Krieg, C. P., Dacosta-Calheiros, E., Bueno, E., Sessa, E., Penmetsa, R. V. & Von Wettberg, E. 2020. The impact of domestication on aboveground and belowground trait responses to nitrogen fertilization in wild and cultivated genotypes of Chickpea (Cicer sp.). *Frontiers in Genetics,* 11**,** 576338

Martín-Robles, N., Morente-López, J., Freschet, G. T., Poorter, H., Roumet, C., Milla, R. & Tjoelker, M. 2018. Root traits of herbaceous crops: Pre-adaptation to cultivation or evolution under domestication? *Functional Ecology,* 33**,** 273-285.10.1111/1365-2435.13231

Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., Korol, A., David, M., Reiter, E. & Riehl, S. 2016. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nature Genetics,* 48**,** 1089-1093

Matesanz, S. & Milla, R. 2018. Differential plasticity to water and nutrients between crops and their wild progenitors. *Environmental and Experimental Botany,* 145**,** 54-63.10.1016/j.envexpbot.2017.10.014

Mayr, E. 1963. *Animal species and evolution*, Harvard University Press

Mcclintock, B. 1953. Induction of instability at selected loci in maize. *Genetics,* 38**,** 579

Mccullers, T. J. & Steiniger, M. 2017. Transposable elements in Drosophila. *Mobile genetic elements,* 7**,** 1-18

Mcnally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., Zeller, G., Clark, R. M., Hoen, D. R. & Bureau, T. E. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences,* 106**,** 12273-12278

Mehra, M., Gangwar, I. & Shankar, R. 2015. A Deluge of Complex Repeats: The Solanum Genome. *PLoS One,* 10**,** e0133962

Ménard, L., Mckey, D., Mühlen, G. S., Clair, B. & Rowe, N. P. 2013. The evolutionary fate of phenotypic plasticity and functional traits under domestication in manioc: Changes in stem biomechanics and the appearance of stem brittleness. *PloS one,* 8**,** e74727

Messer, P. W. & Petrov, D. A. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution,* 28**,** 659-669

Meyer, R. S. & Purugganan, M. D. 2013. Evolution of crop species: genetics of domestication and diversification. *Nature reviews genetics,* 14**,** 840-852

Milla, R. 2023. Phenotypic evolution of agricultural crops. *Functional Ecology,* 37**,** 976-988

Milla, R., Bastida, J. M., Turcotte, M. M., Jones, G., Violle, C., Osborne, C. P., Chacón-Labella, J., Sosinski Jr, Ê. E., Kattge, J. & Laughlin, D. C. 2018. Phylogenetic patterns and phenotypic profiles of the species of plants and mammals farmed for food. *Nature Ecology & Evolution,* 2**,** 1808-1817

Milla, R., Morente-López, J., Alonso-Rodrigo, J. M., Martín-Robles, N. & Stuart Chapin Iii, F. 2014. Shifts and disruptions in resource-use trait syndromes during the evolution of herbaceous crops. *Proceedings of the Royal Society B: Biological Sciences,* 281**,** 20141429

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M. & Kliebenstein, D. 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature,* 602**,** 101-105

Moura, D. S., Bergey, D. R. & Ryan, C. A. 2001. Characterization and localization of a wound-inducible type I serine-carboxypeptidase from leaves of tomato plants (Lycopersicon esculentum Mill.). *Planta,* 212**,** 222-230

Moyers, B. T., Morrell, P. L. & Mckay, J. K. 2018. Genetic costs of domestication and improvement. *Journal of Heredity,* 109**,** 103-116

Mueller, N. G. 2017. Evolutionary “Bet-Hedgers” under cultivation: investigating the domestication of erect knotweed (Polygonum erectum L.) using growth experiments. *Human Ecology,* 45**,** 189-203.https://www.jstor.org/stable/44202545

Mueller, N. G., Horton, E. T., Belcher, M. E. & Kistler, L. 2023. The taming of the weed: Developmental plasticity facilitated plant domestication. *Plos one,* 18**,** e0284136

Muñoz-Bertomeu, J., Miedes, E. & Lorences, E. 2013. Expression of xyloglucan endotransglucosylase/hydrolase (XTH) genes and XET activity in ethylene treated apple and tomato fruits. *Journal of plant physiology,* 170**,** 1194-1201

Murray, C., Sutherland, P. W., Phung, M. M., Lester, M. T., Marshall, R. K. & Christeller, J. T. 2002. Expression of biotin-binding proteins, avidin and streptavidin, in plant tissues using plant vacuolar targeting sequences. *Transgenic Research,* 11**,** 199-214

Murren, C. J., Auld, J. R., Callahan, H., Ghalambor, C. K., Handelsman, C. A., Heskel, M. A., Kingsolver, J., Maclean, H. J., Masel, J. & Maughan, H. 2015. Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity. *Heredity,* 115**,** 293-301

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., Okumoto, Y., Tanisaka, T. & Wessler, S. R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature,* 461**,** 1130–1134

Nakazaki, T., Okumoto, Y., Horibata, A., Yamahira, S., Teraishi, M., Nishida, H., Inoue, H. & Tanisaka, T. 2003. Mobilization of a transposon in the rice genome. *Nature,* 421**,** 170-172

Naver, H., Boudreau, E. & Rochaix, J.-D. 2001. Functional studies of Ycf3: its role in assembly of photosystem I and interactions with some of its subunits. *The Plant Cell,* 13**,** 2731-2745

Ndjiondjop, M. N., Alachiotis, N., Pavlidis, P., Goungoulou, A., Kpeki, S. B., Zhao, D. & Semagn, K. 2019. Comparisons of molecular diversity indices, selective sweeps and population structure of African rice with its wild progenitor and Asian rice. *Theoretical and Applied Genetics,* 132**,** 1145-1158

Nei, M. & Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences,* 76**,** 5269-5273

Nelson, M., Linheiro, R. & Bergman, C. 2017. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertion in Whole-Genome Shotgun Sequencing Data. *G3,* 7**,** 2763-2778

Nesbitt, T. C. & Tanksley, S. D. 2002. Comparative sequencing in the genus Lycopersicon: implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics,* 162**,** 365-379

Nitasaka, E. 2003. Insertion of an En/Spm-related transposable element into a floral homeotic gene DUPLICATED causes a double flower phenotype in the Japanese morning glory. *The Plant Journal,* 36**,** 522-531

Niu, X.-M., Xu, Y.-C., Li, Z.-W., Bian, Y.-T., Hou, X.-H., Chen, J.-F., Zou, Y.-P., Jiang, J., Wu, Q. & Ge, S. 2019. Transposable elements drive rapid phenotypic variation in Capsella rubella. *Proceedings of the National Academy of Sciences,* 116**,** 6908-6913

Nuez, F., Prohens, J. & Blanca, J. M. 2004. Relationships, origin, and diversity of Galapagos tomatoes: implications for the conservation of natural populations. *American Journal of Botany,* 91**,** 86-99

Oliver, K. R. & Greene, W. K. 2012. Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. *Ecology and evolution,* 2**,** 2912-2933

Oliver, K. R., Mccomb, J. A. & Greene, W. K. 2013. Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biol Evol,* 5**,** 1886–1901

Olson-Manning, C. F., Wagner, M. R. & Mitchell-Olds, T. 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews Genetics,* 13**,** 867-877

Osmanski, A. B., Paulat, N. S., Korstian, J., Grimshaw, J. R., Halsey, M., Sullivan, K. A., Moreno-Santillán, D. D., Crookshanks, C., Roberts, J. & Garcia, C. 2023. Insights into

mammalian TE diversity through the curation of 248 genome assemblies. *Science,* 380**,** eabn1430

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D. & Lynch, M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *science,* 327**,** 92-94

Ou, S. & Jiang, N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology,* 176**,** 1410-1422

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N. & Hufford, M. B. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology,* 20**,** 275

Pailles, Y., Ho, S., Pires, I. S., Tester, M., Negrão, S. & Schmöckel, S. M. 2017. Genetic diversity and population structure of two tomato species from the Galapagos Islands. *Frontiers in Plant Science,* 8**,** 138

Palacio-López, K., Beckage, B., Scheiner, S. & Molofsky, J. 2015. The ubiquity of phenotypic plasticity in plants: a synthesis. *Ecology and evolution,* 5**,** 3389-3400

Park, C., Qian, W. & Zhang, J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO reports,* 13**,** 1123-1129

Park, J.-S., Park, J.-H. & Park, Y.-D. 2019. Construction of pseudomolecule sequences of Brassica rapa ssp. pekinensis inbred line CT001 and analysis of spontaneous mutations derived via sexual propagation. *PLoS One,* 14**,** e0222283

Paudel, S., Lin, P.-A., Foolad, M. R., Ali, J. G., Rajotte, E. G. & Felton, G. W. 2019. Induced plant defenses against herbivory in cultivated and wild tomato. *Journal of Chemical Ecology,* 45**,** 693-707

Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biol,* 14**,** e1002379.10.1371/journal.pbio.1002379

Pedro, D. L. F., Lorenzetti, A. P. R., Domingues, D. S. & Paschoal, A. R. 2018. PlaNC-TE: a comprehensive knowledgebase of non-coding RNAs and transposable elements in plants. *Database,* 2018**,** bay078

Peischl, S. & Kirkpatrick, M. 2012. Establishment of new mutations in changing environments. *Genetics,* 191**,** 895-906

Peralta, I. E., Knapp, S. And Spooner, D.M. 2006. Nomenclature for wild and cultivated tomatoes. *Tomato genetics cooperative report,* 56**,** 6-12

Pereira, L., Sapkota, M., Alonge, M., Zheng, Y., Zhang, Y., Razifard, H., Taitano, N. K., Schatz, M. C., Fernie, A. R. & Wang, Y. 2021. Natural genetic diversity in tomato flavor genes. *Frontiers in plant science,* 12**,** 642828

Pfennig, D. W. 2021. *Phenotypic plasticity & evolution: causes, consequences, controversies*, Taylor & Francis

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D. & Moczek, A. P. 2010. Phenotypic plasticity's impacts on diversification and speciation. *Trends in ecology & evolution,* 25**,** 459-467

Pigliucci, M. & Murren, C. J. 2003. Perspective: genetic assimilation and a possible evolutionary paradox: can macroevolution sometimes be so fast as to pass us by? *Evolution,* 57**,** 1455-1464

Pigliucci, M., Tyler, G. A. & Schlichting, C. D. 1998. Mutational effects on constraints on character evolution and phenotypic plasticity in Arabidopsis thaliana. *Journal of Genetics,* 77**,** 95-103

Pimpinelli, S., Piacentini, L. & Herrel, A. 2019. Environmental change and the evolution of genomes: Transposable elements as translators of phenotypic plasticity into genotypic variability. *Functional Ecology,* 34**,** 428-441.10.1111/1365-2435.13497

Piperno, D. R. 2011. The origins of plant cultivation and domestication in the New World tropics: patterns, process, and new developments. *Current anthropology,* 52**,** S453-S470

Piperno, D. R. 2017. Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. *PNAS,* 114**,** 6429-6437

Piperno, D. R., Holst, I., Moreno, J. E. & Winter, K. 2019. Experimenting with domestication: Understanding macro- and micro-phenotypes and developmental plasticity in teosinte in its ancestral pleistocene and early holocene environments. *Journal of Archaeological Science,* 108.10.1016/j.jas.2019.05.006

Piperno, D. R., Holst, I., Winter, K. & Mcmillan, O. 2015. Teosinte before domestication: Experimental study of growth and phenotypic variability in Late Pleistocene and early Holocene environments. *Quaternary International,* 363**,** 65-77

Pnueli, L., Carmel-Goren, L., Hareven, D., Gutfinger, T., Alvarez, J., Ganal, M., Zamir, D. & Lifschitz, E. 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

Preece, C., Clamp, N. F., Warham, G., Charles, M., Rees, M., Jones, G. & Osborne, C. P. 2018. Cereal progenitors differ in stand harvest characteristics from related wild grasses. *J Ecol,* 106**,** 1286-1297.10.1111/1365-2745.12905

Preece, C., Jones, G., Rees, M. & Osborne, C. P. 2021. Fertile Crescent crop progenitors gained a competitive advantage from large seedlings. *Ecology and Evolution,* 11**,** 3300-3312

Preece, C., Livarda, A., Christin, P. A., Wallace, M., Martin, G., Charles, M., Jones, G., Rees, M. & Osborne, C. P. 2017. How did the domestication of Fertile Crescent grain crops increase their yields? *Functional ecology,* 31**,** 387-397

Preece, C., Livarda, A., Wallace, M., Martin, G., Charles, M., Christin, P. A., Jones, G., Rees, M. & Osborne, C. P. 2015. Were Fertile Crescent crop progenitors higher yielding than other wild species that were never domesticated? *New Phytol,* 207**,** 905-13.10.1111/nph.13353

Prjna713664, B. I.:

Purugganan, M. D. 2019. Evolutionary Insights into the Nature of Plant Domestication. *Curr Biol,* 29**,** R705-R714.10.1016/j.cub.2019.05.053

Purugganan, M. D. 2022. What is domestication? *Trends in Ecology & Evolution,* 37**,** 663-671

Purugganan, M. D. & Fuller, D. Q. 2009. The nature of selection during plant domestication. *Nature,* 457**,** 843-848

Purugganan, M. D. & Fuller, D. Q. 2011. Archaeological data reveal slow rates of evolution during plant domestication. *Evolution,* 65**,** 171-183

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini Silveira, A., Engelen, S., Baillet, V. & Wincker, P. 2019. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nature Communications,* 10**,** 3421

Ramakrishnan, M., Satish, L., Kalendar, R., Narayanan, M., Kandasamy, S., Sharma, A., Emamverdian, A., Wei, Q. & Zhou, M. 2021. The dynamism of transposon methylation for plant development and stress adaptation. *International Journal of Molecular Sciences,* 22**,** 11387

Ramakrishnan, M., Satish, L., Sharma, A., Kurungara Vinod, K., Emamverdian, A., Zhou, M. & Wei, Q. 2022. Transposable elements in plants: Recent advancements, tools and prospects. *Plant Molecular Biology Reporter,* 40**,** 628-645

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 biology,* 8**,** 1-16

Raquin, A. L., Brabant, P., Rhoné, B., Balfourier, F., Leroy, P. & Goldringer, I. 2008. Soft selective sweep near a gene that increases plant height in wheat. *Molecular Ecology,* 17**,** 741-756

Razifard, H., Ramos, A., Della Valle, A. L., Bodary, C., Goetz, E., Manser, E. J., Li, X., Zhang, L., Visa, S., Tieman, D., Van Der Knaap, E. & Caicedo, A. L. 2020. Genomic Evidence for Complex Domestication History of the Cultivated Tomato in Latin America. *Mol Biol Evol,* 37**,** 1118-1132.10.1093/molbev/msz297

Reimer, J. J., Thiele, B., Biermann, R. T., Junker-Frohn, L. V., Wiese-Klinkenberg, A., Usadel, B. & Wormit, A. 2021. Tomato leaves under stress: A comparison of stress response to mild abiotic stress between a cultivated and a wild tomato species. *Plant molecular biology,* 107**,** 177-206

Remigereau, M.-S., Lakis, G., Rekima, S., Leveugle, M., Fontaine, M. C., Langin, T., Sarr, A. & Robert, T. 2011. Cereal domestication and evolution of branching: evidence for soft selection in the Tb1 orthologue of pearl millet (Pennisetum glaucum [L.] R. Br.). *PloS one,* 6**,** e22404

Renaut, S. & Rieseberg, L. H. 2015. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular biology and evolution,* 32**,** 2273-2283

Rindos, D. 2013. *The origins of agriculture: an evolutionary perspective*, Academic Press

Rivera, H. E., Aichelman, H. E., Fifer, J. E., Kriefall, N. G., Wuitchik, D. M., Smith, S. J. & Davies, S. W. 2021. A framework for understanding gene expression plasticity and its influence on stress tolerance. *Molecular Ecology,* 30**,** 1381-1397

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29**,** 24-26

Romero, A. J., Kolesnikova, A., Ezard, T. H., Charles, M., Gutaker, R. M., Osborne, C. P. & Chapman, M. A. 2025. 'Domesticability': were some species predisposed for domestication? *Trends in Ecology & Evolution,*

Roquis, D., Robertson, M., Yu, L., Thieme, M., Julkowska, M. & Bucher, E. 2021. Genomic impact of stress-induced transposable element mobility in Arabidopsis. *Nucleic acids research,* 49**,** 10431-10447

Roselius, K., Stephan, W. & StäDler, T. 2005. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics,* 171**,** 753-763

Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences,* 104**,** 8641-8648

Rowley-Conwy, P. & Layton, R. 2011. Foraging and farming as niche construction: stable and unstable adaptations. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 366**,** 849-862

Sadras, V. O., Lake, L., Li, Y., Farquharson, E. A. & Sutton, T. 2016. Phenotypic plasticity and its genetic regulation for yield, nitrogen fixation and δ13C in chickpea crops under varying water regimes. *Journal of Experimental Botany,* 67**,** 4339-4351

Sage, R. F. 1995. Was low atmospheric CO2 during the Pleistocene a limiting factor for the origin of agriculture? *Global Change Biology,* 1**,** 93-106

Sah, S. K., Reddy, K. R. & Li, J. 2016. Abscisic acid and abiotic stress tolerance in crop plants. *Frontiers in plant science,* 7**,** 571

Sahu, P. P., Sharma, N., Puranik, S. & Prasad, M. 2014. Post-transcriptional and epigenetic arms of RNA silencing: a defense machinery of naturally tolerant tomato plant against Tomato leaf curl New Delhi virus. *Plant molecular biology reporter,* 32**,** 1015-1029

Salazar-Mendoza, P., Magalhães, D. M., Lourenção, A. L. & Bento, J. M. S. 2023. Differential defensive and nutritional traits among cultivated tomato and its wild relatives shape their interactions with a specialist herbivore. *Planta,* 257**,** 76

Salazar, M., González, E., Casaretto, J. A., Casacuberta, J. M. & Ruiz-Lara, S. 2007. The promoter of the TLC1. 1 retrotransposon from Solanum chilense is activated by multiple stress-related signaling molecules. *Plant Cell Reports,* 26**,** 1861-1868

Salman-Minkov, A., Sabath, N. & Mayrose, I. 2016. Whole-genome duplication as a key factor in crop domestication. *Nature plants,* 2**,** 1-4

Sampath, P., Murukarthick, J., Izzah, N. K., Lee, J., Choi, H.-I., Shirasawa, K., Choi, B.-S., Liu, S., Nou, I.-S. & Yang, T.-J. 2014. Genome-Wide Comparative Analysis of 20 Miniature Inverted-Repeat Transposable Element Families in Brassica rapa and B. oleracea. *PLoS ONE,* 9**,** e94499.10.1371/journal.pone.0094499

Sandler, G., Bartkowska, M., Agrawal, A. F. & Wright, S. I. 2020. Estimation of the SNP mutation rate in two vegetatively propagating species of duckweed. *G3: Genes, Genomes, Genetics,* 10**,** 4191-4200

Sato, S., Peet, M. M. & Thomas, J. F. 2002. Determining critical pre-and post-anthesis periods and physiological processes in Lycopersicon esculentum Mill. exposed to moderately elevated temperatures. *Journal of Experimental Botany,* 53**,** 1187-1195

Sauvage, C., Rau, A., Aichholz, C., Chadoeuf, J., Sarah, G., Ruiz, M., Santoni, S., Causse, M., David, J. & Glémin, S. 2017. Domestication rewired gene expression and nucleotide diversity patterns in tomato. *The Plant Journal,* 91**,** 631-645

Schmalhausen, I. I. 1949. Factors of evolution: the theory of stabilizing selection.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S. & Al., E. 2009. The B73 maize genome: complexity, diversity and dynamics. *Science,* 326**,** 1112-1115

Schneider, H. M. 2022. Characterization, costs, cues and future perspectives of phenotypic plasticity. *Annals of botany,* 130**,** 131-148

Schneider, H. M. & Lynch, J. P. 2020. Should root plasticity be a crop breeding target? *Frontiers in Plant Science,* 11**,** 546.https://doi.org/10.3389/fpls.2020.00546

Schrader, L. & Schmitz, J. 2019. The impact of transposable elements in adaptive evolution. *Molecular Ecology,* 28**,** 1537-1549

Schwahn, K., De Souza, L. P., Fernie, A. R. & Tohge, T. 2014. Metabolomics-assisted refinement of the pathways of steroidal glycoalkaloid biosynthesis in the tomato clade. *Journal of integrative plant biology,* 56**,** 864-875

Schwander, T. & Leimar, O. 2011. Genes as leaders and followers in evolution. *Trends in Ecology & Evolution,* 26**,** 143-151

Scott, M. F., Botigué, L. R., Brace, S., Stevens, C. J., Mullin, V. E., Stevenson, A., Thomas, M. G., Fuller, D. Q. & Mott, R. 2019. A 3,000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. *Nature plants,* 5**,** 1120-1128

Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., Pleban, T., Lippman, Z., Gur, A. & Zamir, D. 2006. Overdominant quantitative trait loci for yield and fitness in tomato. *PNAS,* 103 (35) 12981-12986

Seymour, M., Räsänen, K. & Kristjánsson, B. K. 2019. Drift versus selection as drivers of phenotypic divergence at small spatial scales: The case of Belgjarskógur threespine stickleback. *Ecology and evolution,* 9**,** 8133-8145

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A. & Bard, E. 2012. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature,* 484**,** 49-54

Sharp, P. M., Emery, L. R. & Zeng, K. 2010. Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 1203-1212

She, H., Hao, Y., Song, G., Luo, X., Lei, F., Zhai, W. & Qu, Y. 2024. Gene expression plasticity followed by genetic change during colonization in a high-elevation environment. *Elife,* 12**,** RP86687

Shen, W., Le, S., Li, Y. & Hu, F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one,* 11**,** e0163962

Shi, J. & Liang, C. 2019. Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant physiology,* 180**,** 1803-1815

Shinozaki, Y., Nicolas, P., Fernandez-Pozo, N., Ma, Q., Evanich, D. J., Shi, Y., Xu, Y., Zheng, Y., Snyder, S. I. & Martin, L. B. 2018. High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. *Nature communications,* 9**,** 364

Shukla, V., Upadhyay, R. K., Tucker, M. L., Giovannoni, J. J., Rudrabhatla, S. V. & Mattoo, A. K. 2017. Transient regulation of three clustered tomato class-I small heat-shock chaperone genes by ethylene is mediated by SlMADS-RIN transcription factor. *Scientific reports,* 7**,** 6474

Smit, A., Hubley, R. & Green, P. 2015. *RepeatMasker Open-4.0* [Online]. Available: http://www.repeatmasker.org [Accessed].

Smith, O., Nicholson, W. V., Kistler, L., Mace, E., Clapham, A., Rose, P., Stevens, C., Ware, R., Samavedam, S. & Barker, G. 2019. A domestication history of dynamic adaptation and genomic deterioration in Sorghum. *Nature plants,* 5**,** 369-379

Sohrab, V., López-Díaz, C., Di Pietro, A., Ma, L.-J. & Ayhan, D. H. 2021. TEfinder: a bioinformatics pipeline for detecting new transposable element insertion events in next-generation sequencing data. *Genes,* 12**,** 224

Sommer, R. J. 2020. Phenotypic plasticity: from theory and genetics to current and future challenges. *Genetics,* 215**,** 1-13

Song, Y., Ji, D., Li, S., Wang, P., Li, Q. & Xiang, F. 2012. The dynamic changes of DNA methylation and histone modifications of salt responsive transcription factor genes in soybean. *PloS one,* 7**,** e41274

Sousa, A., Bourgard, C., Wahl, L. M. & Gordo, I. 2013. Rates of transposition in Escherichia coli. *Biology letters,* 9**,** 20130838

Soyk, S., Lemmon, Z. H., Oved, M., Fisher, J., Liberatore, K. L., Park, S. J., Goren, A., Jiang, K., Ramos, A., Van Der Knaap, E., Van Eck, J., Zamir, D., Eshed, Y. & Lippman, Z. B. 2017. Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. *Cell,* 169**,** 1142-1155 e12.10.1016/j.cell.2017.04.032

Stam, R., Nosenko, T., Hörger, A. C., Stephan, W., Seidel, M., Kuhn, J. M. M., Haberer, G. & Tellier, A. 2019. The de Novo Reference Genome and Transcriptome Assemblies of the Wild Tomato Species Solanum chilense Highlights Birth and Death of NLR Genes Between Tomato Species. *G3 (Bethesda),* 9**,** 3933-3941

Stapley, J., Santure, A. W. & Dennis, S. R. 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular ecology,* 24**,** 2241-2252

Stephan, W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 1245-1253

Stetter, M. G. 2020. Limits and constraints to crop domestication. *American Journal of Botany,* 107

Stetter, M. G., Gates, D. J., Mei, W. & Ross-Ibarra, J. 2017a. How to make a domesticate. *Current Biology,* 27**,** R896-R900

Stetter, M. G., Müller, T. & Schmid, K. J. 2017b. Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (Amaranthus caudatus). *Molecular ecology,* 26**,** 871-886

Stetter, M. G., Vidal-Villarejo, M. & Schmid, K. J. 2020. Parallel seed color adaptation during multiple domestication attempts of an ancient new world grain. *Molecular Biology and Evolution,* 37**,** 1407-1419

Stinchcombe, J. R., Rutter, M. T., Burdick, D. S., Tiffin, P., Rausher, M. D. & Mauricio, R. 2002. Testing for environmentally induced bias in phenotypic estimates of natural selection: theory and practice. *The American Naturalist,* 160**,** 511-523

Strickler, S. R., Bombarely, A., Munkvold, J. D., York, T., Menda, N., Martin, G. B. & Mueller, L. A. 2015. Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. *PeerJ,* 3**,** e793

Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics,* 43**,** 1160-U164.10.1038/ng.942

Su, W., Gu, X. & Peterson, T. 2019. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Molecular plant,* 12**,** 447-460

Su, X., Wang, B., Geng, X., Du, Y., Yang, Q., Liang, B., Meng, G., Gao, Q., Yang, W. & Zhu, Y. 2021. A high-continuity and annotated tomato reference genome. *BMC genomics,* 22**,** 898

Suganami, M., Kojima, S., Yoshida, H., Mori, M., Kawamura, M., Koketsu, E. & Matsuoka, M. 2024. Low mutation rate of spontaneous mutants enables detection of causative genes by comparing whole genome sequences. *Frontiers in Plant Science,* 15**,** 1366413

Sultan, S. E. 2015. *Organism and environment: ecological development, niche construction, and adaptation*, Oxford University Press

Sun, H., Fan, H.-J. & Ling, H.-Q. 2015. Genome-wide identification and characterization of the bHLH gene family in tomato. *BMC genomics,* 16**,** 1-12

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics,* 105**,** 437-460

Tanksley, S. D. 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *The plant cell,* 16**,** S181-S189

Tanno, K.-I. & Willcox, G. 2006. How fast was wild wheat domesticated? *Science,* 311**,** 1886-1886

Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y. & Shatte, T. 2021. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants,* 7**,** 766-773

The Tomato Genome Consortium 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature,* 485**,** 635-41.10.1038/nature11119

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics,* 14**,** 178-192

Tian, P., Lin, Z., Lin, D., Dong, S., Huang, J. & Huang, T. 2021. The pattern of DNA methylation alteration, and its association with the changes of gene expression and alternative splicing during phosphate starvation in tomato. *The Plant Journal,* 108**,** 841-858

Tieman, D., Zhu, G., Resende Jr, M. F., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M. & Zhang, B. 2017. A chemical genetic roadmap to improved tomato flavor. *Science,* 355**,** 391-394

Tohge, T., Scossa, F., Wendenburg, R., Frasse, P., Balbo, I., Watanabe, M., Alseekh, S., Jadhav, S. S., Delfin, J. C. & Lohse, M. 2020. Exploiting natural variation in tomato to define pathway structure and metabolic regulation of fruit polyphenolics in the lycopersicum complex. *Molecular plant,* 13**,** 1027-1046

Trucchi, E., Benazzo, A., Lari, M., Iob, A., Vai, S., Nanni, L., Bellucci, E., Bitocchi, E., Raffini, F. & Xu, C. 2021. Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. *Nature Plants,* 7**,** 123-128

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. & Kakutani, T. 2009. Bursts of retrotransposition reproduced in Arabidopsis. *Nature,* 461**,** 423-426

Uller, T., Moczek, A. P., Watson, R. A., Brakefield, P. M. & Laland, K. N. 2018. Developmental bias and evolution: a regulatory network perspective. *Genetics,* 209**,** 949-966

Uzunović, J., Josephs, E. B., Stinchcombe, J. R. & Wright, S. I. 2019. Transposable elements are important contributors to standing variation in gene expression in Capsella grandiflora. *Mol Biol Evol,* 36 1734-1745

Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., Martínez González, J., García Cook, A., Montiel, R. & Vielle-Calzada, J.-P. 2016. The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proceedings of the National Academy of Sciences,* 113**,** 14151-14156

Van Tassel, D. L., Dehaan, L. R. & Cox, T. S. 2010. Missing domesticated plant forms: can artificial selection fill the gap? *Evolutionary Applications,* 3**,** 434-452

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M. & Castanera, R. 2019. A benchmark of transposon insertion detection tools using real data. *Mobile DNA,* 10**,** 53

Vitte, C., Fustier, M. A., Alix, K. & Tenaillon, M. I. 2014. The bright side of transposons in crop evolution. *Briefings in Functional Genomics,* 13**,** 276-295.10.1093/bfgp/elu002

Vosman, B., Van't Westende, W. P., Henken, B., Van Eekelen, H. D., De Vos, R. C. & Voorrips, R. E. 2018. Broad spectrum insect resistance and metabolites in close relatives of the cultivated tomato. *Euphytica,* 214**,** 1-14

Vukich, M., Giordani, T., Natali, L. & Cavallini, A. 2009. Copia and Gypsy retrotransposons activity in sunflower (Helianthus annuus L.). *BMC Plant Biology,* 9**,** 1-12

Waddington, C. H. 1977. The evolution of an evolutionist. *Journal of the History of Biology,* 10

Wallace, M., Jones, G., Charles, M., Forster, E., Stillman, E., Bonhomme, V., Livarda, A., Osborne, C. P., Rees, M. & Frenck, G. 2019. Re-analysis of archaeobotanical remains from pre-and early agricultural sites provides no evidence for a narrowing of the wild plant food spectrum during the origins of agriculture in southwest Asia. *Vegetation History and Archaeobotany,* 28**,** 449-463

Wang, J., Hu, Z., Zhao, T., Yang, Y., Chen, T., Yang, M., Yu, W. & Zhang, B. 2015a. Genome-wide analysis of bHLH transcription factor and involvement in the infection by yellow leaf curl virus in tomato (Solanum lycopersicum). *BMC genomics,* 16**,** 1-14

Wang, J., Santiago, E. & Caballero, A. 2016. Prediction and estimation of effective population size. *Heredity,* 117**,** 193-206

Wang, M.-S., Zhang, J.-J., Guo, X., Li, M., Meyer, R., Ashari, H., Zheng, Z.-Q., Wang, S., Peng, M.-S. & Jiang, Y. 2021a. Large-scale genomic analysis reveals the genetic cost of chicken domestication. *BMC biology,* 19**,** 1-16

Wang, X., Kang, J., Wang, H., Wang, S., Tang, B. & Lu, J. 2023a. Phenotypic plasticity plays an essential role in the confrontation between plants and herbivorous insects. *CABI Agriculture and Bioscience,* 4**,** 58

Wang, Y., Liu, X. Y., Yang, Y. Z., Huang, J., Sun, F., Lin, J., Gu, Z. Q., Sayyed, A., Xu, C. & Tan, B. C. 2019a. Empty Pericarp21 encodes a novel PPR-DYW protein that is required for mitochondrial RNA editing at multiple sites, complexes I and V biogenesis, and seed development in maize. *PLoS Genet,* 15**,** e1008305.10.1371/journal.pgen.1008305

Wang, Y., Mcneil, P., Abdulazeez, R., Pascual, M., Johnston, S. E., Keightley, P. D. & Obbard, D. J. 2023b. Variation in mutation, recombination, and transposition rates in Drosophila melanogaster and Drosophila simulans. *Genome Research,* 33**,** 587-598

Wang, Y., Mostafa, S., Zeng, W. & Jin, B. 2021b. Function and mechanism of jasmonic acid in plant responses to abiotic and biotic stresses. *International Journal of Molecular Sciences,* 22**,** 8568

Wang, Y. & Obbard, D. J. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. *Evolution Letters,* 7**,** 216-226

Wang, Y., Tang, X., Cheng, Z., Mueller, L., Giovannoni, J. & Tanksley, S. D. 2006. Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. *Genetics,* 172**,** 2529-2540

Wang, Y., Zhang, J., Hu, Z., Guo, X., Tian, S. & Chen, G. 2019b. Genome-wide analysis of the MADS-box transcription factor family in Solanum lycopersicum. *International journal of molecular sciences,* 20**,** 2961

Wang, Z., Wang, Y., Hong, X., Hu, D., Liu, C., Yang, J., Li, Y., Huang, Y., Feng, Y. & Gong, H. 2015b. Functional inactivation of UDP-N-acetylglucosamine pyrophosphorylase 1 (UAP1) induces early leaf senescence and defence responses in rice. *Journal of experimental botany,* 66**,** 973-987

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical population biology,* 7**,** 256-276

Wei, L. & Cao, X. 2016. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Science China Life Sciences,* 59**,** 24-37

Weng, M.-L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., Weigel, D. & Fenster, C. B. 2019. Fine-grained analysis of spontaneous mutation spectrum and frequency in Arabidopsis thaliana. *Genetics,* 211**,** 703-714

West-Eberhard, M. J. 1989. Phenotypic plasticity and the origins of diversity. *Annual review of Ecology and Systematics***,** 249-278

West-Eberhard, M. J. 2003. *Developmental plasticity and evolution*, Oxford University Press

Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., Smit, A. F. & Finn, R. D. 2012. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic acids research,* 41**,** D70-D82

Whitlock, M. C. 2000. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution,* 54**,** 1855-1861

Wicker, T., Gundlach H., Spannagi M., Uauy C., Borrill P., Ramirez-Gonzalez H., R., D. O. & Et Al. 2018. The impact of transposable elements on genome structure and evolution in bread wheat. *BioRxiv,*

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmiguel, P. & Schulman, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics,* 8**,** 973–982

Williams, G. C. 1966. *Adaptation and Natural Selection*, Princeton University Press, Princeton

Willis, K. 2017. *State of the world's plants 2017*, Royal Botanics Gardens Kew

Wiser, M. J., Ribeck, N. & Lenski, R. E. 2013. Long-term dynamics of adaptation in asexual populations. *Science,* 342**,** 1364-1367

Woltereck, R. 1909. Weitere experimentelle Untersuchungen uber Artveranderung, speziell uberdas Wesen quantitativer Artunterschyiede bei Daphniden. *Verh. D. Tsch. Zool. Ges,* 1909**,** 110-172

Wood, D. & Lenné, J. M. 2018. A natural adaptive syndrome as a model for the origins of cereal agriculture. *Proceedings of the Royal Society B: Biological Sciences,* 285**,** 20180277

Wood, D. P., Holmberg, J. A., Osborne, O. G., Helmstetter, A. J., Dunning, L. T., Ellison, A. R., Smith, R. J., Lighten, J. & Papadopulos, A. S. T. 2023. Genetic assimilation of ancestral plasticity during parallel adaptation to zinc contamination in Silene uniflora. *Nature Ecology & Evolution,* 7**,** 414-423.10.1038/s41559-022-01975-w

Woodrow, P., Pontecorvo, G., Ciarmiello, L. F., Fuggi, A. & Carillo, P. 2011. Ttd1a promoter is involved in DNA–protein binding by salt and light stresses. *Molecular biology reports,* 38**,** 3787-3794

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., Mcmullen, M. D. & Gaut, B. S. 2005. The effects of artificial selection on the maize genome. *Science,* 308**,** 1310-1314

Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. 2008. Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences,* 169**,** 105-118

Wu, J., Liu, S., He, Y., Guan, X., Zhu, X., Cheng, L., Wang, J. & Lu, G. 2012. Genome-wide analysis of SAUR gene family in Solanaceae species. *Gene,* 509**,** 38-50

Wu, Q., Han, T.-S., Chen, X., Chen, J.-F., Zou, Y.-P., Li, Z.-W., Xu, Y.-C. & Guo, Y.-L. 2017. Long-term balancing selection contributes to adaptation in Arabidopsis and its relatives. *Genome biology,* 18**,** 1-15

Wund, M. A. 2012. Assessing the impacts of phenotypic plasticity on evolution. Oxford University Press

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & Van Der Knaap, E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science,* 319**,** 1527-1530.10.1126/science.1153040

Xiaoyang, W., Dan, C., Yuqing, L., Weihua, L., Xinming, Y., Xiuquan, L., Juan, D. & Lihui, L. 2017. Molecular characteristics of two new waxy mutations in China waxy maize. *Molecular breeding,* 37**,** 1-7

Xie, Z., Wang, L., Wang, L., Wang, Z., Lu, Z., Tian, D., Yang, S. & Hurst, L. D. 2016. Mutation rate analysis via parent–progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proceedings of the Royal Society B: Biological Sciences,* 283**,** 20161016

Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences,* 111**,** 10263-10268

Xu, Z. & Wang, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research,* 35**,** W265-W268

Xue, S., Bradbury, P. J., Casstevens, T. & Holland, J. B. 2016. Genetic architecture of domestication-related traits in maize. *Genetics,* 204**,** 99-113

Yamada, K., Fukao, Y., Hayashi, M., Fukazawa, M., Suzuki, I. & Nishimura, M. 2007. Cytosolic HSP90 regulates the heat shock response that is responsible for heat acclimation in Arabidopsis thaliana. *Journal of Biological Chemistry,* 282**,** 37794-37804

Yamasaki, M., Wright, S. I. & Mcmullen, M. D. 2007. Genomic screening for artificial selection during domestication and improvement in maize. *Annals of Botany,* 100**,** 967-973

Yang, N., Xu, X.-W., Wang, R.-R., Peng, W.-L., Cai, L., Song, J.-M., Li, W., Luo, X., Niu, L. & Wang, Y. 2017. Contributions of Zea mays subspecies mexicana haplotypes to modern maize. *Nature communications,* 8**,** 1874

Yang, Q., Li, Z., Li, W. Q., Ku, L. X., Wang, C., Ye, J. R., Li, K., Yang, N., Li, Y. P., Zhong, T., Li, J. S., Chen, Y. H., Yan, J. B., Yang, X. H. & Xu, M. L. 2013. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 16969-16974.10.1073/pnas.1310949110

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D. & Tian, D. 2015. Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature,* 523**,** 463-467

Yasuda, K., Ito, M., Sugita, T., Tsukiyama, T., Saito, H., Naito, K., Teraishi, M., Tanisaka, T. & Okumoto, Y. 2013. Utilization of transposable element mPing as a novel genetic tool for modification of the stress response in rice. *Mol Breed.,* 32**,** 505-516

Yeats, T. H., Buda, G. J., Wang, Z., Chehanovsky, N., Moyle, L. C., Jetter, R., Schaffer, A. A. & Rose, J. K. 2012. The fruit cuticles of wild tomato species exhibit architectural and chemical diversity, providing a new model for studying the evolution of cuticle function. *The Plant Journal,* 69**,** 655-666

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N. & Korneliussen, T. S. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *science,* 329**,** 75-78

Yin, S., Ao, Q., Tan, C. & Yang, Y. 2021. Genome-wide identification and characterization of YTH domain-containing genes, encoding the m6A readers, and their expression in tomato. *Plant Cell Reports,* 40**,** 1229-1245

Yin, S., Wan, M., Guo, C., Wang, B., Li, H., Li, G., Tian, Y., Ge, X., King, G. J. & Liu, K. 2020. Transposon insertions within alleles of BnaFLC. A10 and BnaFLC. A2 are associated with seasonal crop type in rapeseed. *Journal of Experimental Botany,* 71**,** 4729-4741.https://doi.org/10.1093/jxb/eraa237

You, F. M., Cloutier, S., Shan, Y. & Ragupathy, R. 2015. LTR annotator: automated identification and annotation of LTR retrotransposons in plant genomes. *International Journal of Bioscience, Biochemistry and Bioinformatics,* 5**,** 165

Yu, Z., Zhang, F., Friml, J. & Ding, Z. 2022. Auxin signaling: Research advances over the past 30 years. *Journal of Integrative Plant Biology,* 64**,** 371-392

Zaki, H. E. & Yokoi, S. 2016. A comparative in vitro study of salt tolerance in cultivated tomato and related wild species. *Plant Biotechnology,* 33**,** 361-372

Zeder, M. A. 2015. Core questions in domestication research. *Proceedings of the National Academy of Sciences,* 112**,** 3191-3198

Zeven, A. C. & De Wet, J. M. 1982. *Dictionary of cultivated plants and their regions of diversity: excluding most ornamentals, forest trees and lower plants*, Pudoc

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics,* 35**,** 1786-1788

Zhang, L., Kang, J., Xie, Q., Gong, J., Shen, H., Chen, Y., Chen, G. & Hu, Z. 2020. The basic helix-loop-helix transcription factor bHLH95 affects fruit ripening and multiple metabolisms in tomato. *Journal of Experimental Botany,* 71**,** 6311-6327

Zhong, L., Yang, Q., Yan, X., Yu, C., Su, L., Zhang, X. & Zhu, Y. 2017. Signatures of soft sweeps across the Dt1 locus underlying determinate growth habit in soya bean [Glycine max (L.) Merr.]. *Molecular Ecology,* 26**,** 4686-4699

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M. & Yang, C. 2018. Rewiring of the fruit metabolome in tomato breeding. *Cell,* 172**,** 249-261. e12

Zhu, J.-Y., Sae-Seaw, J. & Wang, Z.-Y. 2013. Brassinosteroid signalling. *Development,* 140**,** 1615-1620.10.1242/dev.060590

Zohary, D. 2004. Unconscious selection and the evolution of domesticated plants. *Economic botany,* 58**,** 5-10