Haplotype analyses reveal novel insights into tomato history and domestication to show long-distance migrations and latitudinal adaptations

Long-distance migrations and adaptations in the history of tomato

Authors

Jose Blanca[1]*, David Sanchez-Matarredona[1], Peio Ziarsolo[1], J Montero-Pau[2], Esther van der Knaap[3,4], Mª José Díez[1], Joaquín Cañizares[1]

Affiliations

[1] Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana, COMAV, Universitat Politècnica de València, Spain.

[2] Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universitat de València, Spain.

[3] Institute of Plant Breeding, Genetics and Genomics, University of Georgia, GA, USA.

[4] Department of Horticulture, University of Georgia, GA, USA.

Author for correspondence: *Jose Blanca

jblanca@upv.es

Abstract

A novel haplotype-based approach that uses Procrustes analysis and automatic classification was used to provide further insights into tomato history and domestication. Agrarian cultures domesticated species of interest by introducing complex genetic modifications. For tomato, two species, and for one of them two botanical varieties: are thought to be involved in its domestication: the fully wild *Solanum pimpinellifolium* (SP), the wild and semidomesticated *S. lycopersicum* var. *cerasiforme* (SLC) and the cultivated *S. l.* var. *lycopersicum* (SLC). The new approach showed that SP evolved into SLC during a gradual migration from the Peruvian deserts to the Mexican rainforests and that Peruvian and Ecuadorian SLCs were the result of more recent hybridizations. This model was supported by independent evidence that included accession

collection site ecological data and morphological data. In further support of this model, we show that photosynthesis-, flowering time- and circadian rhythm-related genes were selected during the latitudinal migrations.

Teaser

Tomato was domesticated in Ecuador and Peru from plants originally arrived from Mesoamerica after modifying genes related to latitudinal adaptations.

Introduction

Cultivated plants are the result of domestication processes that alter the morphology, physiology, and genetics of wild species to benefit human needs and preferences. These processes usually involve a domestication syndrome, which involves the modification of a set of traits (Hammer, 1984) favored by humans and/or that provide growth advantages under cultivation or adaptations to disturbed habitats (Meyer and Purugganan, 2013). In horticultural crops, these traits usually include larger and more nutritious fruits, robust stems, and reduction in seed dormancy (Yang et al., 2019). Selection, bottlenecks, and outcrossing with wild and feral populations are common during domestication. Moreover, the domestication histories are intertwined with the history of the agrarian cultures that carried them out, and complex migrations and interchanges between different geographic regions can also occur. Thus, the extant population genetic structure and pattern of morphological diversity are often complex. In addition to historical interest, the study of these processes has practical implications, as they generate knowledge about genes and pathways of agronomic interest (Zsögön et al., 2018).

The fully wild *Solanum pimpinellifolium* L. (SP) and *S. lycopersicum* L. (SL) are two sister Solanaceae species (genus *Solanum* L., section *Lycopersicon* (Peralta et al., 2008)) that are capable of interbreeding. SL is split into two botanical varieties: *S. l.* var. *lycopersicum* L. (SLL) and *S. l.* var. *cerasiforme* (Dunal) Spooner, G.J. Anderson & R.K. Jansen (SLC) (Peralta et al., 2008). SP has been proposed to be the species from which the cultivated tomato forms were domesticated (Peralta et al., 2008). SP is divided in several

populations associated with different climates and ecological niches: the dry Peruvian coast, the northern Peruvian and southern Ecuadorian Andean valleys and the wet northern Ecuadorian coast (Gibson and Moyle, 2020). However, recent studies have shown that SP likely evolved into SLC before human colonization of the Americas (Razifard et al., 2020). Therefore, an ancestral wild SLC population might have been involved in the tomato domestication. SLL is cultivated, whereas SLC is composed of a complex mix of wild, semidomesticated, and vintage Peruvian, Ecuadorian and Mesoamerican varieties (Blanca et al., 2015; Rick and Holle, 1990). Furthermore, as a feral and weedy species, SLC has colonized subtropical regions all over the world after the arrival of Europeans in America (Rick and Holle, 1990).

The most accepted model for tomato domestication is a two-step process (Blanca et al., 2015, 2012; Gao et al., 2019; Lin et al., 2014; Razifard et al., 2020). According to this model, the desert-dwelling, most diverse and wild Peruvian SP (SP Pe) composes the most ancient population. In a slow process unrelated to human activity, SP Pe adapted to the climate conditions in the Peruvian and Ecuadorian Andean valleys (SP Montane) and to the northern humid Ecuadorian regions (SP Ec). Ecuadorian SLC (SLC Ec), which composes the most diverse SLC population and whose genetic distance to any SP is the shortest, would be the first SL type derived from SP (Blanca et al., 2012). SLC Ec occupies the humid Amazonian regions in Ecuador closer to the Andes, also known as Ceja de la Montaña (Rick and Holle, 1990). Afterwards, SLC would have moved south from Ecuador to northern Peru, where early farmers might have started domestication. In Ecuador and northern Peru, there are SLC accessions with a domesticated fruit morphology, and several Peruvian and Ecuadorian cultivated SLCs were collected in markets in the early part of the last century. Finally, the Peruvian cultivated tomatoes would have migrated to Mesoamerica, and there, in a second phase of improvement, SLL would have emerged.

However, there are complexities about the tomato domestication history that remain to be explained (Blanca et al., 2012; Razifard et al., 2020). For instance, SLC Ec is the most genetically diverse SLC, but it is proposed to be derived from a less diverse population—SP Ec. The conclusions of prior evolutionary genetic analyses were supported, at least partially, on traditional population indices such genetic diversities

or linkage disequilibria (LDs). Despite being informative, these values can be misleading because they may be explained by different hypotheses. For instance, high genetic diversity is typically found in ancient and well-established populations as well as in recent admixtures. When several measures are combined, certain hypotheses can be effectively refuted. However, when the evolutionary history is complex, the conclusions based on these traditional indices remain tentative at best. As alternatives to these nonparametric approaches, complex statistical models may be used. However, the nonparametric methods also have limitations, as they tend to make unrealistic assumptions such as no Hardy-Weinberg violations and/or depend on parameters that are difficult to establish, such as the number of ancestral migrations (Pickrell and Pritchard, 2012; Raj et al., 2014; Razifard et al., 2020).

Recently, Razifard *et al*. (Razifard et al., 2020) interpreted the high genetic diversity of the SLC Ec key population by assuming that it was an ancient and wild population closely related to SP. However, their own data is not consistent with this hypothesis: SLC Ec appears to be an admixture according to their fastSTRUCTURE result, and, it has higher LD than Mexican SLC, but we would expect LD to be lower in the older population.

Linguistic and historical evidence might complement the genetic data in the study of domestication histories. However for tomato, there is scant linguistic and historical evidence, and they are ambiguous and subject to different interpretations (Iris Peralta and David Spooner, 2011). Moreover, few archeological remains of tomato have been uncovered. This might be due, at least in part, to the perishability of the tomato (Kiple et al., 2000; Pickersgill, 2016). As far as we are aware, there are just two reports of tomato archeological remains, both seeds in coprolites, one in Southern Texas, close to the Mexican border, dated ~2500 B. C (Reinhard et al., 2008), and one in the Peruvian Ica valley, from ~500 A. D. In both cases the tomatoes were ingested, but there is no information regarding the degree of domestication (Beresford-Jones et al., 2011).

Thus, despite past efforts, none of the tomato domestication models capture all molecular, morphological and passport data coherently. To gain deeper insights into the domestication history of tomato, we developed a new approach based on a nonparametric unsupervised automatic classification of haplotypic principal coordinate analyses (PCoAs) aligned via Procrustes (Krzanowski, 2000). The results from this analysis implied that SP evolved into SLC during a slow migration from Peru to Mesoamerica and that Peruvian and Ecuadorian SLCs are admixed populations originating from Mesoamerican SLC and SP. Subsequently, Peruvian domesticated SLCs migrated north to Mexico, and then evolved into SLL. Despite differing from the previous proposed evolutionary models, our model is in agreement with all evidence available, such as LD, genetic diversity and distance, the results from fastSTRUCTURE (Raj et al., 2014) and TreeMix (Pickrell and Pritchard, 2012), as well as morphological and collection site ecological data. Moreover, other analyses showed evidence that selection of genes associated with photosynthesis and flowering time might have been involved in these long-distance migrations between different climate types and latitudes.

Results

Haplotypic PCoAs and automatic classification

The euchromatic regions of the entire genome were divided into 440 segments of 0.5 Mb each, resulting in a total of 526,240 haplotypes (two per plant accession per 440 segments). The output of the PCoAs carried out with the haplotypes of each segment were aligned using Procrustes and resulted in a triangular-like structure that matched the structure of the main taxonomic groups (Fig. 1). The haplotypes were automatically classified by an unsupervised clustering algorithm into three types. The three haplotype types were named according to the taxonomic groups in which they were usually found: hPe (most abundant in Peruvian SP), hEc (most abundant in Ecuadorian SP and in Ecuadorian SLC), and hSL (commonly found in SLC and SLL). The chosen number of haplotype types, 3, was in good agreement with the number of ancestral

populations suggested by the fastSTRUCTURE marginal likelihoods (Fig. S1). However, the Calinski-Harabasz (CH) clustering score, an index commonly used to evaluate clustering performance, suggested a lower number of types (Fig. S1). This might be due to the intermediate haplotypes found between the main clusters, which is likely due to recombinant and intermediate haplotypes, and to the uneven representation of the different haplotype types: the SLL accessions were overrepresented, with 44% of the sequenced accessions, whereas the SP Ec, a traditionally ignored population, was underrepresented with just 2.5% of the accessions. To check the effect of the number of haplotype types, the haplotype classification was also carried out with 2, 3, 4, and 5 haplotype types (Fig. S2). When 2 types were used, the haplotype split divided mainly SP from SL, whereas when more types were allowed SP was divided in subtypes, like, for instance, Ecuadorian and Peruvian SP for 3 types. The relationship between SL and SP, the main focus of the current analysis, remained unchanged when the number of haplotype types altered.

Haplotypic classification was used to analyze the genomic composition of each accession (Fig. S3 and Fig. S4). As an example, the haplotype composition for Cervil, a modern accession, is shown in Fig. S5 (Causse et al., 2013). The haplotype classification and location in the aligned PCoA results showed that this accession comprised hPe and hSL haplotypes. Therefore, it is likely that Cervil is the result of a hybridization between Peruvian SP and a cultivated tomato plant.

Accession classification and haplotype composition

The accessions were classified into genetic groups through a series of hierarchical PCoAs, calculated from the genetic distance matrix (Fig. 2 and Fig. S6), and the information provided by the geographic and taxonomic passport data (Table S1) and their haplotype composition (Figs. S3 and S4). SP was split into three populations: SP Pe (Peru), SP Montane, and SP Ec (Ecuador). The most abundant SLC populations were SLC MA (Mesoamerica), SLC Pe, and SLC Ec, and

SLL was composed of SLL Mx (Mexico), SLL vintage, and SLL modern (Table S1). Other minor populations were noted, such as SLC Co (Colombia), but they were represented by only a few accessions, thus not all analyses could be performed on them. The overall genetic group separation was similar to the population separations used in previous studies (Fig. S7) (Razifard et al., 2020).

The haplotype composition of each population was assessed by plotting the aligned PCoA of the haplotypes of the accessions that belonged to each population (Fig. 3 A). The ancestral population composition calculated from the fastSTRUCTURE results was similar to the haplotype composition both at the population (Fig. 3 B and C) and at the individual level (Fig. S4).

To check if the size chosen for the genome segments could affect the haplotype analyses, they were repeated with different genome segment sizes (100 kb, 500 kb, and 1000 kb), and the results were in all cases very similar (Fig. S8).

The genetic diversity of the haplotypes that belong to one or the other haplotype type (hPe, hEc, hSL) was calculated for each population (Fig. 4). SP Pe was the most diverse population for haplotype type hPe; SP Ec for hEc; and Mesoamerican SLC, for hSL. These diversity results were in agreement with the fastSTRUCTURE results, which also suggested that three haplotype types (hPe, hEc, hSL) corresponded to these three ancestral populations (Fig. 3 C). Furthermore, the haplotypes associated with hPe and hEc were most abundant in the extant SP Pe and SP Ec populations, respectively (Fig. 3 B). SL populations were genetically close, and they contained mostly hSL haplotypes. However, SLC MA presented the highest hSL haplotype diversity (Fig. 4), suggesting a Mesoamerican origin for the typical hSL haplotype type.

Overlapping with the divisions among Peru, Ecuador and SL, the population haplotype composition made evident a pattern of secondary contacts. For instance, modern SLL carried certain hPe haplotypes that were less common in Mexican SLL and in vintage SLL (Fig. 3 A and B). Even more striking was the mixture of haplotypes found in the Andean SLC populations. Although most haplotypes found in SLC Pe were hSL, certain hPe and hEc were also present in that population. Remarkably, a quarter of the haplotypes found in Ecuadorian SLC were not hSL but instead were hEc; thus, they might have been introgressed from SP Ec. To further investigate the complex pattern of gene flow, we employed TreeMix (Fig. S9). According to the TreeMix analysis (Fig. S9 A) SLC MA appeared to be basal to all other SLCs. SLC Pe and SLC Ec were then derived from SLC MA by acquiring introgressions from SP Pe and SP Ec, respectively. These results were also in agreement with the ABBA-BABA analyses (Green et al., 2010). When SLC Ec, SLC MA, SP Ec, and SP Pe were used as populations the value calculated for the D statistic was -0.23, whereas for the populations SLC Pe, SLC MA, SP Ec, and SP Pe the D value was 0.43. In both cases the p-value was 0. These D statistics are compatible with SLC Ec receiving introgressions mainly from SP Ec, whereas SLC Pe would have introgressed genomic segments mainly from SP Pe.

Diversity and linkage disequilibrium

The overall number of polymorphic (95% criteria) genetic variants (Fig. 5 A), the mean number of variants found in a genomic region (Fig. 5 B) and the unbiased Nei genetic diversity (Fig. S10) all gave similar results. According to these indexes, the most diverse populations were the Peruvian and Ecuadorian SP and SLC, as well as modern SLL. SLC MA, SLC World, and, particularly, SLL Mx were characterized by a relatively low genetic diversity. These results matched those of previous analyses (Blanca et al., 2015, 2012) but contrasted with those of another diversity measure—the mean number of different haplotypes found in a given genomic

region (Fig. 5 D). According to this index, the SP populations were the most diverse, whereas all SLC populations showed lower levels of diversity. Although SLC Ec and SLC Pe are seemingly highly variable according to the number of polymorphic variants, these populations do not have many different haplotypes. These results also indicate that SLC Ec and SLC Pe could be the result of an admixture of an ancient SLC population together with a SP, something already suggested by the haplotype composition, the fastSTRUCTURE, the TreeMix and the ABBA-BABA analyses. LDs of recently created populations or of populations that recently incorporated genetic material are usually high and the SLC population with the lowest LD is SLC MA (Fig. 5 C).

Latitude-related selection

SP Pe is according to all previous researchers, and all the evidence presented, the oldest population, thus, SLC MA would be the result of a northward migration. In these migrations some genomic regions could have been selected, and to study the selective sweep possibility the expected heterozygosity was calculated along the genome for SLC MA (Fig. 6 A). SLC Ec and SLC Pe seem to be derived from SLC MA, but, according to the haplotype, TreeMix, fastSTRUCTURE and ABBA-BABA analyses, they have introgressions from SP. Thus, we calculated the D, BABA, and ABAA indexes along the genome assuming the following evolutionary schemas: 1) SLC Ec, SLC MA, SP Ec, and SP Pe, 2) SLC Pe, SLC MA, SP Ec, and SP Pe (Fig. 6 B, Fig. S11). There are some genomic regions that have a high introgression frequency, like the chromosome 1 end or the chromosome 4 just before the pericentromeric region. We analyzed the possible relation between the regions with lowest diversity in SLC MA and with the highest introgression rate in SLC Ec (Fig. 6 C). The regions more introgressed in SLC Ec have lower diversity in SLC MA. This result seems to indicate that the selection process experienced during the northward migration might have be partially reversed by introgessing from SP after the Ecuadorian colonization by SLC. Additionally, we tested this hypothesis by

calculating in SLC Ec the frequency of SP alleles not present in SLC MA, and we related these possible introgressions with the SLC MA diversity (Fig. S12). In this analysis the genomic regions with an abundance of SP alleles are also correlated with the regions with lower diversities in Mesoamerican SLCs.

We inspected the genes found in regions with high introgression rate in SLC Ec and SLC Pe and low diversity in SLC MA (Table S2). Some regions, like the ones in chromosome 2, 8 or 11 were large and comprised hundreds of genes, whereas other, like the one in chromosome 7 were smaller. In this region only 5 genes were found and one of them was Solyc07g043270, a FAR-red elongated Hypocotyl 3-like protein encoding gene. This FAR-red genes genes respond to the light and have been related to the flowering time and other processes regulated by light conditions (Li et al., 2011; Xie et al., 2020). The FAR-red genes were found in three of these regions. In chromosome 4 two of the 22 genes found were a spermidine synthase (Solyc04g026030) that might be involved in the flowering time regulation (Imamura et al., 2015) and an Agamous protein possibly involved in flowering and fruit development (Pan et al., 2010). In total three Agamous genes were found in different regions. Also, in chromosome 2, Solyc02g021650, a component of the light signal transduction machinery involved in repression of photomorphogenesis (Lieberman et al., 2004), and in chromosome 6, Solyc06g050620, a reticulata related family gene associated, among other processes, with the chloroplast development were found (Pérez-Pérez et al., 2013).

Mexican SLL origin

To further determine the relationship between the Mesoamerican and Peruvian SLCs and the Mexican SLL, a detailed haplotype-based analysis was carried out (Fig. 7). The populations with the most private haplotypes were the Mesoamerican SLC and Peruvian SLC populations. Many

private Peruvian SLC haplotypes appeared to be the result of introgressions from Peruvian and Ecuadorian SP. Mexican SLL had the fewest private haplotypes. The two population pairs that shared the most haplotypes were Mesoamerican and Peruvian SLC and Peruvian SLC and Mexican SLL; however, despite their geographic closeness, SLC MA and SLL Mx shared fewer haplotypes. Thus, according to these results, it is plausible that Mexican SLL originated from Peruvian SLC and not from the geographically closer Mesoamerican SLC.

Morphological analysis

A morphological characterization of SP, SLC and SLL was also carried out (Table S3). A Principal component analysis (PCA) for leaf-, inflorescence-, fruit- and stem-related traits was used to cluster accessions into several morphological types (Fig. 8 A, Fig. S13, Fig. S14, Fig. S15). Three morphological types were found in SP, of which the first featured longer inflorescences, wider petals, and, frequently, striped fruits, as well as curved and exerted styles compared with those of the other types. This first morphological type was characteristic of Peruvian SP. Compared with the other types, the second type comprised mostly northern Ecuadorian SP accessions and featured slightly larger fruits, which usually showed a distinct transversally elongated shape (peanut-like shape). Finally, an intermediate SP morphological type found in Peru and Ecuador was characteristic of accessions found in the mountainous valleys located between Peru and Ecuador. SLC was mainly divided into three morphological types: SLC small, SLC big and SLC Ecu. SLC small had small fruits, with sizes that ranged from small SP fruit size to cherry size, whereas SLC big had relatively large fruits that varied in size from cherry size to full-size commercial fruits. These relatively large fruits were frequently ribbed and almost always flat. Fruit size was found to be associated with stem width. Finally, the endemic Ecuadorian SLC form was characterized as having a mixture of characteristics between different SLC and SP types: it had cherry-sized fruits like those of SLC, but folded back petals and longer

inflorescences, like the intermediate SP type did, and no stem pilosity and some transversally elongated fruits, like those of Ecuadorian SP.

These morphological types were associated with the molecularly defined populations (Fig. 8 B). For this analysis, SLC Pe was divided into two genetic subpopulations, northern Peruvian SLC (SLC Pe N) and southern Peruvian SLC (SLC Pe S) (Fig. S6 B and C), which exhibited differential morphological characteristics despite being closely genetically related. The SP Pe and SP Ec genetic groups corresponded roughly to the Peruvian and Ecuadorian SP morphological types. The SP Montane population was characterized by the intermediate type that was also sometimes found in other SP populations. Mesoamerican SLC produced mainly small fruits, whereas northern Peruvian SLC was characterized mainly by SLC big or even SLL morphological types. The other American SLC populations, southern Peruvian SLC and Ecuadorian SLC, were morphologically variable and included morphological types that produced large and small fruits and, in the Ecuadorian case, also the typical Ecuadorian SLC-type.

Additionally, an analysis of the collection sites taken from the passport data was carried out (Fig. S16). Undisturbed environments were labeled natural, whereas accessions collected in human-altered environments such as roadsides were considered ruderal. Semi-cultivated or cultivated accessions were collected mainly in backyards. The collected data were manually curated, and sometimes, these passport data were complemented by the information obtained from images of the collection sites. In total, 262 accessions had associated collection data. Most SP accessions (160 vs 10) and 65% (17 vs 9) of the SLC MA accessions were wild or ruderal while the rest were weedy or semicultivated. These findings contrast with the data concerning Ecuadorian SLC, which was semi-cultivated or cultivated on 89% of the occasions (25 vs 3 accessions).

Discussion

Previous studies have reported a complex history of SLC, the wild and semidomesticated variety related to cultivated tomato (SLL) (Fig. 9) (Blanca et al., 2015, 2012; Razifard et al., 2020; Rick and Holle, 1990). However, even with the genomic evidence available, no detailed tomato evolutionary model capable of accounting for all the empirical evidence has ever been produced. In the current analysis, the wild and cultivated genetic diversity present among SP, SLC and SLL were analyzed considering all the publicly available whole-genome resequencing data mapped to the latest tomato genome reference (v4.0) (Hosmani et al., 2019), as well as the morphological and passport data gathered from various gene banks. Moreover, we developed a novel nonparametric method capable of performing a genome-wide haplotypic analysis by combining Procrustes-aligned PCoA output with an automatic unsupervised classification. Following this approach, we classified the more than half a million haplotypes present in all the individuals. This new method allows a detailed and quantitative inspection of the haplotype composition of each accession and population, and thus, this method is useful for studying gene flow, introgressions and migrations without the need for any assumptions related to Hardy-Weinberg equilibrium or the reproductive system of the species involved. The only limitation is to have enough LD; otherwise, the haplotypes would not be informative. To our knowledge, Procrustes has never been used for this purpose before, and its use in population genetics has been restricted to the alignment of genomic PCA data and geographic maps (Wang et al., 2010) or to the alignment of PCA data generated from different SNP datasets (Wang et al., 2015). The domestication model obtained is backed by traditional population genetic indexes, parametric statistical models, morphological and passport data. We hope that similar approaches could be used to study the complex domestication histories of other species.

Three wild original populations

This novel analysis provided empirical evidence that suggests a tomato evolutionary model that accounts for all previous problematic results (Fig. 9 A). The automatic classification of the Procrustes-aligned haplotypic PCoA results suggested the existence of three haplotype types (hPe, hEc, and hSL) related to the main tomato taxonomic groups, in agreement with the fastSTRUCTURE results (Fig. 3 B and C and Fig. S4). The haplotype composition, fastSTRUCTURE, and genetic diversity of each type of haplotype (Fig. 4) suggest that the three haplotype types (hPe, hEc, and hSL) originated in three ancient populations related to the three extant SP Pe, SP Ec, and SL MA populations (Fig 9. A). SP Pe, the only population composed mainly of hPe haplotypes, was also the most diverse population. Owing to its diversity and abundance, this population has been considered in all previous studies to be SP's center of origin (Rick and Fobes, 1975). hEc haplotypes were mainly present in the two Ecuadorian populations, namely, SP Ec and SLC Ec (Fig. 3 A and B), but they were more common and diverse in SP Ec (Fig. 4), so they seem to be ancient Ecuadorian SP haplotypes and not SLC haplotypes.

All SLC populations, which were distributed from Mexico to southern Peru, were genetically close and composed mainly of hSL haplotypes. However, SLC MA was found to be the most ancient SLC (Fig. 9, A): its hSL haplotypes were the most diverse (Fig. 4) and it showed the lowest LD (Fig. 5 C), as is to be expected from an older population.

Passport data represent additional completely independent evidence in favor of SLC MA's ancient origin (Fig. S16). In Mexico, SLC has been collected mostly as a wild species in natural or disturbed environments and not under cultivation. Unfortunately, collectors were more interested in the plants' degree of cultivation and did not differentiate between wild and ruderal environments. However, a recent study published involving an on-site evaluation reported wild

SLC in tropical and mesophilic forests as well as in shrublands (Álvarez-Hernández et al., 2009). This abundance of wild Mesoamerican SLC contrasted with its absence in the Andean region. In northern Peru and Ecuador, SLC is mainly cultivated or semicultivated in backyards. Semicultivated plants are those that, despite seldomly being planted, are cared for by backyard owners, as Rick and Holle reported (Rick and Holle, 1990). To our knowledge, Ecuadorian and northern Peruvian SLCs have never been reported in a natural environment. Moreover, some Ecuadorian SLCs are commercially cultivated and have on some occasions been mistakenly collected as vintage SLLs because they produced large fruits and are sold in markets (Blanca et al., 2015). In southern Peru SLC, the ecology of SLC is different and is commonly found in disturbed environments. This SLC behavior is also observed in most other subtropical areas worldwide. SLC has become a very successful invader associated with human-modified environments.

The genetic distance between hSL haplotypes to the SP haplotypes is similar to the distance between hPe and hEc haplotypes. This would suggest that SLC MA would have appeared after a slow migration from South America to Mesoamerica. Whereas, Mesoamerican and Andean hSL haplotypes had no time to differentiate, indicating that the migration from Mesoamerica to Peru and Ecuador was the result of a fast and recent event. Ruderal and weedy behavior could explain why SLC migrated back from Mesoamerica to Ecuador. SLC could have arrived in the Andean region from Mesoamerica as a weed associated with the import of other crop species, such as Mesoamerican maize. Archeological records indicate that maize arrived in Ecuador approximately ∼7000–5500 calibrated years before present (Grobman et al., 2012; Meyer and Purugganan, 2013). If this hypothesis is true, it implies that the weedy SLC that arrived in the Andean region would have produced small fruits. The morphological characterization of the extant Mesoamerican SLC is in agreement with this hypothesis. Mesoamerican SLC typically

produces small fruits. However, without direct archeological evidence, any hypotheses regarding the fruit phenotypes of populations thousands of years old are, by necessity, somewhat speculative, especially in populations that have suffered extensive migrations.

Moreover, haplotype composition, fastSTRUCTURE, TreeMix and ABBA-BABA results clearly showed that the Ecuadorian and Peruvian SLC populations are admixtures between SLC MA and SP probably created after the older SLC MA migrated south to the Amazonian region and introgressed some Ecuadorian and Peruvian alleles. The TreeMix analysis carried out with less stringent LD thresholds (Fig. S9 B) showed different tree topologies, thus caution is advised when interpreting the TreeMix results. This is lack of robustness is a limitation acknowledged by the TreeMix authors (Pickrell and Pritchard, 2012), and it might have caused problems in the Razifard et al. TreeMix analysis carried out with a LD threshold that overrepresented the highly linked tomato pericentromeric regions (Razifard et al., 2020).

Based on the high Nei diversity found in SLC Ec all previous studies proposed SLC Ec and not SLC MA to be the oldest SLC population (Blanca et al., 2012; Lin et al., 2014; Razifard et al., 2020) (Fig. 9 B and C). However, this diversity index was calculated using all haplotypes, and it is a result also found in the current study (Fig. 5 A and B). Moreover, the previous models struggled to explain how the highly diverse SLC Ec could have been derived from the genetically close but less diverse SP Ec, and they all proposed additional secondary contacts between SP and SLC to explain the high SLC Ec Nei diversity. However, this proposal undermined the main evidence used to appoint SLC Ec as the oldest SLC.

There and back

The wide range of latitudes covered by the wild SP and SLC plants suggests that, in the northward migration of SP from Peru to Mexico in which became SLC, there should have been some selection related to latitudinal adaptation. Moreover, some of the adaptations associated with the northern latitudes could have been detrimental in Ecuador and might have been reverted back in the southward trip of SLC to Ecuador. However, given that the SLC southbound migration was too fast for many new haplotypes to arise, we hypothesized that the alleles used in the readaptation of SLC to Ecuadorian latitudes would have been mainly introgressed from Ecuadorian SP. This hypothesis was successfully tested by calculating the ABBA-BABA statistics. The introgressions detected are concentrated in specific genomic region (Fig 6 and Fig S11). Moreover, many of these regions with SP introgresssions have lower expected heterozygosity in SLC MA (Fig 6, A). That result indicates that many regions that suffered selective sweeps in the slow northward migrations recovered the original SP allele in the fast southward migration by introgressing it.

We manually inspected the genes located in these genomic regions (Table S2). Some regions might have been selected to adapt the plants to the new lighting conditions. For example, in the region detected in chromosome 7 there are only 5 genes, and one is a FAR like gene involved in light detection (Xie et al., 2020). In total of the 13 analyzed regions 3 included FAR-like genes and one included a light response gene (Solyc02g021650). Other regions showed flowering related genes: 3 had Agamous-like genes (Solyc06g161130, Solyc05g056620, Solyc04g160300) possibly involved in flowering regulation (Pan et al., 2010), and one had a possible flowering time regulation gene (Imamura et al., 2015). These regions also included a chloroplast development gene (Solyc06g050620) (Pérez-Pérez et al., 2013) and a photosystem protein (Solyc06g009950). Some of these regions are quite large, they include many genes, and it is not

possible to know for sure which gene was selected, because most of the introgressed genes would have just been carried over along with the selected ones. However, the abundance of these kind of biological functions in these regions seems to indicate that they were involved in the latitudinal adaptations. The genetic studies required to check these possibilities, gene by gene, might be accelerated by the public availability of hundreds of F2 populations of many of the accessions involved in the current study with SP, SLC and SP parents (Mata-Nicolás et al., 2020).

This relation between flowering genes and latitudinal adaptation has also been detected in other species. For instance, in potato, the southern wild species eased the introduction of cultivated potatoes to southern latitudes in Chile (Hardigan et al., 2017). Additionally, in rice, Cui et al. (Cui et al., 2020) observed selection of genes related to heading date in the northward expansion of rice cultivation. In wheat and barley, photoperiod sensitivity arose when these crop species emerged from the Fertile Crescent (Meyer and Purugganan, 2013). Unfortunately, this resource does not seem to be available for all species. With respect to pumpkins and gourds, for instance, another American domesticate, were not able to be transferred effectively between latitudes, and these were domesticated independently from different species in different regions (Kates et al., 2017).

Therefore, the original wide range of latitudes covered by wild tomato plants might have created a wealth of allelic diversity in key flowering time-related genes that could be actively used by breeders to adapt tomato varieties to different latitudes worldwide. This might prove to be a key characteristic in the adaptation to climate change in the near future.

Two-step domestication

Both Blanca et *al*. (Blanca et al., 2015) and Razifard et *al*. (Razifard et al., 2020) proposed a two-step domestication evolutionary model (Fig 9. B and C). According to these models SLC would

have been domesticated in northern Peru and then moved to Mesoamerica, where, finally, it would have been improved and transformed into SLL. The analyses and evidences shown in the current study are still in agreement with this two-step domestication. Mexican SLL is closely related to Peruvian SLC. SLL Mx has few novel alleles and shared most of its haplotypes with SLC Pe, with some being SP Pe introgressions. Therefore, SLL Mx could be the result of improvements of local plants originally imported from northern Peru (Fig. 7). This close relationship between SLC Pe and SLL Mx constitutes indirect evidence in favor of an Andean domestication. If there would have been semidomesticated SLC in Mexico, Mexican growers would have probably derived SLL from it and not from imported Peruvian SLC.

Moreover, northern SLC Pe and SLC Ec, which, according to the proposed evolutionary model, are related to the oldest cultivated populations, were collected mainly from backyards, whereas in Mesoamerica, SLC was found to be mainly wild and ruderal. Additionally, according to the morphological PCA (Fig. 8 A) the large-fruited SLC morphological type, typical of northern Peruvian SLC, was very close to Mexican SLL, whereas the small-fruited SLC type, typically found in SLC MA, was located between SP Ec and the large-fruited SLC. Thus, the sequence suggested by this morphological analysis would be as follows: SP Ec, small SLC (typical of Mesoamerican SLC), large-fruited SLC (typically found in northern Peru) and, finally, SLL. Therefore, there is a match between the evolutionary model suggested by both morphological and genetic evidence. Most traits, such as style exertion, petal position and width, and fruit shape varied monotonically along this sequence (Fig. 8 C). Assuming this progression, leaf type and margin would have acquired its typical cultivated forms already in SLC MA; however, petal width would have decreased until its minimum was reached in SLC Pe N, and this population would be also the first one without folded back petals. Inflorescences would have become gradually more irregular starting with SLC MA and reaching a maximum in SLL Mx, a trend

shared by the flat and ribbed fruits that would have appeared in SLC MA and in SLC Pe N, respectively. Razifard et al. noticed this same pattern of small-fruited SLC in Mesoamerica, but since they thought Peruvian and Ecuadorian SLC to be older than Mesoamerican SLC, they proposed a reduction in fruit size in the migration out of the Andean region and a redomestication in Mesoamerica (Razifard et al., 2020). The model presented in the current study proposes a smoother domestication process (Fig. 8), and it explains why the same domesticated alleles were found in Ecuador, Peru and Mesoamerica (Blanca et al., 2015).

Conclusions

The new analysis method based on Procrustes and automatic haplotype classification has allowed us to propose a new hypothesis for the complex evolution of wild and cultivated tomato plants. The wild populations would be Peruvian and Ecuadorian SP and Mesoamerican SLC. SLC after migrating back to Ecuador and Peru was domesticated and the Mexican SLL would have been derived from these improved materials. This model is backed by traditional population genetic indexes, parametric statistical models, morphological and passport data and the new haplotypic analysis. We hope that similar approaches could be used to study the complex domestication histories of other species. Finally, we have identified genomic regions associated with the latitudinal migration experienced by tomato plants that could be useful to adapt the currently cultivated varieties to new latitudes, a trait particularly useful in a world affected by climate change.

Materials and Methods

Accessions and sequences

A total of 628 sequenced accessions, obtained from the Varitome project (Gao et al., 2019) and six other previous studies (Causse et al., 2013; Lin et al., 2014; Sato et al., 2012; Strickler et al., 2015; Zhu et al., 2018), were included in the study. The sequences were obtained specifically from all SP, SLC and SLL resequencing experiments publicly available in the Sol Genomics database (solgenomics.net) and the NCBI Sequence Read Archive as well as 3 new sequenced accessions of Ecuadorian origin. The passport data and mean coverages are listed in Table S3.

Mapping and SNPs

All reads were mapped with BWA-MEM (Li, 2013) against the latest available tomato reference genome (v.4.0) (Hosmani et al., 2019). After mapping, duplicate reads marked by Picard Tools (https://broadinstitute.github.io/picard/), and the reads with a mapping quality lower than 57 were removed. For variant calling, the first and last 3 bases of every mapped read were ignored. SNP calling was carried out by FreeBayes, with a minimum coverage of 10, a minimum alternative allele count of 2, an alternative allele frequency of 0.1, and a maximum number of searched alleles of 5 (Garrison and Marth, 2012).

After SNP calling, genotypes with a coverage lower than 5 were set to missing, and variants that had a calling rate lower than 0.6, that had an observed heterozygosity higher than 0.1 or that were located on chromosome 0 were removed. Afterwards, accessions with a calling rate lower than 0.85 were removed from the analysis. Of the 25.3 million variants initially generated by FreeBayes, 11.8 million were kept after filtering out the low-quality variants and 2.02 million

were present in the euchromatic regions. The variant filtering code is available in create_tier1_snps_excluding_low_qual_samples_and_chrom0.py.

Heterochromatic region determination was based on the recombination rate, and this rate was determined by interpolating data of the publicly available SolCAP genetic map constructed with 3,503 genotyped markers in the EXPEN 2000 F2 population (Sim et al., 2012). A region was considered heterochromatic when its physical distance-to-genetic distance ratio was lower than 1e-6.

We included all the code used to perform the analysis, from variant filtering to figure creation, in the supplementary materials and in a public GitHub repository (https://github.com/bioinfcomav/tomato_haplotype_paper). The code developed was thoroughly tested and the tests are also available in GitHub and can be run to check the correctness of the implementation. The tests for the indexes calculated were also checked with standard population genetic software (Peakall and Smouse, 2012).

For most haplotypic analyses, the highly correlated heterochromatic regions were ignored, and to speed up the computations, only variants with at most 10% missing genotypes were used, resulting in a working set of 33,790 variants. These haplotype analyses the variants were phased and imputed using Beagle (Browning et al., 2018), and the relevant script used was phase_and_impute_with_beagle.py.

Aligned haplotypic PCoAs and automatic haplotypic classification

The euchromatic regions were divided in segments. For each segment that had at least 20 markers, a PCoA was carried out using the haplotypic alleles (two alleles per diploid individual)

reconstructed after imputing. First, a pairwise distance matrix was constructed by calculating the edit distances. PCoAs were done according to the methods of Krzanowski (Krzanowski, 2000). PCoAs from different genomic segments were then aligned using the SciPy orthogonal Procrustes function (Krzanowski, 2000). The Procrustes algorithm takes two sets of points in a space and then calculates and applies a linear transformation (e.g., rotates, translates, reflects) to one of them to align the second set of points as best as possible to the first set. In this case, all the PCoA results were aligned to obtain the final set of aligned PCoA data. The code that implements this functionality is located haplo_auto_classification.py, haplo_pca.py, and procrustes.py.

Once all PCoA data were aligned, the haplotypes were automatically classified by an unsupervised classification algorithm. Prior to classification, the outlier haplotypes were removed by using the Isolation Forest algorithm implemented by the sklearn Python library, with the contamination parameter set to 0.070 (Liu et al., 2012). Due to memory allocation limitations, outlier detection and unsupervised classification could not be carried out with the whole aligned PCoA haplotype matrix. This matrix had more than half a million rows, and these algorithms require a memory allocation that grows geometrically with the number of rows. To solve this problem, thinned input matrices were inputted into the algorithms. The thinned matrices were constructed by calculating the euclidean distance between haplotypes, and when several haplotypes were closer than 0.0015 only one was retained in the thinned matrix. The automatic classification depended only on the haplotype location in the PCoA, so once the classification was completed, all haplotypes that were close to the one present in the thinned matrix shared that same classification. We are aware that excessive thinning could alter the haplotype cluster density in the aligned PCoA data, which would affect the unsupervised classification. Therefore, care was taken to keep the thinning to a minimum. The thinning distance chosen was the minimum

distance that created a matrix capable of being held in the computer's memory. In total, 45,982 of the original 526,240 haplotypes were ultimately present in the thinned matrix.

Haplotypic unsupervised classification was carried out in two steps. First, the haplotypes were classified by using the agglomerative algorithm implemented by scikit-learn (Pedregosa et al., 2011). This agglomerative approach has one limitation—it forces a classification of each haplotype. To solve this problem, the automatic classification was refined in a second step by the KNeighbors supervised classification algorithm that uses the classification generated by the agglomerative step as an input. KNeighbors was configured to use 30 neighbors (Pedregosa et al., 2011).

Population structure

The accessions were classified into populations by taking into account their taxonomic and collection origin passport data as well as a series of variant-based (not haplotypic) hierarchical PCoAs. A population was defined when a set of contiguous accessions in a PCoA were collected from the same geographic area or were classified in the same taxon and shared a similar haplotype composition inferred from the previously calculated haplotypic-based PCoAs. The variant-based PCoAs were carried out only with variants with a major allele frequency lower than 0.95 and a missing genotype rate lower than 0.1 and present in the euchromatin. To avoid overrepresenting any genomic regions, the euchromatin was divided into segments of 100 kb, and from each segment, only the most variable SNP or indel was retained. With the resulting variants, the Kosman distances were calculated (Kosman and Leonard, 2005), and PCoA was carried out (Krzanowski, 2000). The relevant scripts used include both pcas_do.py and pca.py.

FastSTRUCTURE was also used to infer the population composition (Raj et al., 2014). This method was run with the Ks value ranging from 2 to 11 and with the default settings. For this analysis, only the variants that had a maximum major allele frequency of 0.95 and were located in the euchromatic regions were used. The code used is embedded in fastructure_run.py.

Parametric population history reconstruction

We used TreeMix to reconstruct parametric population history using variants characterized by a low linkage disequilibrium (Pickrell and Pritchard, 2012). Thus, the variants were selected by generating haplotype blocks of consecutive SNPs in the genome that allowed only a certain maximum correlation threshold between its genotypes. From every haplotype block, only one SNP was selected at random to be used by TreeMix. Branch support was calculated by bootstrapping, and in each bootstrap iteration, random variants were taken from every linked genomic block. Different TreeMix analyses with different linkage disequilibrium thresholds were performed to test the robustness of the results. The code can be found within the treemix.py file.

Population diversity and LD

To evaluate the population diversity, various parameters were calculated, such as the number of polymorphic variants (95% threshold) and unbiased Nei diversity (Nei and Roychoudhury, 1974). Additionally, to reflect the haplotypic diversity, several indexes were evaluated calculated for each 500 kb genome segment: the mean number of different haplotypes or the mean number of variants found in a genomic segment.

Some of these parameters could be potentially affected by the number of individuals available in each population. This potential pitfall was accounted for by the following two strategies. In the first strategy, diversities were calculated with the same number of accessions for each population. This number of accessions was chosen to be 75% of the number of accessions of the population

with the fewest number of accessions in the analysis. The indexes were then calculated 100 times, with different accessions chosen at random for each population in each iteration. When all 100 values were obtained, the mean and the confidence interval of the mean were used to represent the diversity of each population. In the second approach, a complementary rarefaction analysis, in which accessions were added one by one to each population, was carried out. The script files used to perform these calculations were diversities_vars.py and diversities_haplos.py.

LD was calculated by using only the polymorphic variants (95% threshold). LDs between markers that were up to 1000 kb apart were calculated, and LD decay was estimated with locally weighted scatterplot smoothing (LOWESS) implemented by the statsmodels Python library (www.statsmodels.org). Finally, the LD at 10 kb was obtained from the adjusted curve. The relevant code is located in ld.py, ld_decay_plot.py and ld_bar_plot.py.

Introgression detection and functional analysis

ABBA-BABA indexes were calculated following (Green et al., 2010) using the euchromatic variants. Additionally, the mean values per genome region were obtained and plotted to look for introgressions in Ecuadorian and Peruvian SLC. These genomic regions were 50 kb long. The relevant scripts are abba.py and there_and_back_abba.py.

Additionally, the alleles potentially introgressed from SP into SL were looked for by using the following method. The alleles in the SLC MA population were considered reference SL alleles, and the alleles found in all other SLC populations were labeled as introgressed when they were not found in SLC MA but were present in a SP population at a frequency higher than 10%. SLC MA was used as a reference because the results showed that this population had the fewest Peruvian and Ecuadorian haplotypes. Once each allele for each variant was labeled as

introgressed or not introgressed, the allele introgression frequency was calculated for each variant. The implementation can be found in the introgressions.py file.

Morphological analysis

Morphological analysis based on a characterization of 375 accessions available at the Tomato Genetic Resource Center (TGRC) and the COMAV genebank was carried out using the accessions for which images were available. Each accession was evaluated for basic inflorescence, leaf, stem and fruit traits (Table S3, Fig. S15).

The morphological classification was carried out via principal component analysis (PCA). Morphological traits were treated as ordinal traits in the analysis. For the PCA, accessions with more than 7 missing values were removed, and any missing data in the remaining accessions were filled with the means. The data for each trait were standardized using the StandardScaler function of the scikit-learn Python library, and PCA was carried out by the PCA functionality implemented in the same library (Pedregosa et al., 2011). The code is located in the morphological.py file. Finally, morphological classification was performed manually by inspecting the morphological PCA data as well as the taxonomic and collection site passport information.

The passport data were manually curated to check the collection source information. The final collection source for each accession was obtained by combining the information stated in the collection source passport field as well as the annotations and images made during the collection expeditions (when available).

References

Álvarez-Hernández, J.C., Cortez-Madrigal, H., García-Ruiz, I., 2009. Exploración y caracterización de poblaciones silvestres de jitomate (Solanaceae) en tres regiones de Michoacán, México. Polibotánica 139–159.

Beresford-Jones, D.G., Whaley, O., Ledesma, C.A., Cadwallader, L., 2011. Two millennia of changes in human ecology: archaeobotanical and invertebrate records from the lower Ica valley, south coast Peru. Veget Hist Archaeobot 20, 273. https://doi.org/10.1007/s00334-011-0292-4

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. https://doi.org/10.1371/journal.pone.0048198

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M.J., Francis, D., Causse, M., van der Knaap, E., Cañizares, J., 2015. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics 16, 257. https://doi.org/10.1186/s12864-015-1444-1

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics 103, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Causse, M., Desplat, N., Pascual, L., Le Paslier, M.-C., 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, 791. https://doi.org/10.1186/1471-2164-14-791

Cui, Y., Wang, J., Feng, L., Liu, S., Li, J., Qiao, W., Song, Y., Zhang, Z., Cheng, Y., Zhang, L., Zheng, X., Yang, Q., 2020. A Combination of Long-Day Suppressor Genes Contributes to the Northward Expansion of Rice. Front Plant Sci 11, 864. https://doi.org/10.3389/fpls.2020.00864

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., Thannhauser, T.W., Foolad, M.R., Diez, M.J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H.J., Giovannoni, J.J., Fei, Z., 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature Genetics 51, 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio].

Gibson, M.J.S., Moyle, L.C., 2020. Regional differences in the abiotic environment contribute to genomic divergence within a wild tomato species. Molecular Ecology 29, 2204–2217. https://doi.org/10.1111/mec.15477

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722. https://doi.org/10.1126/science.1188021

Grobman, A., Bonavia, D., Dillehay, T.D., Piperno, D.R., Iriarte, J., Holst, I., 2012. Preceramic maize from Paredones and Huaca Prieta, Peru. PNAS 109, 1755–1759. https://doi.org/10.1073/pnas.1120270109

Hammer, K., 1984. Das Domestikationssyndrom. Die Kulturpflanze 32, 11–34. https://doi.org/10.1007/BF02098682

Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J.C., Douches, D.S., Farré, E.M., Veilleux, R.E., Buell, C.R., 2017. Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. PNAS 114, E9999–E10008. https://doi.org/10.1073/pnas.1714380114

Hosmani, P.S., Flores-Gonzalez, M., Geest, H. van de, Maumus, F., Bakker, L.V., Schijlen, E., Haarst, J. van, 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 767764. https://doi.org/10.1101/767764

Imamura, T., Fujita, K., Tasaki, K., Higuchi, A., Takahashi, H., 2015. Characterization of spermidine synthase and spermine synthase--The polyamine-synthetic enzymes that induce early flowering in Gentiana triflora. Biochem Biophys Res Commun 463, 781–786. https://doi.org/10.1016/j.bbrc.2015.06.013

Iris Peralta, David Spooner, 2011. History, Origin and Early Cultivation of Tomato (Solanaceae), in: Maharaj K. Razdan, K. Mattoo (Eds.), Genetic Improvement of Solanaceous Crops. CRC Press, Boca Raton.

Kates, H.R., Soltis, P.S., Soltis, D.E., 2017. Evolutionary and domestication history of Cucurbita (pumpkin and squash) species inferred from 44 nuclear loci. Molecular Phylogenetics and Evolution 111, 98–109. https://doi.org/10.1016/j.ympev.2017.03.002

Kiple, K.F., Ornelas, K.C., Press, C.U., 2000. The Cambridge World History of Food. Cambridge University Press.

Kosman, E., Leonard, K.J., 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. Molecular Ecology 14, 415–424. https://doi.org/10.1111/j.1365-294X.2005.02416.x

Krzanowski, W., 2000. Principles of Multivariate Analysis: A User's Perspective: 23. Oxford Oxfordshire□; New York.

Li, G., Siddiqui, H., Teng, Y., Lin, R., Wan, X., Li, J., Lau, O.-S., Ouyang, X., Dai, M., Wan, J., Devlin, P.F., Deng, X.W., Wang, H., 2011. Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis. Nature Cell Biology 13, 616–622. https://doi.org/10.1038/ncb2219

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio].

Lieberman, M., Segev, O., Gilboa, N., Lalazar, A., Levin, I., 2004. The tomato homolog of the gene encoding UV-damaged DNA binding protein 1 (DDB1) underlined as the gene that causes the high pigment-1 mutant phenotype. Theor Appl Genet 108, 1574–1581. https://doi.org/10.1007/s00122-004-1584-1

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, Junming, Zhang, C., Wang, T., Zhang, Yuyang, Wang, A., Zhang, Yancong, Lin, K., Li, C., Xiong, G., Xue, Y., Mazzucato, A., Causse, M., Fei, Z., Giovannoni, J.J., Chetelat, R.T., Zamir, D., Städler, T., Li, Jingfu, Ye, Z., Du, Y., Huang, S., 2014. Genomic analyses provide insights into the history of tomato breeding. Nature Genetics 46, 1220–1226. https://doi.org/10.1038/ng.3117

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-Based Anomaly Detection. ACM Trans. Knowl. Discov. Data 6, 3:1-3:39. https://doi.org/10.1145/2133360.2133363

Mata-Nicolás, E., Montero-Pau, J., Gimeno-Paez, E., Garcia-Carpintero, V., Ziarsolo, P., Menda, N., Mueller, L.A., Blanca, J., Cañizares, J., van der Knaap, E., Díez, M.J., 2020. Exploiting the diversity of tomato: the development of a phenotypically and genetically detailed germplasm collection. Hortic Res 7. https://doi.org/10.1038/s41438-020-0291-7

Meyer, R.S., Purugganan, M.D., 2013. Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 14, 840–852. https://doi.org/10.1038/nrg3605

Nei, M., Roychoudhury, A.K., 1974. Sampling Variances of Heterozygosity and Genetic Distance. Genetics 76, 379–390.

Pan, I.L., McQuinn, R., Giovannoni, J.J., Irish, V.F., 2010. Functional diversification of AGAMOUS lineage genes in regulating tomato flower and fruit development. J Exp Bot 61, 1795–1806. https://doi.org/10.1093/jxb/erq046

Peakall, R., Smouse, P.E., 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics 28, 2537–2539. https://doi.org/10.1093/bioinformatics/bts460

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peralta, I.E., Spooner, D.M., Knapp, S., 2008. Taxonomy of Wild Tomatoes and Their Relatives (Solanum Sect. Lycopersicoides, Sect. Juglandifolia, Sect. Lycopersicon; Solanaceae). American Society of Plant Taxonomists.

Pérez-Pérez, J.M., Esteve-Bruna, D., González-Bayón, R., Kangasjärvi, S., Caldana, C., Hannah, M.A., Willmitzer, L., Ponce, M.R., Micol, J.L., 2013. Functional Redundancy and Divergence within the Arabidopsis RETICULATA-RELATED Gene Family1[W][OA]. Plant Physiol 162, 589–603. https://doi.org/10.1104/pp.113.217323

Pickersgill, B., 2016. Domestication of Plants in Mesoamerica: An Archaeological Review with Some Ethnobotanical Interpretations, in: Lira, R., Casas, A., Blancas, J. (Eds.), Ethnobotany of Mexico: Interactions of People and Plants in Mesoamerica, Ethnobiology. Springer, New York, NY, pp. 207–231. https://doi.org/10.1007/978-1-4614-6669-7_9

Pickrell, J.K., Pritchard, J.K., 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genetics 8, e1002967. https://doi.org/10.1371/journal.pgen.1002967

Raj, A., Stephens, M., Pritchard, J.K., 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 197, 573–589. https://doi.org/10.1534/genetics.114.164350

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. Molecular Biology and Evolution 37, 1118–1132. https://doi.org/10.1093/molbev/msz297

Reinhard, K.J., Chaves, S.M., Jones, J.G., Iñiguez, A.M., 2008. Evaluating chloroplast DNA in prehistoric Texas coprolites: medicinal, dietary, or ambient ancient DNA? Journal of Archaeological Science 35, 1748–1755. https://doi.org/10.1016/j.jas.2007.11.013

Rick, C.M., Fobes, J.F., 1975. Allozyme Variation in the Cultivated Tomato and Closely Related Species. Bulletin of the Torrey Botanical Club 102, 376–384. https://doi.org/10.2307/2484764

Rick, C.M., Holle, M., 1990. Andean lycopersicon esculentum var. cerasiforme: genetic variation and its evolutionary significance. Econ Bot 44, 69. https://doi.org/10.1007/BF02860476

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, Changbao, Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., Ren, Y., Wang, J., Albiero, A., Dal Pero, F., Todesco, S., Van Eck, J., Buels, R.M., Bombarely, A., Gosselin, J.R., Huang, M., Leto, J.A., Menda, N., Strickler, S., Mao, L., Gao, S., Tecle, I.Y., York, T., Zheng, Y., Vrebalov, J.T., Lee, J., Zhong, S., Mueller, L.A., Stiekema, W.J., Ribeca, P., Alioto, T., Yang, W., Huang, Sanwen, Du, Y., Zhang, Z., Gao, Jianchang, Guo, Y., Wang, X., Li, Y., He, J., Li, Chuanyou, Cheng, Z., Zuo, J., Ren, J., Zhao, J., Yan, L., Jiang, H., Wang, B., Li, H., Li, Z., Fu, F., Chen, B., Han, B., Feng, Q., Fan, D., Wang, Y., Ling, H., Xue, Y., Ware, D., Richard McCombie, W., Lippman, Z.B., Chia, J.-M., Jiang, K., Pasternak, S., Gelley, L., Kramer, M., Anderson, L.K., Chang, S.-B., Royer, S.M., Shearer, L.A., Stack, S.M., Rose, J.K.C.,

Xu, Y., Eannetta, N., Matas, A.J., McQuinn, R., Tanksley, S.D., Camara, F., Guigó, R., Rombauts, S., Fawcett, J., Van de Peer, Y., Zamir, D., Liang, C., Spannagl, M., Gundlach, H., Bruggmann, R., Mayer, K., Jia, Z., Zhang, J., Ye, Z., Bishop, G.J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, R., Singh, M., Singh, A., Kumar Pal, J., Pandit, A., Kumar Singh, P., Kumar Mahato, A., Dogra, V., Gaikwad, K., Raj Sharma, T., Mohapatra, T., Kumar Singh, N., Causse, M., Rothan, C., Schiex, T., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P., Vautrin, S., Zouine, M., Latché, A., Rousseau, C., Regad, F., Pech, J.-C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., Fernandez del Carmen, A., Monforte, A., Granell, A., Fernandez-Muñoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S., Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perrotta, G., Daddiego, L., Bryan, G., Orozco, M., Pastor, X., Torrents, D., van Schriek, M.G.M., Feron, R.M.C., van Oeveren, J., de Heer, P., daPonte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., van Eijk, M.J.T., Janssen, A., van Haaren, M.J.J., Jo, S.-H., Kim, J., Kwon, S.-Y., Kim, Sangmi, Koo, D.-H., Lee, S., Hur, C.-G., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jöcker, A., Warfsmann, J., Göbel, U., Kawamura, S., Yano, K., Sherman, J.D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E.G.W.M., van de Belt, J., van Haarst, J.C., Peters, S.A., van Staveren, M.J., Henkens, M.H.C., Mooyman, P.J.W., Hesselink, T., van Ham, R.C.H.J., Jiang, G., Droege, M., Choi, D., Kang, B.-C., Dong Kim, B., Park, M., Kim, Seungill, Yeom, S.-I., Lee, Y.-H., Choi, Y.-D., Li, G., Gao, Jianwei, Liu, Y., Huang, Shengxiong, Fernandez-Pedrosa, V., Collado, C., Zuñiga, S., Wang, G., Cade, R., Dietrich, R.A., Rogers, J., Knapp, S., Fei, Z., White, R.A., Thannhauser, T.W., Giovannoni, J.J., Angel Botella, M., Gilbert, L., Gonzalez, R., The Tomato Genome Consortium, Kazusa DNA Research Institute, 454 Life Sciences, a R. company, Amplicon Express Inc., Beijing Academy of Agriculture and Forestry Sciences, BGI-Shenzhen, BMR-Genomics SrL, Boyce Thompson Institute for Plant Research, Centre for BioSystems Genomics, Centro Nacional de Análisis Genómico (CNAG), China Agricultural University, Chinese Academy of Agricultural Sciences, Chinese Academy of Sciences, Cold Spring Harbor Laboratory and United States Department of Agriculture – Agricultural Research Service, Colorado State University, Cornell University, Genome Bioinformatics Laboratory GRIB–IMIM/UPF/CRG, Ghent University-VIB, Hebrew University of Jerusalem, Heilongjiang Academy of Agricultural Sciences, Helmholtz Center for Health and Environment, Henan Agricultural University, Huazhong Agricultural University, Imperial College London, Indian Agricultural Research Institute, INRA Avignon, INRA Bordeaux, INRA Toulouse, Institut National Polytechnique de Toulouse, Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM-UMA-CSIC), Instituto Nacional de Tecnología Agropecuaría (IB-INTA) and Consejo Nacionalde Investigaciones Científicas y Técnicas (CONICET), Italian National Res Council, I. for B.T., Italian National Research Council, I. of P.G., Research Division Portici, Italian National Agency for New technologies, E. and S.D., James Hutton Institute, Joint IRB-BSC program on Computational Biology, Keygene N.V., Korea Research Institute of Bioscience and Biotechnology, Life Technologies, Max Planck Institute for Plant Breeding Research, Meiji University, Montana State University, NARO Institute of Vegetable and Tea Science, National Institute of Plant Genome Research, Plant Research International, Qingdao Agricultural University, Roche Applied Science, Seoul National University, Shandong Academy of Agricultural Sciences, Sichuan University, Sistemas Genomicos, South China Agricultural University, Syngenta Biotechnology, The Genome Analysis Centre, The Natural History Museum, United States Department of Agriculture – Agricultural Research Service, R.W.H.C. and B.T.I. for P.R., Universidad de Malaga-Consejo Superior de Investigaciones Cientificas, Universitat Pompeu Fabra, 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641. https://doi.org/10.1038/nature11119

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M.W., Deynze, A.V., Hamilton, J.P., Buell, C.R., Causse, M., Wijeratne, S., Francis, D.M., 2012. Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato. PLOS ONE 7, e40563. https://doi.org/10.1371/journal.pone.0040563

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. https://doi.org/10.7717/peerj.793

Wang, C., Szpiech, Z.A., Degnan, J.H., Jakobsson, M., Pemberton, T.J., Hardy, J.A., Singleton, A.B., Rosenberg, N.A., 2010. Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis. Statistical Applications in Genetics and Molecular Biology 9. https://doi.org/10.2202/1544-6115.1493

Wang, C., Zhan, X., Liang, L., Abecasis, G.R., Lin, X., 2015. Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation. Am J Hum Genet 96, 926–937. https://doi.org/10.1016/j.ajhg.2015.04.018

Xie, Y., Zhou, Q., Zhao, Y., Li, Q., Liu, Y., Ma, M., Wang, B., Shen, R., Zheng, Z., Wang, H., 2020. FHY3 and FAR1 Integrate Light Signals with the miR156-SPL Module-Mediated Aging Pathway to Regulate Arabidopsis Flowering. Molecular Plant, Photobiology 13, 483–498. https://doi.org/10.1016/j.molp.2020.01.013

Yang, Z., Li, G., Tieman, D., Zhu, G., 2019. Genomics Approaches to Domestication Studies of Horticultural Crops. Horticultural Plant Journal, Advance of Horticultural Plant Genomes 5, 240–246. https://doi.org/10.1016/j.hpj.2019.11.001

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., Cao, X., Han, X., Wang, X., Knaap, E. van der, Zhang, Z., Cui, X., Klee, H., Fernie, A.R., Luo, J., Huang, S., 2018. Rewiring of the Fruit Metabolome in Tomato Breeding. Cell 172, 249-261.e12. https://doi.org/10.1016/j.cell.2017.12.019

Zsögön, A., Čermák, T., Naves, E.R., Notini, M.M., Edel, K.H., Weinl, S., Freschi, L., Voytas, D.F., Kudla, J., Peres, L.E.P., 2018. De novo domestication of wild tomato using genome editing. Nature Biotechnology 36, 1211–1216. https://doi.org/10.1038/nbt.4272

Funding

This research was supported by the National Natural Science Foundation of USA Varitome Project (NSF IOS 1564366).

Author contributions:

JB and JC wrote the manuscript and designed the methodology. DM, PZ, JMP, JB and JC performed the statistical analysis. MJD performed the morphological characterizations. JMP, MJD and EK participated in the discussion and manuscript revisions. All the authors have read and approved the final manuscript.

Competing interests: Authors declare that they have no competing interest

Data and materials availability

The new genome reads supporting the conclusions of this article are available in theSRA repository, under bioproject PRJNA702633 (https://www.ncbi.nlm.nih.gov/bioproject/702633)

All the Python code used is available in the Github public repositories: tomato_haplotype_paper and variation5 (https://github.com/bioinfcomav/tomato_haplotype_paper, https://github.com/bioinfcomav/variation5) under the GNU GPL license.

Figures

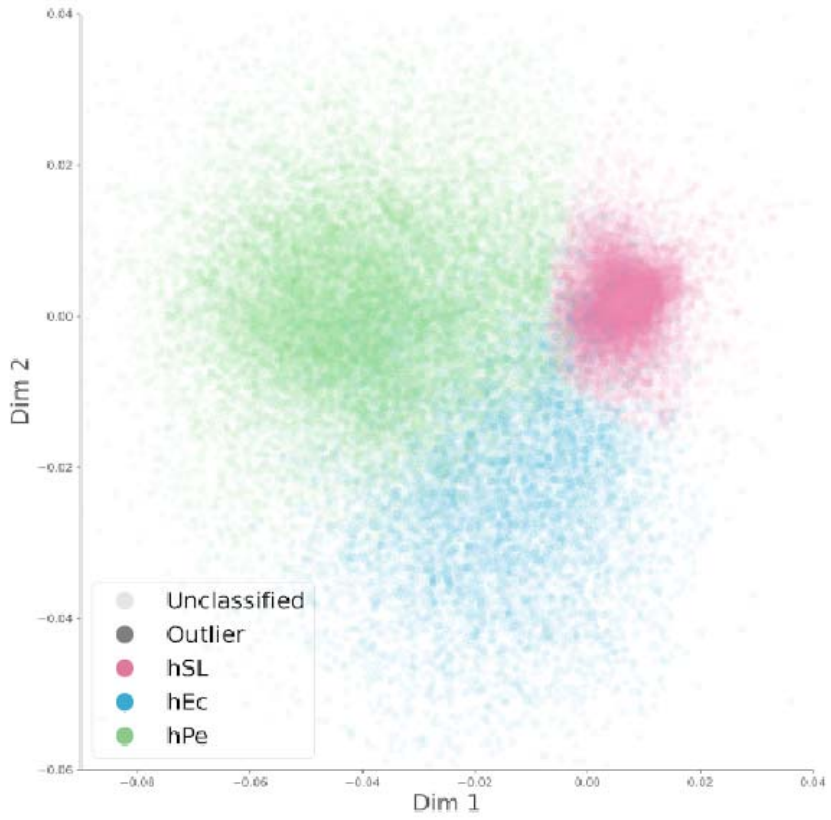

Fig. 1. Haplotypic PCoAs. A PCoA was carried out for every 500 kb genome segment using edit distances between haplotypes. The resulting PCoAs were aligned using Procrustes and automatically classified into three haplotype types.

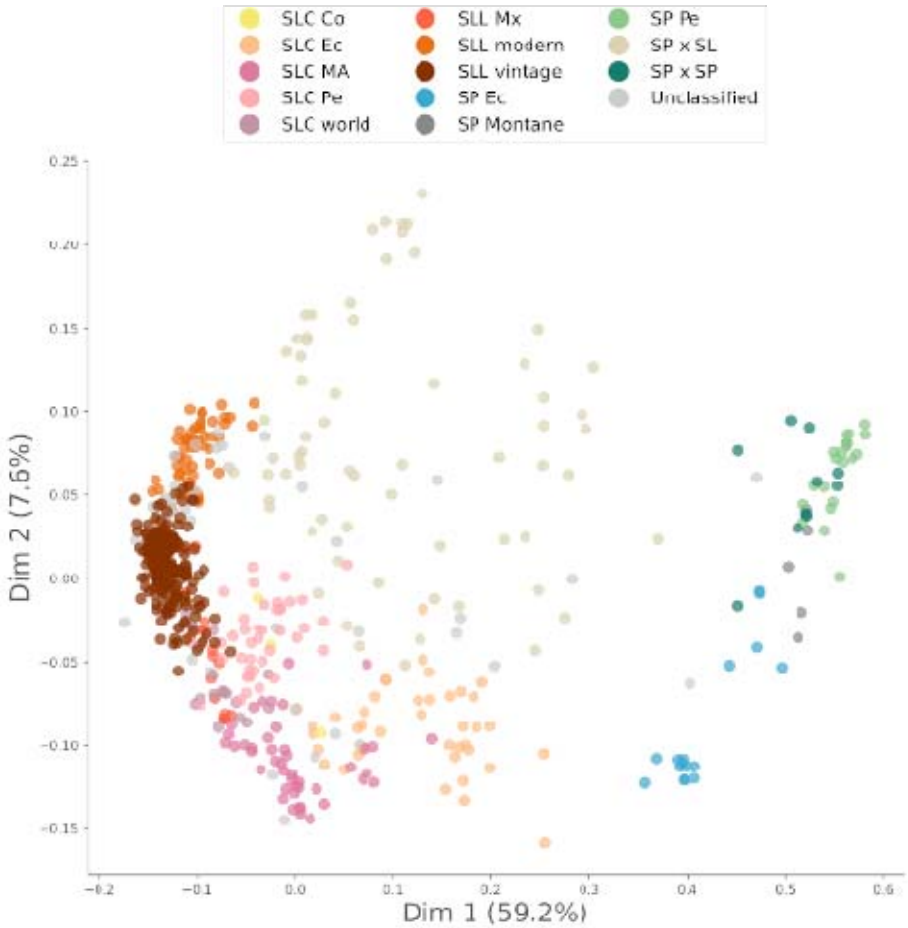

Fig. 2. Accession PCoAs. Pairwise Kosman genetic distances between accessions were calculated using variants characterized by a major allele frequency lower than 0.95 and a missing genotype rate lower than 0.1. Only the most variable variant from every 100 kb genomic segment was used to calculate the distances. The PCoA was based on the obtained pairwise genetic distance matrix. The accessions are colored according to the population in which they were classified.

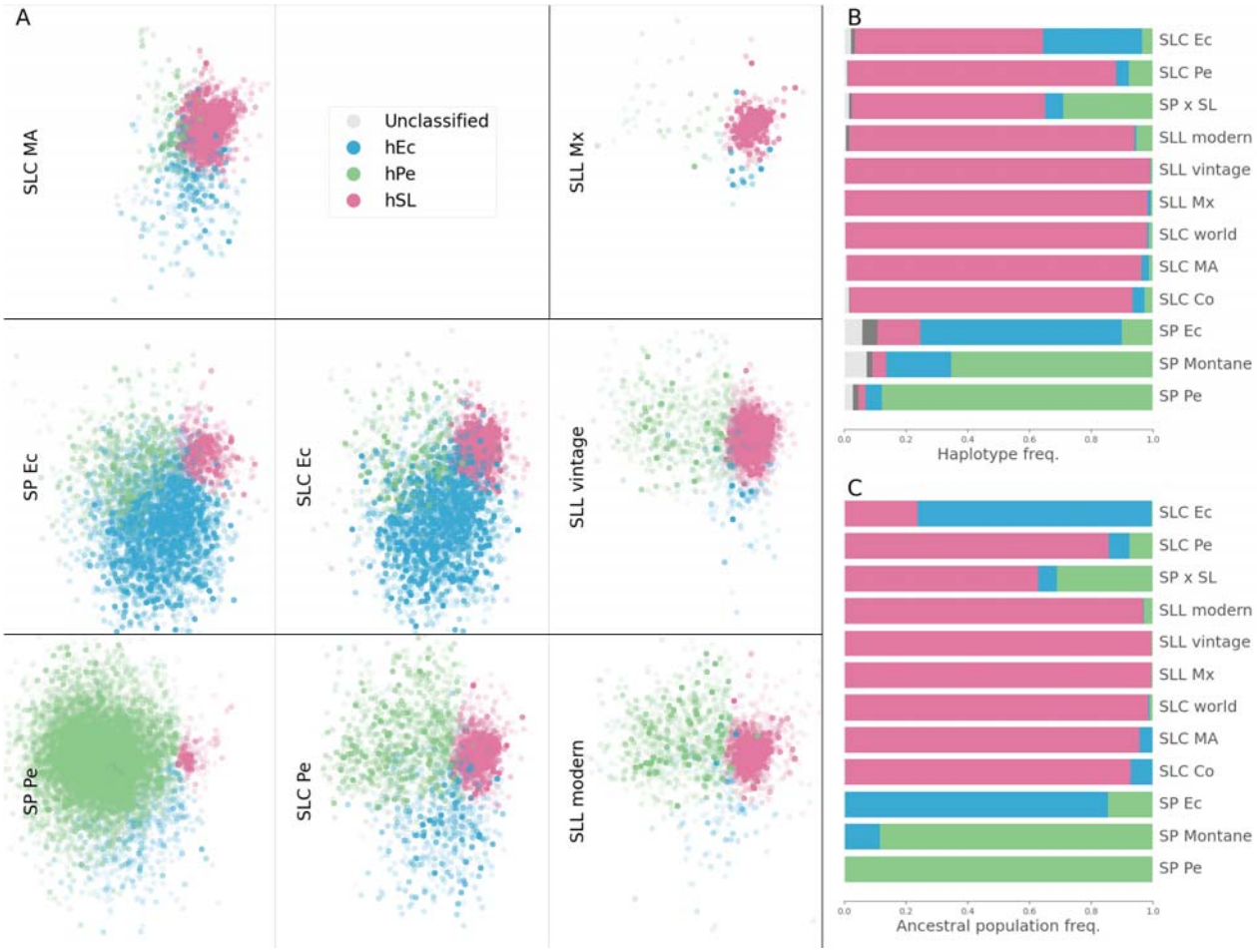

Fig. 3. Population haplotype composition. A) The classified haplotype PCoA obtained for figure 1 was divided into several figures, one per population. In each figure only the haplotypes that belong to each population accessions were included. B) Frequencies of each haplotype type found in each population. C) FastSTRUCTURE ancestral population composition per population.

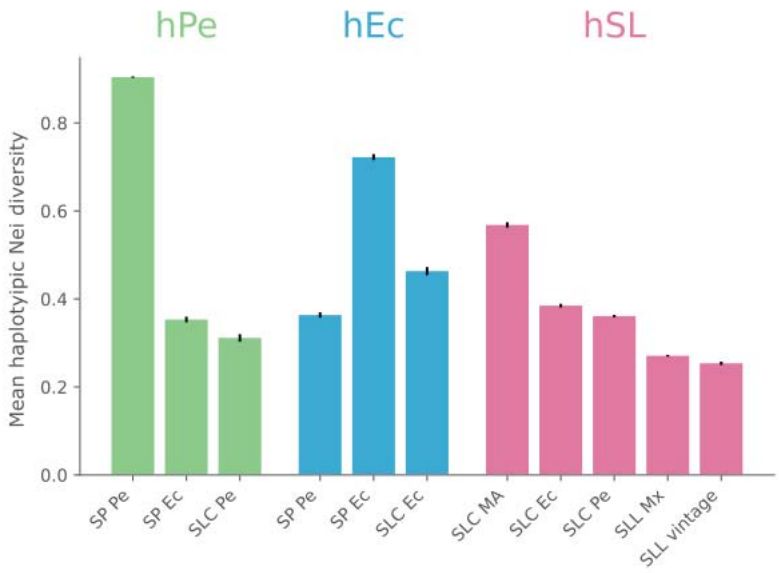

Fig. 4. Mean Nei haplotypic diversity per haplotype type. Euchromatic regions were split into 500 kb segments; for each segment, the haplotypic alleles were determined and classified. The unbiased expected heterozygosity per variant was calculated using only the genotypes that corresponded to the haplotypes classified as: hPe, hPEc, and hSL, and the mean of the expected heterozygosity for the whole genome was calculated. To calculate the indexes, the number of accessions was the same for each population, 75% of the population with less individuals. The analysis was repeated 100 times choosing the accessions representative of each population at random. The bar represents the mean result obtained in these 100 repeats and the error bars are the confidence intervals of the means.

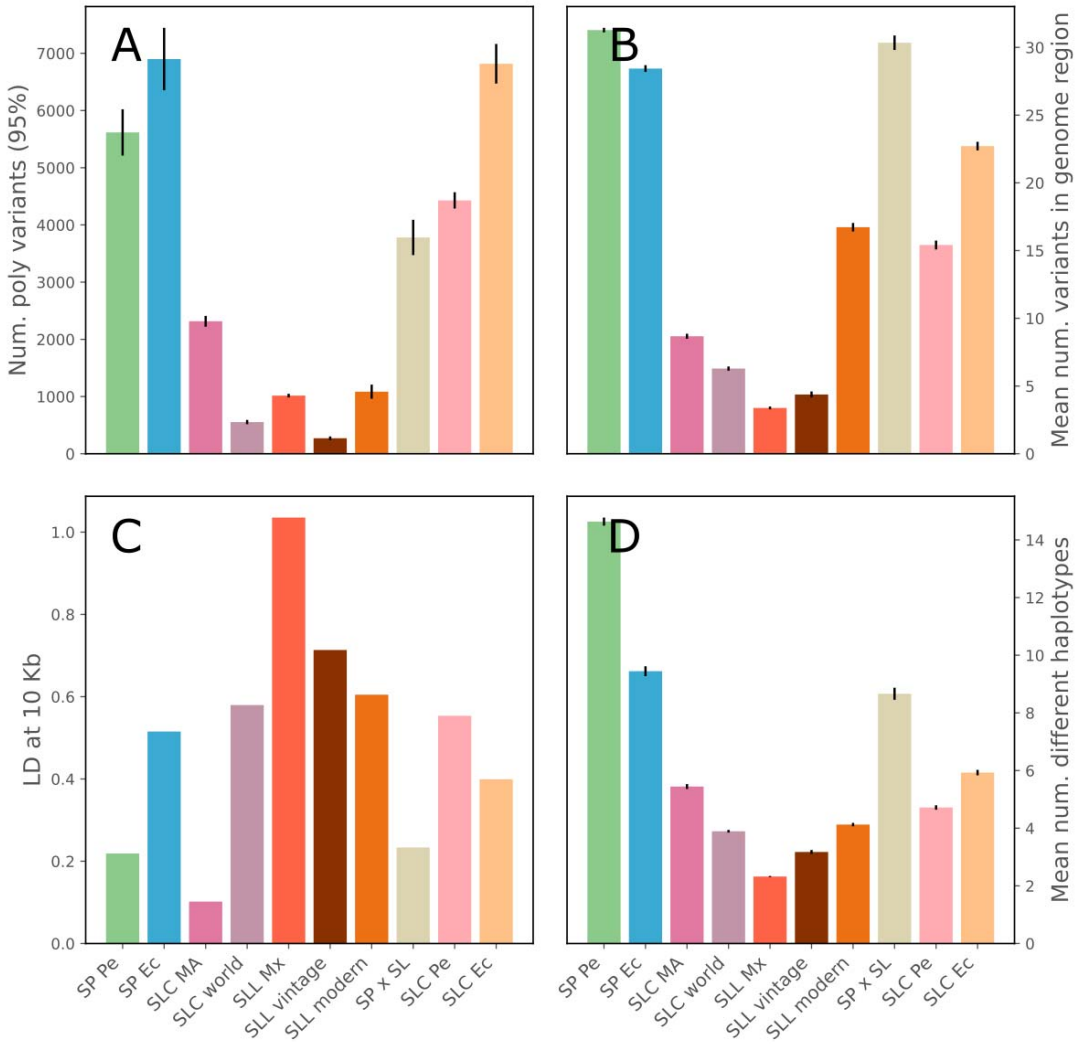

Fig. 5. Diversity and linkage disequilibrium per population. To calculate the indexes, the number of accessions was the same for each population, 75% of the population with less individuals. The analysis was repeated 100 times choosing the accessions representative of each population at random. The bar represents the mean result obtained in these 100 repeats and the error bars are the confidence intervals of the means. A) Number of polymorphic variants (95% threshold). B) Mean number of variants found in a 500 kb euchromatic segment. C) Linkage disequilibrium between variants at 10 kb. D) Mean number of haplotypic alleles (500 kb euchromatic segments).

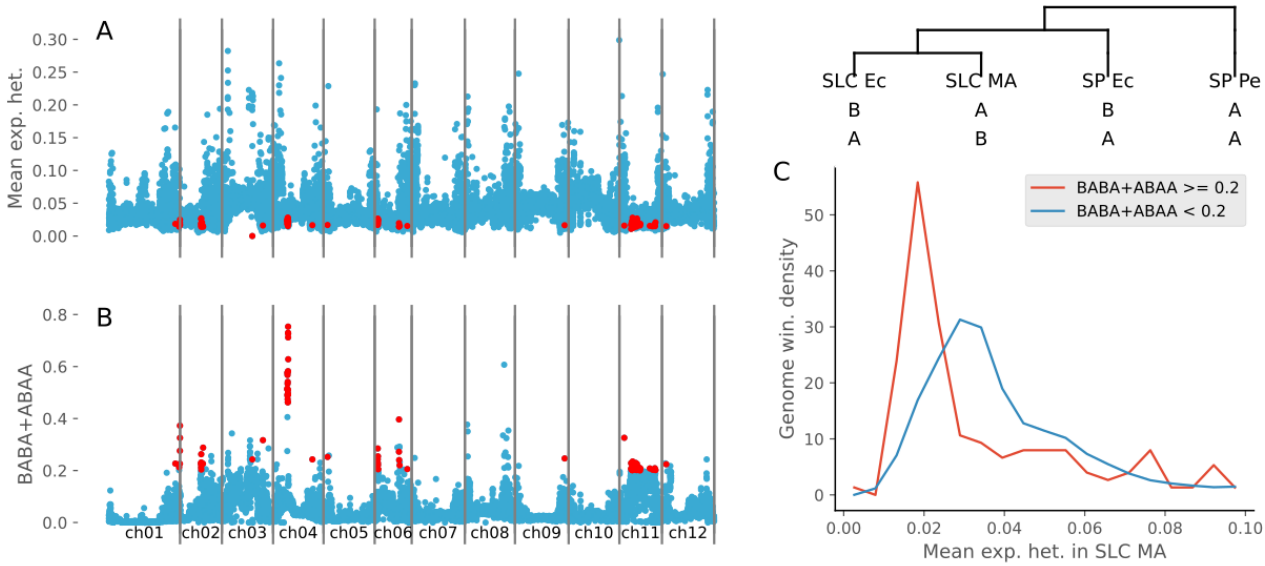

Fig. 6. Mesoamerican Nei diversity and introgression detection in SLC Ec along the genome. A) Mesoamerican SLC expected heterozygosity along the genome in 500 kb segments. B) Sum of BABA and ABAA products calculated in 500 kb segments along the genome using the evolution model: SLC Ec, SLC MA, SP Ec, SP Pe. The genome segments with a BABA + ABAA value higher than 0.2 and an expected heterozygosity lower than 0.03 are represented in red. C) SLC MA expected heterozygosity genome segment distributions with BABA + ABAA lower and higher than 0.2.

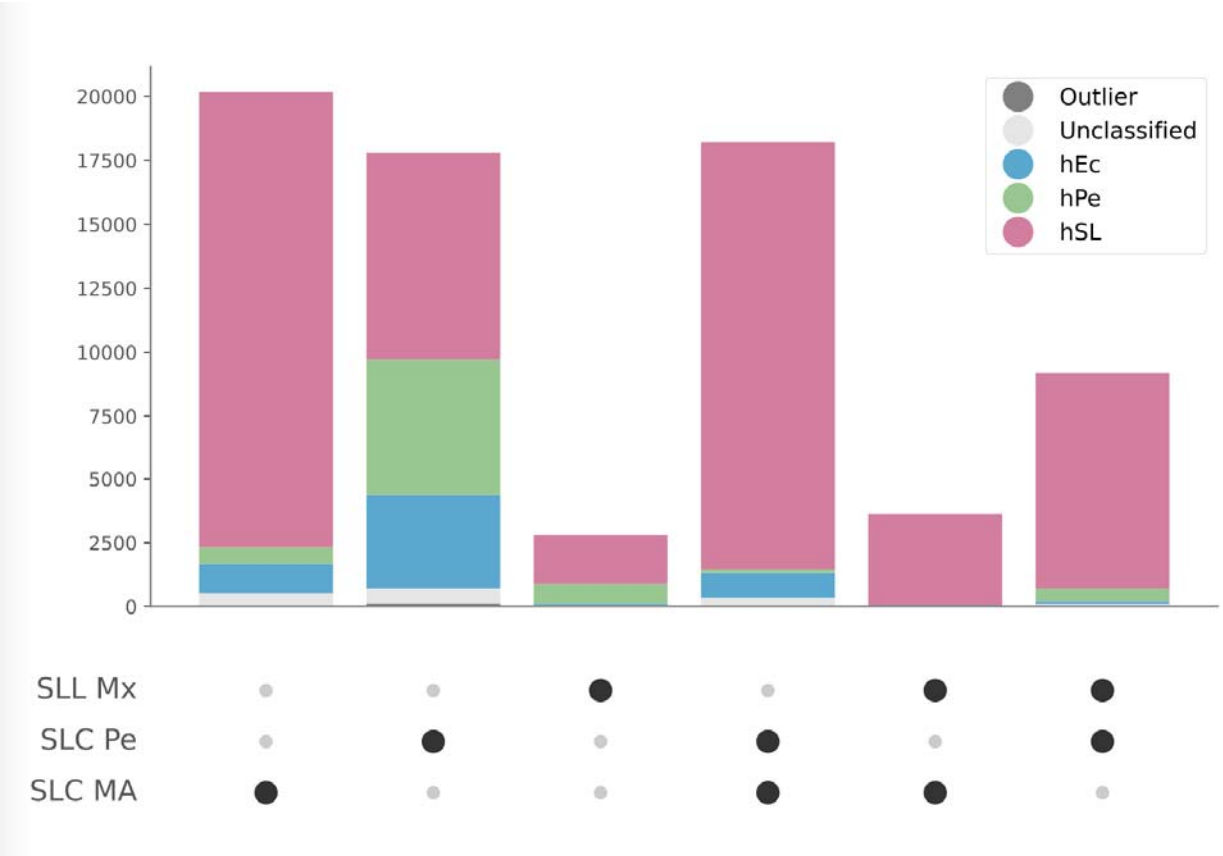

Fig. 7. Shared and private haplotypic alleles among Mesoamerican SLC, Peruvian SLC and Mexican SLL. The bars represent the number of the different haplotypes shared between the populations depicted by the bigger dots in the lower panel.

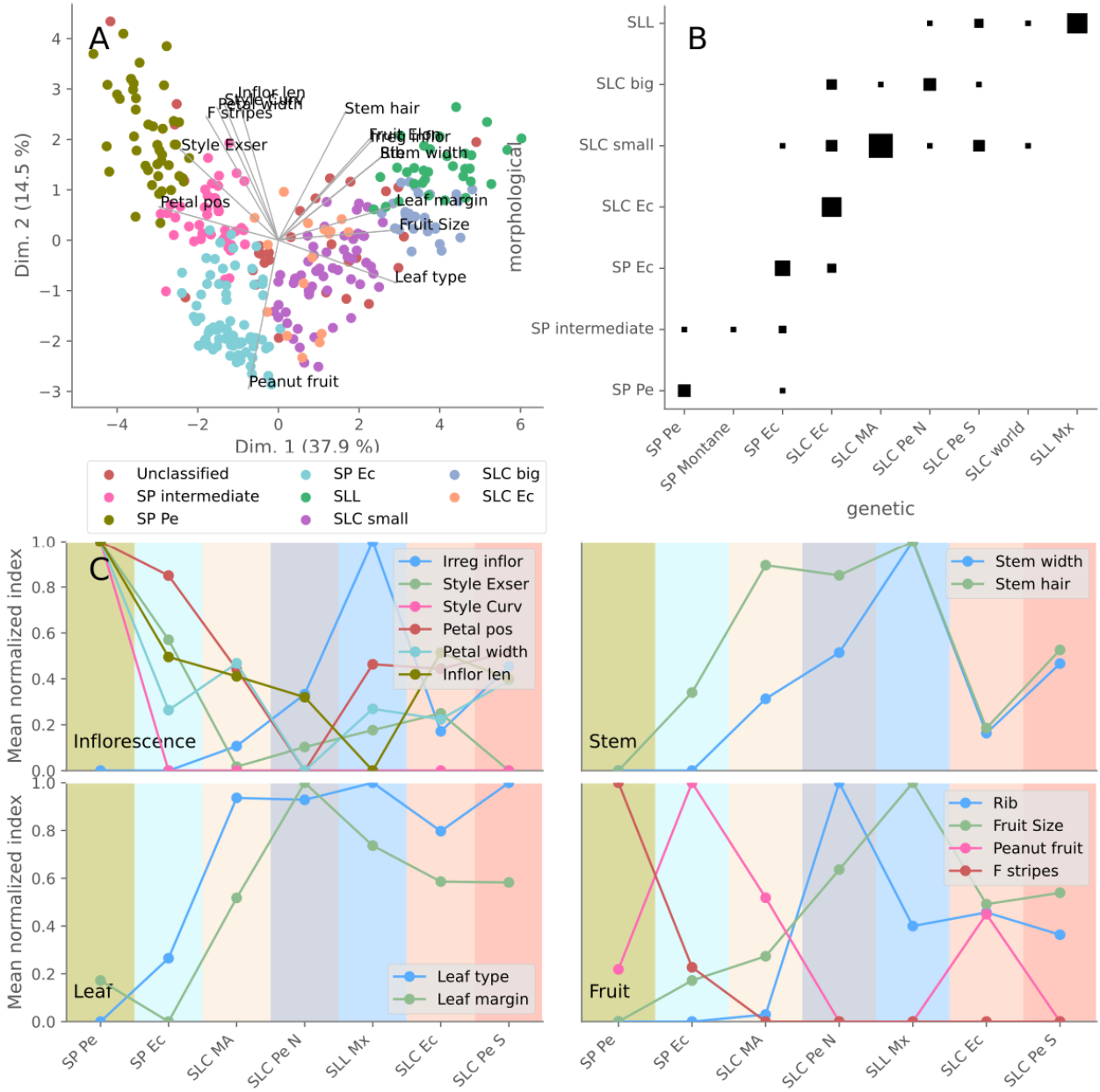

Fig. 8. Morphological analysis. A) Accession based Principal Component Analysis calculated using the morphological traits. The marker colors represent different morphological types. B) Comparisons between morphological types and populations. The marker size represents the number of accessions. C) Morphological trait mean values calculated for different populations.

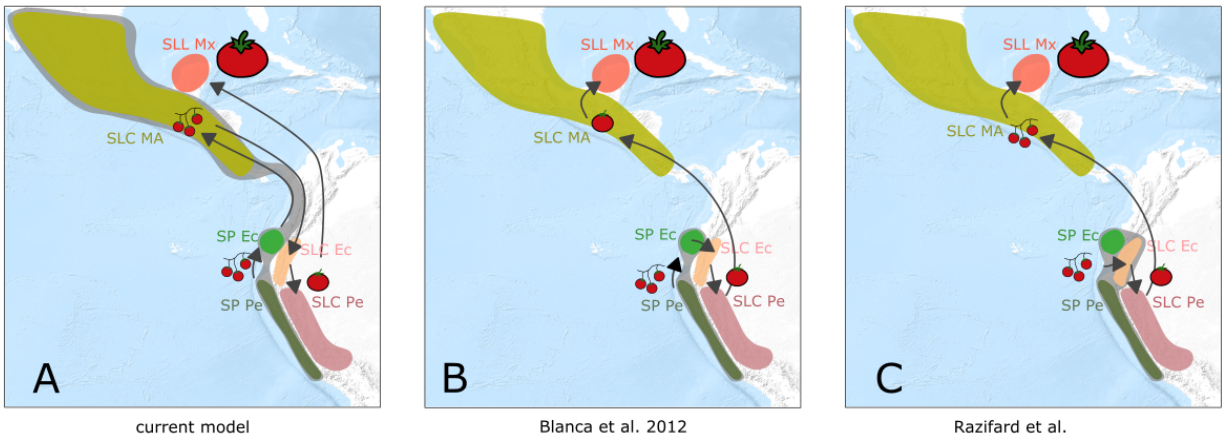

Fig. 9. Tomato evolution hypotheses. Genetic population ranges are represented by the coloring of the different geographical areas. The areas showed in grey include populations supposed to have evolved before any human alteration. Arrows indicate migrations. Three fruit sizes are represented: wild-like small fruits, semi-domesticated forms and SLL fruit types. A) Evolutionary model proposed in the current study. B) Hypothesis proposed by Blanca et al. 2012. C) Hypothesis proposed by Razifard et al. 2020.