# A *Solanum lycopersicoides* reference genome facilitates insights into tomato specialized metabolism and immunity

Adrian F. Powell[1,#], Ari Feder[1,#,§], Jie Li[2,#], Maximilian H.-W. Schmidt[3,4], Lance Courtney[1,5], Saleh Alseekh[6,7], Emma M. Jobson[1], Alexander Vogel[3], Yimin Xu[1], David Lyon[8], Kathryn Dumschott[4], Marcus McHale[9], Ronan Sulpice[9], Kan Bao[1], Rohit Lal[1], Asha Duhan[1], Asis Hallab[4], Alisandra K. Denton[3], Marie E. Bolger[4], Alisdair R. Fernie[6,7], Sarah R. Hind[10], Lukas A. Mueller[1], Gregory B. Martin[1,11], Zhangjun Fei[1,12], Cathie Martin[2], James J. Giovannoni[1,12], Susan R. Strickler[1*], Björn Usadel[3,4*]

[1] Boyce Thompson Institute, Ithaca, NY 14853, USA

[2] Department of Biochemistry and Metabolism, The John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

[3] Institute for Biology I, BioSC, RWTH Aachen University, 52474 Aachen, Germany

[4] IBG-4 Bioinformatics, Forschungszentrum Jülich, 52428 Jülich, Germany

[5] Plant Biology Section, School of Integrative Plant Sciences, Cornell University, Ithaca, NY 14853, U.S.A.

[6] Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476, Potsdam-Golm, Germany

[7] Center of Plant Systems Biology and Biotechnology, 4000 Plovdiv, Bulgaria

[8] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, CA 94720

[9] Plant Systems Biology Lab, Ryan Institute, National University of Ireland, H91 TK33 Galway, Ireland.

[10] Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[11] Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, U.S.A.

[12] US Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA.

*for correspondence srs57@cornell.edu and b.usadel@fz-juelich.de

§Present address: Institute of Evolution, University of Haifa, Israel

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.15770

#These authors contributed equally

**Running title:** A *Solanum lycopersicoides* reference genome

**Key Words:** *Solanum lycopersicoides,* genome*,* Carotenoids, anthocyanin, disease resistance, drought

Summary

Wild relatives of tomato are a valuable source of natural variation in tomato breeding, as many can be hybridized to the cultivated species (*Solanum lycopersicum*). Several, including *S. lycopersicoides,* have been crossed to *S. lycopersicum* for the development of ordered introgression lines (ILs), facilitating breeding for desirable traits. Despite the utility of these wild relatives and their associated ILs, few finished genome sequences have been produced to aid genetic and genomic studies. Here we report a chromosome-scale genome assembly for *S. lycopersicoides* LA2951, which contains 37,938 predicted protein-coding genes. With the aid of this genome assembly, we have precisely delimited the boundaries of the *S. lycopersicoides* introgressions in a set of *S. lycopersicum* cv. VF36 x LA2951 ILs. We demonstrate the usefulness of the LA2951 genome by identifying several quantitative trait loci (QTLs) for phenolics and carotenoids, including underlying candidate genes, and by investigating the genome organization and immunity-associated function of the clustered *Pto* gene family. In addition, syntenic analysis of R2R3MYB genes sheds light on the identity of the *Aubergine* locus underlying anthocyanin production. The genome sequence and IL map provide valuable resources for studying fruit nutrient/quality traits, pathogen resistance, and environmental stress tolerance.

Introduction

Tomato (*Solanum lycopersicum*) is one of the most widely consumed fruit crops with great world-wide value (fao.org). It is rich in essential nutrients, particularly provitamin A, folate, vitamin C, vitamin E and vitamin K, and calcium (Ali *et al.*, 2020). Tomato yield is, however, often reduced significantly by losses caused by adverse environmental conditions, disease, pest damage, and post-harvest processes (Panno *et al.*, 2021). The narrow germplasm base currently deployed in most breeding programs limits the potential for tomato improvement (Bauchet and Causse, 2012).

Fortunately, close wild relatives present opportunities to add enormous genetic diversity to tomato breeding programs and the means to identify and study genes that underpin useful novel variation. At least 14 wild relatives can be crossed to the cultivated tomato, and have been used for decades in breeding programs (Grandillo *et al.*, 2011). Breeding with *S. lycopersicoides* was until recently prevented by strong crossing barriers and hybrid sterility. While introgression lines (ILs) are now available that can assist identification of alleles and mapping of loci conferring beneficial traits from *S. lycopersicoides* (Canady *et al.*, 2005), few loci have been cloned due, in part, to a lack of a reference genome for this species. Despite the importance of wild accessions in tomato breeding, only two such species, *S. pennellii* (A., Bolger *et al.*, 2014; Schmidt *et al.*, 2017) and *S. pimpinellifolium (Wang et al., 2020)*, have reference-quality genome assemblies. Three species, *S. galapagense (Strickler et al., 2015)*, *S. arcanum (The 100 Tomato Genome Sequencing Consortium et al., 2014)*, and *S. habrochaites* (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014) have draft *de novo* assemblies with small contig size and no gene annotation, whereas *S. sitiens* (Molitor *et al.*, 2021) and *S. lycopersicum* var. *cerasiforme* (Takei *et al.*, 2021) have been scaffolded using similarity to reference genomes of related species. Of the biparental tomato IL populations developed to date, the *S. lycopersicum* cv. VF36 x *S. lycopersicoides* LA2951 ILs (Canady *et al.*, 2005) represents one of the widest crosses possible, because *S. lycopersicoides* belongs to a group that is sister to the tomato clade. Due to the lack of a *S. lycopersicoide*s genome, introgression boundaries of these lines are typically defined based on the published tomato reference genome (Tomato Genome Consortium, 2012). Genetic mapping resolution depends on these boundaries being well-defined. Traditionally, boundaries have been defined by DNA markers dispersed across the genome, but more recent approaches increase precision by using single nucleotide polymorphisms (SNPs) derived from resequencing or RNA-seq data (Gonda *et al.*, 2019); (Chitwood *et al.*, 2013). However, basing introgression coordinates on the 'Heinz 1706' reference genome and annotation in wide crosses may lead to important genomic features in the non-reference genome being overlooked. Indeed, while many differences between genomes are accounted for by repetitive elements, gene gain and loss are also important to consider, as considerable variation in genes involved in tomato improvement, including fruit quality and stress tolerance, was observed in the tomato pan genome (Gao *et al.*, 2019) as well as in efforts to mine structural variation in tomato and wild relatives (Alonge *et al.*, 2020; Wang *et al.*, 2020).

Wild relatives and ILs have been used to investigate flavour and nutritional QTL (Klee and Tieman, 2018; Lee *et al.*, 2012; Tohge *et al.*, 2020). Remarkably, under favorable conditions, *S.*

*lycopersicoides* produces potentially health-beneficial anthocyanins (Martin *et al.*, 2011) resulting in a purple or black fruit, a trait not seen in cultivated tomatoes or their close relatives, but which is also expressed to a lesser degree in *S. cheesmaniae, S. chilense* and *S. peruvianum (Bedinger et al., 2011; Rick et al., 1994)*. The *Aubergine* (*Abg)* locus from chromosome 10 of *S. lycopersicoides* has been introgressed into cultivated *S. lycopersicum* resulting in anthocyanin production, but the identity of the underlying gene remains unknown.

*S. lycopersicoides* has also been investigated for its response to biotic and abiotic stressors. Some *S. lycopersicoides* accessions are susceptible to three nematode species (*Meloidogyne incognita*, *M. javanica*, and *M. hapla*) and to the oomycete responsible for the Irish potato famine, *Phytophthora infestans* (Phills *et al.*, 1977). Interestingly, no symptoms were observed in *S. lycopersicoides* upon inoculation with several viruses including cucumber mosaic virus and tobacco mosaic virus even though the plants were systemically infected and carried a high titer of the virus (Phills, B.R., Provvidenti, R. and Robinson, R.W., 1977; Zhao *et al.*, 2005). Previous studies have demonstrated resistance of *S. lycopersicoides* to the necrotrophic fungal pathogen, *Botrytis cinerea*, and identified resistance to another necrotrophic pathogen *Alternaria solani* (Guimarães *et al.*, 2004; Davis *et al.*, 2009; Smith *et al.*, 2014; Phills, B.R., Robinson, R.W. and Shail, J.W, 1977; Egashira *et al.*, 2000). Investigation of the genetic basis of resistance to *Botrytis* revealed five QTLs, with four decreasing the frequency of the infection and one reducing lesion diameter (Davis *et al.*, 2009). Other studies identified resistance in *S. lycopersicoides* to the biotrophic fungal pathogens *Cladosporium fulvum* and *Fusarium oxysporum* sp. *lycopersici* although genetic analysis of the resistance has not been reported (Phills, B.R., Provvidenti, R. and Robinson, R.W., 1977; Zhao *et al.*, 2005). Recently the *Ptr1* locus was identified in *S. lycopersicoides,* which mediates recognition of the bacterial type III effector AvrRpt2 (Mazo-Molina *et al.*, 2019). This effector is conserved in race 1 strains of the bacterium *Pseudomonas syringae* pv. *tomato* for which no other genetic resistance has been available. The *Ptr1* gene encodes a coiled-coil nucleotide binding leucine rich repeat (CC-NLR) protein (Mazo-Molina *et al.*, 2020). Additionally, *S. lycopersicoides* has been shown to exhibit enhanced cold tolerance (Zhao *et al.*, 2005) and adaptation to arid environments (Peralta *et al.*, 2008), and the IL population has been used to identify QTLs affecting salt tolerance in seedlings (Li *et al.*, 2011).

To improve understanding of the genes responsible for these agriculturally important traits and increase the utility of the introgression lines, we generated a chromosome-scale, reference genome for *S. lycopersicoides*. We compared the genome to those of *S. pennellii* and *S.*

*lycopersicum* to identify unique features that may be key to the understanding of the genetics of abiotic and biotic stress tolerance in *S. lycopersicoides*. Using RNA-seq, we mapped the introgressions of the ILs to both parental genomes to refine the genetic map of the population. We use the genome and IL maps to explore the evolution of tomato fruit color and immunity. The genome and associated map will facilitate understanding of genes responsible for agriculturally important traits, enabling their targeted introduction into tomato breeding lines.

## Results

### Genome assembly and feature prediction

*S. lycopersicoides* LA2951 was selected for sequencing due to its genetic distance from the cultivated tomato and the existence of introgression lines previously generated using this accession as the donor parent (Canady *et al.*, 2005). We assembled approximately 17 million PacBio reads with an average length of 6.29 kb totaling 107 Gb using the two different long-read genome assembly programs, Canu (Koren *et al.*, 2017) and Falcon (Chin *et al.*, 2016). Canu produced an assembly with 17,507 contigs and an N50 length of 139,475 bp, which was more contiguous and the assembly captured a larger proportion of the BUSCO set with only 48 missing genes versus 246 for Falcon (Waterhouse *et al.*, 2017). Thus, the Canu assembly was selected for further scaffolding using Dovetail Chicago and Hi-C. The final assembly is 1.2 Gb in length (Table 1) and captures 97% of the BUSCO set (Supplemental Table 1). The genome assembly is larger than the *S. lycopersicum* assembly, in line with previous observations (Rick *et al.*, 1986; Menzel, 1962; Ji *et al.*, 2004). The final assembly had an N50 scaffold size of 93.9 Mb and 90% of the assembly was assigned to the 12 chromosomes. Approximately 1% of sites were found to be heterozygous.

Figure 1 shows the twelve pseudochromosomes, along with repeat density and gene feature density. About 68% of the sequence contained in the pseudochromosomes consists of repeats (Supplemental table 2). Genome annotation predicted a total of 37,939 putative genes with a mean length of 4,388 bp. Genes had an average of 5.2 exons and a coding sequence (CDS) length of 1,232 bp. We were able to identify 96% of the BUSCO set in the predicted proteins (Supplemental Table 1).

Of all repeat classes, the long terminal repeat retrotransposons (LTR-RTs) constituted the greatest percentage (~57%) of the *S. lycopersicoides* genome, as was the case with *S. lycopersicum* and *S. pennellii* genomes (Supplemental table 2). Among the LTR-RT elements, *Gypsy*-type elements were proportionately more abundant than *Copia*-type elements in all three genomes (Supplemental table 2) (Bolger *et al.*, 2014). *S. lycopersicoides* also had a greater proportional abundance of younger LTR-RTs (with insertion times < 1 million years ago) compared to the other two genome assemblies (Supplemental figure 1), consistent with evidence of expansion of the repetitive fraction based on GISH cytology (Ji *et al.*, 2004).

**Mapping of *S. lycopersicoides* introgressions in IL accessions**

RNA-seq data from 71 unique *S. lycopersicum* cv. VF36 x *S. lycopersicoides* LA2951 IL accessions were used to assemble genotype maps of the population. The map based on the current domesticated tomato reference genome (SL4.0) provides the coordinates delimiting the introgression boundaries within the reference parent background (Supplemental table 3, Supplemental Table 4, Supplemental Figure 2). An additional map based on the present *S. lycopersicoides* reference provides the coordinates which define the introgressed regions within the wild donor parent (Figure 2, Supplemental Table 5). A total of 87% of the *S. lycopersicoides* LA2951 genome is represented across these 71 IL accessions. Our analysis revealed numerous unidentified introgressed segments and chromosomal features when the Heinz 1706 genome was used as the reference. For example, the IL line LA4245 is believed to harbor a single introgression on chromosome 4 (Canady *et al.*, 2005). Our results show that LA4245 harbors an additional ~400 kb introgressed region at the distal end of the short arm of chromosome 4. However, although this introgression appears at the end of the chromosome within LA4245, it is derived from a region closer to the centromere of chromosome 4 in *S. lycopersicoides.* This discovery prompted us to explore structural variation between the parental genomes. Alignment and visualization of paired chromosomal segments of at least 8 kb and 92% sequence identity showed high degrees of synteny between *S. lycopersicoides* and both *S. lycopersicum* and *S. pennellii*. Syntenic dot plots performed at a finer scale identified inversions between LA2951 and the domesticated tomato genome (Supplemental figure 3). One inversion, found on chromosome 10 at approximately 87-94 Mb based on domesticated tomato genome coordinates, is supported by a previously characterized inversion on chromosome 10 in *S. lycopersicoides* relative to *S. lycopersicum* and *S. pennellii (Pertuzé et al., 2002)*. This inversion was not detected cytologically using BAC probes (Szinay *et al.*, 2012). These findings demonstrate the risks of low-resolution

genotype mapping in missing identification of introgressed segments and highlight the potential for error in downstream analyses. In addition, it shows that reference-quality genomes of the introgression donors help shed light on introgression fragments as they aid in deciding on the provenance of regions. Without adequate genotyping, QTL mapping may be misleading. The present genome allowed for high-resolution genotype mapping of associated ILs, providing a resource for the community to map QTLs more efficiently and to characterize the variation available from *S. lycopersicoides*. In addition, we used both the *S. lycopersicum* and *S. lycopersicoides* reference genomes and the introgression coordinates from the IL map to develop in-silico hybrid reference genomes specific to each IL where the introgressed part is derived from the *S. lycopersicoides* sequence and the rest from *S. lycopersicum* sequence as an optimal RNA-seq mapping reference. This minimizes expression anomalies resulting from mapping to a reference other than that from which a transcript was derived (Supplemental Figure 4).

### *S. lycopersicoides* ILs provide insight into the evolution of fruit color in tomato

Carotenoigenic biosynthesis regulation during tomato fruit ripening is mainly governed by transcriptional regulation of the coding genes for structural enzymes along the biosynthesis pathway (Figure 3). During fruit ripening, the accumulation of lycopene, the predominant carotenoid in ripe tomato fruit, is mediated through transcriptional upregulation of coding genes for enzymes upstream of lycopene in the biosynthesis pathway, including PSY1, ZISO, and CRTISO. In addition, there is transcriptional downregulation of genes coding for enzymes downstream of lycopene, CRTL-E and CRTL-B, for which two coding genes were described in tomato: CRTL-B1 (Pecker et al 1996), CRTL-B2 (BETA), which was predominantly associated with chloroplast-containing tissues (Ronen *et al.*, 2000).

Variation in fruit carotenoid accumulation mediated by *S. lycopersicoides* allelic diversity in the ILs was obvious in mature fruits (Supplemental figure 5). IL fruits were generally lighter in color and shifted from red (lycopene) to orange (ß-carotene) pigment. Analysis of mature IL fruit carotenoid content revealed 15 potentially associated loci (Supplemental table 6, Supplemental table 7). For instance, IL4232, shows a significant decrease in lycopene content (Supplemental table 6 associated with c1m18 BIN QTL (Supplemental figure 6, Supplemental table 8).

To pinpoint underlying genes we next analyzed gene expression in mature fruits of different ILs. This analysis revealed that *S. lycopersicoides* alleles of key carotenoid pathway genes upstream of lycopene were expressed at a substantially lower level in ILs containing the *S. lycopersicoides*

alleles than in those harboring the *S. lycopersicum* alleles (Supplemental figure 7). One notable *S. lycopersicoides* allele with reduced expression found in BIN c3m3 encodes phytoene synthase 1 (PSY1), exhibiting 3-fold reduced mRNA accumulation as compared to its ortholog in the cultivated tomato.

PSY1 encodes the first committed step in the carotenoid biosynthesis pathway and is highly induced at ripening initiation in red tomatoes. Loss of function of PSY1 is the basis of the yellow-flesh mutation (*r*) resulting in pale, yellow, low carotenoid fruit (Fray and Grierson, 1993), Pathway inhibition in the ILs was not as severe as in the yellow-flesh mutant due to heterozygosity of the *PSY1* gene maintained in these ILs (Supplemental figure 7). Fruit from overlapping chromosome 12 introgressions harboring a low expression allele of the ζ -carotene isomerase (ZISO), which isomerizes 9,15,9' tri-cis-ζ-carotene to 9,9' di-cis-ζ-carotene (a precursor of prolycopene), also had low lycopene levels (Supplemental figure 6 and (Supplemental figure 7 G). Lycopene is the red carotenoid predominant in ripe *S. lycopersicum* tomatoes. VIGS-mediated silencing of this gene in tomato resulted in similarly pale-red low lycopene fruit (Fantini *et al.*, 2013). Elevated expression of these genes in the red cultivated tomato implies enhanced flux toward carotenoid synthesis. Downstream of lycopene, lycopene ε-cyclase (CRTL-E) was elevated in expression in lines harboring the *S. lycopersicoides* allele implying that fruit containing low lycopene could be due to redirection of carotenoid flux away from lycopene and toward lutein (Supplemental figure 6 and Supplemental figure 7 F). This is similar to the previously described gene from *S. pennellii* whose corresponding allele is termed *Delta* (*Del*) (Ronen *et al.*, 2000; Ronen *et al.*, 1999) for elevated delta carotene accumulation.

Lycopene ß-cyclase is a key enzyme influencing lycopene accumulation as it catalyzes the conversion of lycopene to ß-carotene. Expression of the fruit *lycopene ß-cyclase 2* (*CRTL-B2*) gene is repressed in cultivated tomatoes by ethylene during ripening, facilitating the accumulation of lycopene via reduced conversion to ß-carotene (Alba *et al.*, 2005). The *S. lycopersicoides* allele however is not repressed during ripening (Figure 3), resulting in the conversion of lycopene to ß-carotene leading to orange fruit in IL4254 (Supplemental figure 5 and Supplemental figure 6). As with the gene encoding the ε-cyclase, the *S. lycopersicoides* allele is similar to the previously described *Beta* allele of *S. pennellii* (Ronen *et al.*, 2000) in that it drives flux away from lycopene accumulation due to enduring high *lycopene* ß-cyclase expression. A second *CRTL-B* gene, *CRTL-B1*, resides in an introgression conferring low lycopene on chromosome 4, and showed an elevated expression of this allele compared to *S. lycopersicum* VF36, driving flux from lycopene toward ß-carotene (Supplemental figure 6 and Supplemental figure 7 B). Chromosome 10 harbors

two carotenoid accumulation loci, of which one (c10m5) includes both *CRTR-E*, responsible for ε-hydroxylation of α-carotene toward the synthesis of lutein, and a putative lycopene ß-cyclase, *CRTL-B3,* which has not been functionally characterized. Like *CRTL-B1* and *B2*, the *S. lycopersicoides* allele of *CRTL-B3* is more highly expressed than the *S. lycopersicum* allele (Supplemental figure 7 D), resulting in a shift in the lycopene to ß-carotene ratio of almost 6:1 in the VF36 control to 1:1 in the IL (Supplemental figure 8). As this gene has not been previously characterized, we demonstrated its product's ability to convert lycopene to ß-carotene by in-vitro expression in *E. coli* (Supplemental figure 8). Phylogenetic analysis of the *CRTL-B* genes indicates a duplication resulting in the creation of *CRTL-B3,* which occurred between the separation of tobacco and petunia (Supplemental figure 9), and evolution of red, lycopene accumulating fruit is accompanied by silencing of this gene during ripening (Supplemental figure 10). In summary, analysis of the effects of *S. lycopersicoides* carotenoid QTLs is consistent with the fact that green-fruited wild tomatoes can remain photosynthetic (Kilambi *et al.*, 2017; Cipollini and Levey, 1991). The red-fruited tomatoes, however, show a comparable induction of genes leading to the red pigment lycopene and repression of downstream genes involved in lycopene catabolism during ripening (Figure 3).

**Flavonoid pathway**

*S. lycopersicoides* LA2951 fruit accumulated about 200-fold more polyphenols (phenylpropanoids, chlorogenic acids, and flavonoids) compared to the VF36 cultivated control ( Supplemental table 9). Under high light intensity conditions, *S. lycopersicoides* produces anthocyanins in its otherwise green fruit. This prompted us to analyze the region on chromosome 10 harboring the R2R3MYB transcription factors belonging to subgroup 6, known to regulate anthocyanin biosynthesis in plants (Muñoz-Gómez *et al.*, 2021). On chromosome 10 of *S. lycopersicoides* there were three clustered genes encoding subgroup 6 R2R3MYB transcription factors, in a region syntenic with a cluster of five R2R3MYB subgroup 6 genes in *S. lycopersicum* (*AN2, AN2-like, AN2-like2, ANT1*, *and ANT1-like*) on chromosome 10 (Figure 4B). In addition, another three R2R3-MYB subgroup 6 genes were identified by searching the *S. lycopersicoides* genome sequence in a contig not anchored to the genome. These are likely the result of heterozygosity in the *S. lycopersicoides* genome. Indeed, the ILs IL10-2 (LA4275) and IL10-4 (LA4276), which both contain the clustered R2R3-MYB transcription factors, have been reported as available only as heterozygotes for the *S. lycopersicoides* interval from chromosome 10 (Canady *et al.*, 2005), supporting the view that the additional copies detected in the parental genome are alternative alleles of the annotated genes. If this entire region were heterozygous on

chromosome 10 in the sequenced *S. lycopersicoides* genome, it would affect all the genes in the cluster, which is exactly what was observed.

A phylogenetic tree of the R2R3MYB subgroup 6 genes from *S. lycopersicum*, *S. lycopersicoides*, *S. pennellii*, potato (*S. tuberosum*), and petunia (*P. axillaris*, *P. inflata* and *P. hybrida*), showed that the major gene duplications (in fact a 5-fold multiplication) occurred before the divergence of the *Solanum* family, but independently of the gene multiplications that occurred in petunia (Figure 4A).

A syntenic map was constructed based on the genomic regions harboring subgroup 6 R2R3MYB genes from four species in the Solanaceae family (*S. lycopersicum*, *S. pennellii*, *S. lycopersicoides*, and the diploid *S. tuberosum*) (Figure 4B). In general, the five clustered genes encoding R2R3-MYB transcription factors belonging to subgroup 6 (*ANT1, ANT1-like, AN2, AN2-like*, and *AN2-like2*) are present in different *Solanum* species, derived from the ancestor of the genus *Solanum* (and represented by six genes encoding proteins in the different clades in potato). One clade comprising genes encoding ANT1 and ANT1-like in tomato has been lost in *S. lycopersicoides* (Figure 4A, 4B) leaving just three genes encoding R2R3-MYB subgroup 6 proteins. Based on the genome annotation, in one of the other three clades (*AN2*, *AN2-like*, and *AN2-like2*), a gene (*AN2-like2*) appears to have been lost in tomato but was present in potato, *S. lycopersicoides*, and *S. pennellii*. However, our analysis of synteny showed that a sequence in the tomato genome (*Solyc10g086300*.1), lying closer to the telomere than the other four genes in the subgroup 6 MYB cluster and annotated as an 'unknown protein', was syntenic to the *AN2-like2* genes, *SpAN2-like2*, *SlydAN2-like2*, and *StAN2-like2*. Consequently, we renamed the genomic region containing *Solyc10g086300.1* as *SlAN2-like2*. Furthermore, the sequence upstream of *Solyc10g086300*.1 has a high similarity with the sequences upstream of the coding sequences of the *SlydAN2-like2* and *StAN2-like2* genes (Figure 4C). We found that the predicted AN2-like2 protein in *S. pennellii* is much shorter than the predicted AN2-like2 proteins from *S. lycopersicoides* and *S. tuberosum* and it is truncated at its 5' end encoding the N-terminal DNA binding domain. However, the sequence upstream of *SpAN2-like2* (marked as 'SpAN2-like2 upstream') also aligned well with the upstream regions of the *SlydAN2-like2* and *StAN2-like2* genes (Figure 4C). These results suggested that, in this clade, the *AN2-like2* gene from the ancestor of the genus Solanum (represented by *StAN2-like2*) and *SlydAN2-like2* encode intact MYB proteins, which may also be functional. However, during evolution, polymorphisms causing a premature stop codon resulted in the loss of function of the corresponding AN2-like2 proteins in *S. lycopersicum* and *S. pennellii*. For the other two *Solanum* R2R3-MYB subgroup 6 subclades

comprising *AN2* and *AN2-like,* respectively, there are single copies in each subclade in *S. lycopersicum*, *S. pennellii,* and *S. lycopersicoides*, with phylogenetic distances that support the *S. lycopersicoides* genes being sister to the *S. lycopersicum* and *S. pennellii* sequences.

From the tissue-specific RNA-seq of *S. lycopersicoides*, we found that SlydAN2-like was predominantly expressed in fruit, while no transcripts were detected for SlydAN2 or SlydAN2-like2 in fruit (Supplemental table 10 and online database).

The *Anthocyanin in fruit (Aft)* allele of *SlAN2* (*SlAN2-like*$^{Aft}$) from *S. chilense* also promotes anthocyanin production under appropriate light conditions. SlAN2-like$^{Aft}$ shares high similarity in its protein sequence with SlydAN2-like (89%), compared to SlydAN2 (62%) or SlydAN2-like2 (59%). Therefore, the *SlydAN2-like* gene introgressed into cultivated tomato very likely contributes to the Aubergine phenotype. It has been reported that the expression level of *SlAN2-like*$^{Aft}$ is significantly higher than *SlAN2-like*$^{WT}$ in fruit (Colanero, Tagliani, *et al.*, 2020), so we compared promoter regions of *SlAN2-like*$^{WT}$, *SlAN2-like*$^{Aft}$, *ScAN2-like* and *SlydAN2-like* genes (Figure 4E). There were two blocks of sequence conserved in the *AN2-like* alleles conferring fruit anthocyanin production but absent from the WT allele from *S. lycopersicum*. The gene-proximal sequence includes an AC box on the negative strand, which is a recognition motif for R2R3-MYB proteins, suggesting that fruit-specific expression of *SlydAN2-like* might be subject to autoregulation as has been reported for *MdMYB10* in apples (Espley *et al.*, 2009). The presence of this sequence may cause the elevated transcript levels of *AN2-like* genes in fruit, particularly in response to light, although these *in silico* analyses need further verification with experimental testing of wild type and mutant promoters in fruit.

When *S. lycopersicoides* was grown in our growth chambers (which is not conducive to the light-dependent phenotype as it lacks UV components inducing anthocyanins (Colanero, Perata, *et al.*, 2020)) only green ripe fruit without visible anthocyanin production were produced, although we did note substantial variation in phenolic compounds (Supplemental table 9). Hence, we used the IL population to identify QTLs impacting the accumulation of phenolic compounds. On chromosome 5 (IL4252 and IL4299) we identified a metabolic QTL (mQTL) positively influencing chlorogenic acid contents (Figure 5, Supplemental table 11, Supplemental figure 11), which was correlated to reduced *CHS2* expression.

Another mQTL on chromosome 10 resulted in up to 5-fold elevation of total phenylpropanoids (Figure 5, Supplemental table 11, Supplemental figure 11). *PAL5* resides at this locus as it does

for a *S. neorickii* phenylalanine mQTL (Brog *et al.*, 2019). In contrast to cultivated tomato, wild tomato species accumulate significant phenolics in fruit pericarp (Willits *et al.*, 2005) as do lines with the *S. lycopersicoides* allele of this mQTL locus (Supplemental figure 12).

Differences in flavonoids might also be due to differences in the genes encoding the enzymes of flavonoid synthesis and modification, as we detected an expansion of these genes. Whilst the gene encoding dihydroflavonol 4-reductase was triplicated (Solyd02g073780, Solyd02g073830, Solyd02g073850) in *S. lycopersicoides* compared to the domesticated tomato and *S. pennellii*, the gene encoding flavonol-3-O-glycoside-rhamnosyltransferase exists in two copies on chromosomes 3 and 5 in the genome of domesticated tomato, *S. pennellii* shows a tandem duplication of the gene on chromosome 3 of *S. pennellii* and a (near) tandem duplication on both chromosomes 3 and 5 of *S. lycopersicoides* (Solyd03g076070, Solyd03g076130, Solyd05g056030, Solyd05g056040).

**Immunity-associated genes in the LA2951 genome**

The LA2951 genome sequence was examined for genes encoding potential immunity-associated proteins. Considering only full-length gene models in LA2951 as compared to Heinz 1706, there are more Nod-like receptors (NLRs; 254 vs. 123), but comparable numbers of receptor-like kinases (RLKs; 203 vs. 214) and receptor-like proteins (RLPs; 24 vs. 30) Supplemental table 12). A search in LA2951 for homologs of other genes potentially associated with effector- or pattern-triggered immunity (ETI and PTI), identified homologs, most of which had sequence identities greater than 95% (Supplemental table 13). In several cases, two closely related genes occur in both LA2951 and Heinz 1706 indicating these gene duplications occurred in a common ancestor (e.g., *Bti9a/Bti9b*, *Fls2.1/Fls2.2*, *Pti1a/Pti1b*) (Zeng *et al.*, 2012; Schwizer *et al.*, 2017; Roberts *et al.*, 2020).

The *Pto* resistance gene was originally identified as a member of a clustered gene family on a segment of chromosome 5 derived from *S. pimpinellifolium (Martin et al., 1993)*. Embedded within this cluster is *Prf* which encodes a CC-NLR that acts in concert with the Pto kinase and another member of the Pto family, Fen, to confer resistance to race 0 strains of *Pseudomonas syringae* pv. *tomato* (*Pst*) (Pedley and Martin, 2003). To test whether LA2951 has a functional *Pto* gene, we inoculated three ILs carrying the LA2951 *Pto/Prf* region in the background of VF36 with *Pst* DC3000 (Figure 6A). *Pst* DC3000 has the effectors AvrPto and AvrPtoB, both of which are recognized by the Pto/Prf complex. As expected, tomato line RG-PtoR, which has the *Pto/Prf* genes, developed no disease in response to DC3000, whereas related lines with mutations in *Pto*

or *Prf* (RG-pto11 and RG-prf3) developed disease symptoms. All three ILs developed extensive disease symptoms on stems and leaves in response to DC3000. Another *Pst* strain, DC3000mut5, is recognized by *Fen* in lines lacking a *Pto* gene. An IL containing the *Pto/Prf* region of LA2951 was also susceptible to this *Pst* strain (Supplemental Figure 13). Together these observations suggest that LA2951 lacks functional *Fen* and *Pto* genes or possibly lacks a functional *Prf* gene. Analysis of the LA2951 genome sequence spanning the *Pto/Prf* region revealed that orthologs of *PtoA* (96%) and *Fen* (93%) are present. *PtoD* is present but contains a premature stop codon. A homolog of *Prf* was also identified, but it contains an insertion with a mutation leading to a premature stop codon, likely rendering the gene nonfunctional (Figure 6B). The susceptibility of the ILs to DC3000 and DC3000mut5 is therefore likely due to the lack of functional *Fen*, *Pto*, and *Prf* genes in LA2951.

**Expression analysis of drought resistance**

To analyze the effect of drought stress on *S. lycopersicoides,* water was withheld and transcriptome profiling was conducted at two time points and in a control treatment. Whilst the initial drought condition yielded several thousand differentially expressed genes (DEGs), sustained low water yielded fewer DEGs, but both sets exhibited a large overlap of genes similarly regulated. Interestingly, although a number of genes (172) that were up-regulated at the earlier time point were down-regulated later, many more targets (460) were down-regulated first and up-regulated later. This pattern is consistent with widespread down-regulation of gene expression at the onset of drought followed by recovery under the low-water condition accompanied by sustained activity of a set of up-regulated genes at both time points. The set of genes displaying sustained up-regulation across both time points might play a protective role. Indeed, the most significant among these were two annexin homologs (Solyd04g071760, Solyd04g071720), a protein family with well-established roles in drought stress response in tomato (Ijaz *et al.*, 2017). Analyzing the drought response at a pathway level revealed a consistent down-regulation of many LRR-III protein kinase family members at both time points. This family includes the *AtRDK1* gene in *Arabidopsis*, whose loss of function results in drought hypersensitivity (Kumar *et al.*, 2017). In addition, many other genes encoding post-translational modifiers were upregulated at both time points, including clade A of the PPP Fe-Zn-dependent phosphatase families and glutaredoxins (Supplemental Table 14).

To make the expression dataset readily accessible, we developed a web-based expression atlas for *S. lycopersicoides*, the Lycopersicoides Expression Atlas (http://lycopersicoides-

ea.sgn.cornell.edu), which also includes gene expression in multiple tissues. The website allows users to select a project dataset and explore visualizations and analyses of the data using tools including an expression cube and scatter plot viewer. Additional datasets contributed from the community can be included in the future.

## Discussion

The last few years have seen a tremendous advance in plant genomic sciences driven by technologies such as PacBio long-read and Oxford Nanopore sequencing (Bolger *et al.*, 2019) as well as new technologies to obtain long-range chromatin information, including optical mapping and Hi-C data (Schreiber *et al.*, 2018). Indeed, it has been shown that these technological advances can be used to develop chromosome-scale assemblies for large, complicated plant genomes (Belser *et al.*, 2018); (Li *et al.*, 2021) and that they are particularly useful for unraveling genomes of wild relatives of important crops, providing information about exotic germplasm, and facilitating its use (Wu *et al.*, 2018; Sun *et al.*, 2020). However, while this would be particularly useful in the tomato clade, reference-grade assemblies exist only for the domesticated tomato (Tomato Genome Consortium, 2012; Razali *et al.*, 2017), the red-fruited *S. pimpinellifolium (Wang et al., 2020)*, and the wild species *S. pennellii (A., Bolger et al., 2014; Schmidt et al., 2017)*. Here, we present a novel high-quality chromosome-scale assembly for the genome of *S. lycopersicoides*, which is one of the most distantly related sexually compatible wild species to the cultivated tomato, representing an exotic germplasm donor that, together with *S. sitiens*, comprises the *Solanum* sect. *Lycopersicoides*, sister to the sect. *Lycopersicon* harboring domesticated tomato (Knapp and Peralta, 2016).

Using long-read sequencing technology and Dovetail scaffolding, we provide a genome with >90% of the sequence gathered in 12 chromosome-like scaffolds and very high gene content completeness. This is of particular importance as *S. lycopersicoides* has been used to establish an introgression line population (Canady *et al.*, 2005) mapped for abiotic and biotic stress tolerance QTL underlying the enhanced resilience of this wild species. Thus, this genome provides a reference to facilitate the fine mapping and identification of causative genes underlying IL QTL. In line with previous marker-based data, we also observed an inversion on chromosome 10 of the cultivated tomato relative to the ancestral *S. lycopersicoides* configuration (Pertuzé *et al.*, 2002). We also found an additional inversion between the cultivated tomato and *S. lycopersicoides* on chromosome 4, which had previously been speculated to be a hotspot for rearrangements (Albrecht and Chetelat, 2009). The complete genome sequence enabled

understanding of the evolution of subgroup 6 R2R3-MYB genes in the genus *Solanum*, exemplifying that this genome sequence, together with extant *Solanum* genomes, could be applied to investigate the evolution of subgroup 6 R2R3-MYB genes in other species, as well as to understand the evolution of other gene families in the genus *Solanum*, and facilitate the verification of gene annotation in related plant species.

Fruit of domesticated tomato and its wild progenitor, *S. pimpinellifolium*, does not normally accumulate anthocyanins, either in the fruit flesh or in the peel. However, some green-fruited species produce purple anthocyanins in their fruit, notably *S. lycopersicoides* and *S. chilense*. In both species, purple anthocyanin pigmentation occurs principally in the fruit peel and is strongly dependent on incident light. The *Aubergine* (*Abg*) locus was identified in *S. lycopericum* x *S. lycopersicoides* introgression lines and is responsible for anthocyanin production in purple-peeled fruit in *S. lycopersicum (Rick et al., 1994)*. *Abg* has been mapped to chromosome 10 in *S. lycopersicoides*. However, a paracentric inversion from the cross between cultivated tomato and *S. lycopersicoides* prevents the introgression from being stable (Canady *et al.*, 2006). Therefore, the genetic nature of the *Abg* locus has remained unclear. The phenotype of *Abg* lines is very similar to that of *Aft* (derived by introgression from *S. chilense*) in tomato. *Aft* has been shown to be conferred by the *S. chilense* allele of *SlAN2-like (Yan et al., 2020; Colanero, Tagliani, et al., 2020)*. The two *S. lycopersicoides* genes evolutionarily closest to *SlAN2-like* (*SlydAN2-like* and *SlydAN2-like2*) are therefore the most likely candidates for *Abg (Yan et al., 2020; Colanero, Tagliani, et al., 2020)*. RNA-Seq data from fruit of *S. lycopersicoides* showed elevated expression of *SlydAN2-like* alone among the genes encoding R2R3MYB subgroup 6 transcription factors on chromosome 10, supporting the view that the *SlydAN2-like* gene in *S. lycopersicoides* confers the *Abg* purple fruit phenotype.

Fruits of the *S. lycopersicoides* ILs were generally lighter in color and shifted from red (lycopene) to orange (ß-carotene) accumulation. Using an IL mapping strategy relying on the improved genome, we could show that *CRTL-B2* is not repressed during ripening in *S. lycopersicoides* as is its cultivated tomato ortholog. Ripening-related *CRTL-B2* repression in cultivated tomato contributes to lycopene accumulation and the characteristic red tomato color, while CRTL-B2 activity catalyzes the conversion of lycopene to ß-carotene (orange) and downstream metabolites. This was also the case for the previously uncharacterized *CRTL-B3,* underlying another QTL for carotenoid accumulation. We showed that *CRTL-B3* encodes a gene with lycopene ß-cyclase activity, which together with somewhat reduced expression of *CRTL-B1,* indicates that

transcriptional regulation plays a major role in determining carotenoid accumulation that could be deduced using the improved IL maps enabled by a high-quality *S. lycopersicoides* reference genome.

Many resistance genes used in tomato varietal improvement programs were originally identified in wild relatives and subsequently introgressed into breeding lines (Blanca *et al.*, 2015). Resistance to several microbial pathogens has been reported in *S. lycopersicoides* and the genome sequence, in combination with the IL population, will now facilitate cloning of these and additional underlying genes. Testing the response of LA2951 ILs to two *Pst* strains and analysis of the *Pto/Prf* region in LA2951 showed that it lacks the *Pto, PtoC,* and *PtoF* genes. *Fen* is present although it is unable to detect the truncated AvrPtoB. At present we cannot say whether this is because of its diverged sequence or because *Prf* is nonfunctional due to a premature stop codon. *S. lycopersicoides* may have an ancestral arrangement of the *Pto/Prf* region before multiple gene duplications occurred in this region. Based on molecular analyses of AvrPtoB, we hypothesized previously that *Fen* arose first to recognize variants of AvrPtoB that lack an E3 ligase domain and that *Pto* evolved later to overcome AvrPtoB variants with the E3 ligase domain (Rosebrock *et al.*, 2007; Mathieu *et al.*, 2014). The presence of *Fen,* even though it might be nonfunctional, and not *Pto* in LA2951 supports this hypothesis and it will be interesting to examine *Pto/Prf* genome organization in additional wild relatives of tomato to gain further insight into the evolution of this clustered gene family.

Finally, in addition to improving QTL analyses, we also demonstrated the importance of this genome in assessing gene expression data. Given the large evolutionary distance between *S. lycopersicoides* and the domesticated tomato, simply mapping *S. lycopersicoides* RNA-seq data to the domesticated gene models showed a low mapping rate, resulting in expression results sometimes different from those obtained with the optimal reference. Mapping to closely related relatives has been used as a stop gap (Hekman *et al.*, 2015) and was recommended as a better strategy than *de novo* transcriptome assemblies (Vijay *et al.*, 2013). Our data show that this approach can be problematic especially with wide interspecific crosses. While sophisticated bioinformatics pipelines that allow more flexible read alignment and iterative analysis of data and normalization (Zhou *et al.*, 2019) can alleviate these problems, a high-quality reference genome provides the best solution. To improve gene expression analyses within ILs, we present a reference grafting strategy. Using the coordinates from the IL map and both parental reference genomes, we have assembled synthetic genomes specific for each IL. These grafted reference

genomes dramatically improve RNA-Seq mapping efficiency within introgressed regions of the ILs and yield the most accurate transcriptomic outputs. IL populations can be powerful tools for both breeding and research communities and the wild reference genome improves their utility. The high-quality *S. lycopersicoides* genome presented here provides a tool for plant evolutionary analyses, characterization of QTLs derived from exotic and stress resilient tomato germplasm, gene discovery and crop improvement.

## Experimental Procedures

### Plant Material

Seeds of *S. lycopersicoides* LA2951 and the corresponding IL line seeds were obtained from Tomato Genetics Resource Center (TGRC; https://tgrc.ucdavis.edu/). After germination, they were transferred to soil and further cultivated in a greenhouse supplemented with artificial light to a light intensity of at least 200 µmol $m^{-2}$ $s^{-1}$ generated using Phillips hpi-t plus 400w/645 metal-halide lamps for 16 h a day. To preserve the genotype of the used accession, one plant was chosen and propagated by cuttings. Young fresh leaves were collected and used for DNA extraction. *S. lycopersicoides* plants, used for RNAseq and fruit metabolic analysis, were grown in a growing chamber, equipped with HID illumination, photoperiod of 10 hours light, 14 hours dark, in root limiting conditions to induce flowering. Flowers were crossed using 'mass sib' pollination. Fruit, flowers, and leaves were ground in liquid $N_2$ into very fine powder. An aliquot of 200mg of each of these tissues was used to extract total RNA using Qiagen RNeasy mini kit. RNA-Seq libraries were prepared following the method described in Zhong et al., 2011 (Zhong *et al.*, 2011).

For a drought stress experiment, a total of 40 plants were grown in groups of eight with randomized positions in a phyto chamber with artificial light at 350 µmol $m^{-2}$ $s^{-1}$ for 16 h per day. Humidity was maintained at ~70% and the temperature was maintained at 22℃ and 18℃ in day and night conditions, respectively. Approximately 8-week old plants were transplanted into ~2L pots and pruned apically at ~1cm above the 11th node. At 10 days after transplanting and pruning, any emergent sucker branch >1cm in length was removed and plants were recovered for a further 14 days before the start of the experiment. Third and fourth leaf tissues from the 1$^{st}$ and 2$^{nd}$ apical branches of eight plants were harvested at 11-12 hours after dawn (ZT 11-12) as a time point 0 control (day 0). Pots were all saturated with water at this time by soaking for 5 minutes and for the next 5 days the weight of control group pots was kept constant by adding measured volumes of water every 1-2 days while drought stress pots were not watered at all. The same tissues were

then harvested at ZT 11-12 for time point 1 (day 5) from eight drought-stressed and eight control group plants. The remaining eight control plants were again soaked in water for 5 min and for the next 4 days their weight was again kept constant with measured volumes of water, while eight treatment plants were now supplied with 50 mL water each day. Tissues were then harvested at ZT 11-12 for time point 2 (day 9). Tissue was always pooled for two plants, hence four pools of two plants each representing eight different plants were analyzed. Tissues from all time points were first ground with liquid $N_2$ and stored at -80℃ until further use. RNA was extracted from approximately 50 mg of tissue using a Zymo DirectZol Kit. Approximately 200 mg of the same ground tissue was used for metabolomic profiling (Lisec *et al.*, 2006) (Giavalisco *et al.*, 2009).

**Metabolite Analysis**

Secondary metabolites were profiled by the Waters Acquity UPLC system coupled to the Q Exactive Orbitrap mass detector according to the previously published protocol (Giavalisco *et al.*, 2009). The UPLC system was equipped with a HSS T3 C18 reversed-phase column (100 × 2.1 mm i.d., 1.8-µm particle size; Waters) that was operated at a temperature of 40°C. The mobile phases consisted of 0.1% formic acid in water (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B). The flow rate of the mobile phase was 400 µL/min, and 2 µL of sample was loaded per injection. The UPLC was connected to a Q Exactive Orbitrap (Thermo Fisher Scientific) via a heated electrospray source (Thermo Fisher Scientific). The spectra were recorded using full scan mode of negative ion detection, covering a mass range from m/z 100 to 1500. The resolution was set to 25,000, and the maximum scan time was set to 250 ms. The sheath gas was set to a value of 60, while the auxiliary gas was set to 35. The transfer capillary temperature was set to 150°C, while the heater temperature was adjusted to 300°C. The spray voltage was fixed at 3 kV, with a capillary voltage and a skimmer voltage of 25 and 15 V, respectively. MS spectra were recorded from minute 0 to minute 19 of the UPLC gradient. Molecular masses, retention time, and associated peak intensities were extracted from the raw files using RefinerMS (version 5.3; GeneData), and Xcalibur software (Thermo Fisher Scientific). Metabolite identification and annotation were performed using standard compounds, data dependent method (ddMS2) fragmentation, literature and metabolomics databases (Alseekh *et al.*, 2021).

**RNA Sequencing**

Extracted RNA was treated with Epicentre Baseline-ZERO DNAse. Sample quality and quantity were checked on a Nanodrop and a 1% agarose gel. Within each group, samples were combined

in pools of 2 before sequencing library preparation. Libraries were prepared using the KAPA stranded RNAseq reagents and sequenced on an Illumina NextSeq at Oklahoma State University.

**Genome sequencing and assembly**

A single *S. lycopersicoides* LA2951 plant was chosen for all genome sequencing efforts because this species is self-incompatible and highly heterozygous (Albrecht *et al.*, 2010). DNA was extracted from young fresh leaves as previously described in Bolger et al (A., Bolger *et al.*, 2014) and sequenced using PacBio P6C4 chemistry at Weill Cornell. In addition, one Illumina Nextera library, three Illumina TruSeq PCR-based, and two Illumina PCR-free libraries were generated and sequenced on an Illumina MiSeq at Research Center Jülich following standard Illumina protocols. The PacBio sequence was assembled using Canu (Koren *et al.*, 2017) (version 1.3, with the parameter "genomeSize=1.3g") and Falcon (Chin *et al.*, 2016) assemblers (Table 1). The contigs assembled by Canu were polished using Quiver (https://github.com/PacificBiosciences/GenomicConsensus) with PacBio reads followed by three iterative rounds of correction using Pilon (Walker *et al.*, 2014) with Illumina paired-end reads and then scaffolded with Illumina mate-pair sequences using SSPACE (Boetzer *et al.*, 2011). The assembly was submitted to Dovetail Genomics for further scaffolding using Chicago and Hi-C technologies. Scaffolding gaps were filled with PBJelly (English *et al.*, 2012) using PacBio reads. BUSCO (Waterhouse *et al.*, 2017) was used to assess the quality of the assemblies. Heterozygosity was estimated by mapping Illumina reads from genomic DNA to the final assembly with HISAT2 (Kim *et al.*, 2019) and calling SNPs with the GATK (McKenna *et al.*, 2010) pipeline.

**Annotation**

D*e novo* repeats were predicted in the *S. lycopersicoides* genome assembly using RepeatModeler (Smit AFA, 2008-2015). The predicted repeats containing known protein domains were removed and the remaining repeats were then used with RepeatMasker in conjunction with the Repbase library (Bao *et al.*, 2015). For gene prediction, RNA-Seq reads were mapped to the genome with HISAT2 (Kim *et al.*, 2019). Portcullis (Mapleson *et al.*, 2018) and Mikado (Venturini *et al.*, 2018) were used to process the resulting BAM files. PacBio IsoSeq data was also generated from a pool of leaves, fruits, and flowers, and corrected using the Ice pipeline. AUGUSTUS (Hoff and Stanke, 2019) and SNAP (Korf, 2004) were used for ab initio gene predictions, which were integrated with evidence from Iso-Seq transcripts, proteins from Swiss-Prot, and processed RNA-

seq data, to derive the final consensus gene models using the Maker (Campbell, Holt, *et al.*, 2014) pipeline. Functional annotation and classification was performed using the automated Mercator pipeline (Schwacke *et al.*, 2019).

**Repeat analysis**

The genomes of *S. lycopersicoides*, *S. lycopersicum*, and *S. pennellii* were analyzed for LTR retrotransposons using LTRharvest (Ellinghaus *et al.*, 2008) with the parameters "-seqids yes -minlenltr 100 -maxlenltr 5000 -mindistltr 1000 -maxdistltr 20000 -similar 85 -mintsd 4 -overlaps best". The genomes were then analyzed using LTR_FINDER (Xu and Wang, 2007) with parameters "-D 20000 -d 1000 -L 5000 -l 100 -p 20 -C -M 0.85 -w 0". LTR_retriever was then used to filter the LTR-RT candidates using default parameters, except that the neutral mutation rate was set at $1.0x10^{-8}$ using the -u parameter. This neutral mutation rate was selected as it has been used previously for tomatoes (Lin *et al.*, 2014), assuming one generation per year (Beddows *et al.*, 2017).

Candidate miniature inverted-repeat TEs (MITEs) were obtained using MITE-Hunter (Han and Wessler, 2010), with default parameters except for "-P 0.2". Output candidate MITEs were manually checked for their TSDs and TIRs as suggested in the MITE-Hunter manual. The candidate MITEs were also assigned to superfamilies based on best hits obtained by BLAST against the P-MITE database (http://pmite.hzau.edu.cn/download/), with an e-value cutoff of $1e^{-5}$. Any candidates that could not be unambiguously classified in this way were classified as unknowns.

To arrive at a comprehensive characterization of TEs, the genomes were then masked using the repeat libraries generated by LTR_retriever and MITE-Hunter using Repeatmasker. Additional repeats were then identified *de novo* in the genomes using RepeatModeler. These repeats were classified using BLASTx against the Uniprot and Dfam libraries and protein-coding sequences were excluded using the script ProtExcluder.pl (Campbell, Law, *et al.*, 2014). The masked genomes were then re-masked with Repeatmasker and the corresponding repeat libraries generated by RepeatModeler.

We used TEffectR (Karakülah *et al.*, 2019) to determine cases where expression levels of drought-responsive genes, determined through the differential expression analysis conducted

(i.e., DEGs at an FDR cutoff of 0.05), were correlated with expression of neighboring LTR transposable elements (with no covariates and with a Bonferroni-adjusted p-value < 0.05). This analysis was done using DEGs from the comparison of Day 5 samples to Day 0, as well as from Day 9 to Day 0. The counts generated in the drought DEG analysis were used for the gene count levels in TEffectR. For TE expression levels, reads were mapped to the reference genome using hisat2(Kim *et al.*, 2019)) and SAMtools(Li *et al.*, 2009) was used to sort and merge BAM files corresponding to libraries from the same biological samples. For TE expression, multi- and unmapped reads were also removed from BAM files, as suggested previously (Karakülah *et al.*, 2019).

**Immunity gene analysis**

NLR genes in the Heinz 1706 (SL4.0) and *S. lycopersicoides* genomes (v1.1) and transcriptomes (ITAG v4.2 and v1.1) were identified with NLR-annotator (Steuernagel *et al.*, 2020). RLKs and RLPs were selected from TAIR10 according to their protein functional descriptions. DIAMOND (Buchfink *et al.*, 2015) was used to query this protein database with the Heinz 1706 and *S. lycopersicoides* protein sequences using an e-value cutoff of $10^{-10}$. Structural domains were annotated according to Pfam (Mistry *et al.*, 2021) and Phobius (Käll *et al.*, 2007) was used to predict transmembrane domains and signal peptides.

The *Pto/Prf* region was identified in LA2951 and Heinz 1706 (SL4.0) by querying their genome sequences (Hosmani *et al.*, 2019) with the coding sequences of the *Pto* family members and *Prf* from *PrfRG-PtoR* (NCBI: AF220602.1) using BLASTn (Altschul *et al.*, 1990). A 100 kb region encompassing the identified *Pto/Prf* regions in LA2951 and Heinz 1706 was aligned to that in *RG-PtoR* using the nucmer tool in the MUMmer 3.23 package (Kurtz *et al.*, 2004) with default parameters followed by filtering alignments below 95% identity. The specific coordinates of *PtoA*, *Prf*, *Fen* (*PtoB*), *PtoC*, *PtoD*, and *Pto* (*PtoE*) were determined by querying the 100 kb *Pto/Prf* region of LA2951 and Heinz 1706 with the coding sequence from RG-PtoR for each gene. Multiple sequence alignments of genes were performed with MAFFT version 7 (Katoh and Standley, 2013).

The LA2951 introgression lines were obtained from the Tomato Genetics Resource Center (https://tgrc.ucdavis.edu/) and consist of segments of LA2951 in the background of *S. lycopersicum* (cv. VF36). The lines for this study contained an *S. lycopersicoides* introgression

containing the *Pto/Prf* region. Tomato varieties Rio Grande-PtoR (RG-PtoR; *Pto/Pto, Prf/Prf*) and Rio Grande-prf3 (RG-prf3; *prf3/prf3*) were included as the resistant and susceptible controls respectively. RG-Pto11 was also included as a control for the DC3000mut5 experiments. RG-PtoR carries the resistant *S. pimpinellifolium Pto/Prf* haplotype, and RG-prf3 has a nonfunctional *Prf* gene. *Solanum lycopersicum* cv VF36 and *S. lycopersicoides* LA2951 introgression lines were grown in Cornell Plus Mix (0.16 $m^3$ peat moss, 0.34 $m^3$ vermiculite, 2.27 kg lime, 2.27 kg Osmocote Plus15-9-12 and 0.54 kg Uni-Mix 11-5-11; Everris, Israeli Chemicals Ltd). Plants were grown in a greenhouse without supplemental lighting at 24°C day and 22°C nighttime temperatures. Four-week-old plants were vacuum infiltrated with *Pseudomonas syringae* pv. *tomato* strains DC3000 or DC3000ΔΔ (lacking *avrPto* and *avrPtoB*), or DC3000mut5. Strains were grown on King's B medium for 48 hours and suspended in 10 mM $MgCl_2$ for a final titer of $10^5$ cfu/mL for vacuum infiltration. After inoculation, the plants were kept in a growth chamber (24°C day and 20°C night) for seven days (ten days for the DC3000mut5 experiment) before photographing.

**Genome visualizations, synteny, and structural variants**

Circular genome visualizations were generated using Circos (Krzywinski *et al.*, 2009). For *S. lycopersicoides*, gene and repeat densities were calculated by generating 1MB non-overlapping windows and calculating percent coverage for each feature type (either annotated gene features or repeat elements identified in the repeat analysis) using BEDTools (Quinlan, 2014; Quinlan and Hall, 2010). For synteny analysis, alignment of *S. lycopersicoides* to *S. lycopersicum* (SL4.0) and *S. pennellii* LA0716 v 2 (A., Bolger *et al.*, 2014) was conducted with nucmer using the parameters --maxgap=500 --mincluster=100, followed by delta-filter with parameters '-r -q -i 92 -l 8000 and show-coords'.

**Drought expression**

RNA Seq reads were trimmed using Trimmomatic (A., M., Bolger *et al.*, 2014). Pseudo-alignment was performed using Kallisto (100 bootstraps) to estimate the abundance of each target in each library (Bray *et al.*, 2016). The quality of each library was initially examined in FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Clustering and quality control revealed three libraries that were not considered in further testing. Due to the sampling strategy where two branches were sampled from the same pair of source plants, there was some correlation expected between the two branch samples from a given pool within a treatment group.

We first assessed the potential for an effect of the branch by performing likelihood ratio tests described in Sleuth (Pimentel *et al.*, 2017). Including estimation of a branch effect (i.e. ~ time.treatment + branch) provided no significant improvement to the performance of models for any target. Estimated counts were extracted using TxImport from DESeq2 for further processing (Love *et al.*, 2014; Soneson *et al.*, 2015). A minimum expression level across a minimum number of libraries was enforced and normalization factors (edgeR::calcNormFactors) were applied to generate normalized read counts (McCarthy *et al.*, 2012; Robinson *et al.*, 2010). The extent of correlation between branches from the same source plants was estimated using Limmas duplicateCorrelation function and accounted for in all subsequent testing in mixed linear models by blocking on pool ID and supplying the correlation value to lmFit in Limma (Ritchie *et al.*, 2015). Contrasts were evaluated to estimate typical fold change in response to drought at both time points. MapMan bins (Schwacke *et al.*, 2019) were examined for enrichment in up-regulated, down-regulated or differentially expressed targets in a Fisher's exact test on a contingency table per-bin versus a background of all detected targets Adjustment of p-values to correct for false discovery rate was performed within each set (time point * DEG classification) according to the Benjamini Hochberg procedure (Benjamini and Hochberg, 1995).

**Expression Atlas**

The Lycopersicoides Expression Atlas uses the Tomato Expression Atlas (Fernandez-Pozo et al., 2017) as its template. The main code for the expression atlas is available from GitHub (https://github.com/solgenomics/Tea). The Perl scripts used to format and import data into the database are included in this repository. The code used for customizing the Lycopersicoides Expression Atlas is also publicly available (https://github.com/solgenomics/lycopersicoides_ea).

Two datasets were included in the Lycopersicoides Expression Atlas. In order to generate expression values in counts per million (CPM) for display in the atlas, reads from the two datasets were mapped to the reference genome using hisat2(Kim *et al.*, 2015), SAMtools (Li *et al.*, 2009) was used to sort and merge files from the same biological samples where appropriate and stringtie (Pertea *et al.*, 2016) was used to count reads mapped to gene features. The CPM values were generated from the raw counts using the *cpm* function from the edgeR R package (Robinson *et al.*, 2010). For the correlation values in the atlas, the raw reads were first transformed using the *vst* function from the DESeq2 R package (Love *et al.*, 2014), and Spearman correlation values

were calculated from the transformed counts, following the code included in the TEA repository. For generation of both expression and correlation values, genes with no read counts in any samples were excluded.

**Introgression line maps**

RNA-seq reads from the population were first processed with Trimmomatic-0.36 (A., M., Bolger *et al.*, 2014) to trim and filter low-quality reads. Using Bowtie2 (Langmead and Salzberg, 2012) the reads were then aligned to databases of ribosomal RNA (Quast *et al.*, 2013) and plant virus sequences (Zheng *et al.*, 2017) to filter out non-mRNA reads. The reads were then used to call SNPs following the best practices in the Genome Analysis Toolkit documentation (McKenna *et al.*, 2010). Both *S. lycopersicum* (SL4.0) and *S. lycopersicoides* reference genomes were used to establish a set of SNP markers within the population that distinguish between background and introgressed regions. SNPs that did not match either reference genome were filtered out. Only loci that were homozygous for a single parent or heterozygous with one base from each parent were used. This high-density genotype marker map was then converted into a map of introgressions using SNPbinner (Gonda *et al.*, 2019). Again, reference genomes of both *S. lycopersicum* and *S. lycopersicoides* were used to map the introgression boundaries within the context of the domesticated tomato background as well as to define the introgressed segments within *S. lycopersicoides*.

**Carotenoid quantification and analysis**

Carotenoids were extracted and quantified from 200 mg finely ground frozen pericarp tissue of ILs collected from the greenhouse 2015 experiment as previously described (McQuinn *et al.*, 2018). For the 2015/2016 field experiments, the population was first genotyped (Supplemental Tables 6,7), extracted carotenoids were dried at reduced pressure and re-suspended in ethyl acetate containing 50 μg $ml^{-1}$ of diindolylmethane as an internal standard. Samples were quantified using an Ultra Performance Convergence Chromatography (UPC 2) system, equipped with an HSS C 18 SB column (2.1 × 150 mm, 1.8 μm particle size, Waters), as described (Gonda *et al.*, 2019). Logarithm of the odds (LOD) score analysis was carried out in the R environment, using the R/qtl package (Arends et al 2010). Genome scan was performed with a single QTL

model, using linear marker regression method of phenotypes on marker genotypes. Carotenoid QTL analysis was performed using the field 2015 data set.

**CRTLB 3 functional analysis**

CRTL-B3 CDS was obtained from cv. VF36, by PCR using the following primers F: ATGGATACATTGTTGAAAACCC, R: TTCTGTATCCTTTAACAAATTGTTAATCATAG. The amplicon was transferred to expression vector pEXP5-CT/TOPO (Thermo Fisher Scientific), following confirmation by Sanger sequencing, yielding pEXP-CRTLB3. This construct was transformed into the lycopene-accumulating *E. coli* strain, (pEBI) (Ronen *et al.*, 1999). Positive colonies were incubated in 150 mL liquid LB medium supplemented with IPTG and incubated for 40 hours (25°C, 280 RPM). Cells were harvested, followed by carotenoid extraction and analysis as described previously (Chayut *et al.*, 2017).

Data availability

The genome and raw reads are available under NCBI BioProject ID PRJNA727176, PRJNA526255. In addition the genome its annotation and additional data are available through Solgenomics.net (https://solgenomics.net/organism/Solanum_lycopersicoides/genome). Expression profile data are available at the Lycopersicoides Expression Atlas (http://lycopersicoides-ea.sgn.cornell.edu). The genome and its annotations have been integrated into solgenomics.net. Seeds are available from the Tomato Genetics Resource Center where the IL number corresponds to the LA code.

Acknowledgments

This study was supported by the joint ERA CAPS Regulatome project: DFG US98/7-1 (BU); FE552/29-1 (ARF); BBSRC BB/N005023/1 (CRM); National Science Foundation IOS-1539831 (JJG and ZF); and grants from National Science Foundation IOS-1546625 (GBM, SRS, and ZF); National Science Foundation IOS-1855585 (JJG and ZF); and the BTI Triad Foundation (SRS, AF, SRH, LAM, JJG, and GBM). SA and ARF were additionally supported by funding from the Max- Planck- Society and the European Union project PlantaSyst (SGA- CSA No. 664621 and No. 739582 under FPA No. 664620). BU was supported by BMBF 031A536C. We thank Brian Bell and Steve McKay for plant care, Dr. Theodore Thannhauser, Dr. Yaakov Tadmor, Dr. Mwafaq Ibdah, and Ayala Meir for help with carotenoid experiments, Dr. Surya Saha and Dr. Naama Menda for help with data handling and loading to Solgenomics.net, Dr. Yanna Shi, Jonathan M

Peralta and Jonata Freschi for help with field trials and molecular assistance, Dr. Roger Chetelat for helpful suggestions on the manuscript. Finally, we want to acknowledge the Tomato Genetics Resource Center and Dr. Chetelat for making the introgressions lines available and Rudolf Thomann for donating LA2951 to the Tomato Genetics Resource Center.

Contributions

BU and SRS managed the project. AFP, AF, JL, CM, LC, JJG, BU, MM, and SRS wrote the manuscript with help from all authors. AFP, AF, MHWS, AF, JL, CM, EMJ, SRH, GBM, ZF, JJG, BU, MM and SRS designed the analysis. AF, MHWS, AV, EMJ, KD, MM, AD, YX conducted DNA and RNA preparation and sequencing and plant experiments. AFP, LC, MHS, DL, AD contributed new reagents and analytical tools. SA and ARF conducted metabolic analysis and interpretation. AFP, LC, EMJ, LM, MM, AKD, GBM, ZF, SRS, BU, conducted the data analyses.

Conflict of interest statement

The authors of this article have no conflict of interest to declare.

**Supplemental Figures**

**Supplemental Figure 1 Insertion age of repetitive elements.** The figure shows estimated ages for Copia, Gypsy and other repetitive elements for *S. lycopersicoides* (A), *S. lycopersicum* (B) and *S. pennellii* (C).

**Supplemental Figure 2 Detailed IL map for the *S. lycopersicoides* X *S. lycopersicum* VF36 introgression population**

**Supplemental Figure 3 Syntenic dot plot showing the alignment between *Solanum lycopersicoides* and *Solanum lycopersicum***

The main figure shows the genomic dot plot and two breakouts show regions for chromosome 10 and chromosome 4, respectively.

**Supplemental Figure 4: Mapping efficiency is influenced by parental genome reference**

Overall mapping efficiency of VFf36 (domestic reference parent), *S. lycopersicoides* introgression lines (il, n=12), and *S. lycopersicoides* (slyd) RNAseq reads when mapped to the *S. lycopersicum* reference genome SL4.0. **b,** The percent of reads which fail to map to SL4.0, but successfully map to the *S. lycopersicoides* reference genome (SLYD) **c,** Diagram of line-specific reference

genome grafting using biparental single nucleotide polymorphic (SNP) alignments. **d,** Overall performance of SL4.0 and IL-specific grafted reference assemblies in mapping il reads. **e,** Percent of reads which fail to map to either SL4.0 or the IL-graft, but successfully map to the SLYD reference **f,** Percent of residual (non-SL4.0 or non-IL-graft mapping) reads which map to introgressed segments of SLYD. Error bars represent the standard error of each set: n=6 for vf36 reads, n=4 for slyd reads, n=12 for introgression lines (each line represents an average of at least 3 replicate libraries). One-way ANOVA with Tukey's HSD correction for multiple comparisons was used to determine significance ($p \leq 0.05$).

**Supplemental Figure 5 Selected phenotypes of different IL fruits**. Fruits are shown at breaker +7 stage from a greenhouse experiment.

**Supplemental Figure 6 Fifteen loci (IL bins) affecting carotenoid levels.** Specific carotenoids in a QTL analysis are indicated on the left of each numbered section. The left part of each sub-panel shows carotenoid levels among all genotypes in the IL population. For each assessed carotenoid, the right subpanel shows LOD scores throughout the specified chromosome. In some cases, positions of specific candidate genes are indicated. Blue bars indicate the position of specific ILs associated with the trait. Red marks on the x-axis indicate the location of the bin. AA (VF36 homozygous), AB (heterozygous), BB (*S. lycopersicoides* homozygous).

**Supplemental Figure 7 Expression of candidate genes in genetic BINs.** Expression data of selected candidate genes are shown in the ILs depending on the genotype.

**Supplemental Figure 8 The role of CRTL-B3.** A. Lycopene accumulating *E. coli* cells. B. Lycopene accumulating *E. coli* cells transformed with CmCRTL-B3. C-D. chromatograms of the elution profiles at 450 nm of A-B respectively. E-F. lycopene and beta-carotene absorbance spectra respectively. G. Lycopene/beta carotene ratio in ILs harboring *S. lycopersicoides* CRTL-B3 allele from field 2015 experiment.

**Supplemental Figure 9: CRTL-B3 diversification within the Solanaceae family.** Phylogenetic tree of the lycopene cyclase gene family proteins: CRTL-E (in grey**,** lycopene-epsilon cyclase), CRTL-B (in black lycopene-beta cyclase). Pp (*Physcomitrella patens*), Aco (*Ananas comosus*), Zm (*Zea mays*), Os (*Orysa sativa*), Migut (*Mimulus guttatus*), DCAR (*Daucus carota*), Peaxi (*Petunia axillaris*), Nitab (*Nicotiana tabacum*), CA (*Capsicum annuum*), Sd (S. lycopersicoides), Sl (*S. lycopersicum*), St (*S. tuberosum*), SMEL (*S. melongena*), At (*Arabidopsis thaliana*), Vv (*Vitis vinifera*), Cs (*Citrus sinensis*), Md (*Malus domestica*), Cm (*Cucumis melo*). The occurrence of two CRTL-B homologs in the genome is conserved between Embryophytes (*Physcomitrella*, black) to Angiosperms (*Ananas*; CRTL-B1/2- pink/green), Except grasses (corn, rice) and Arabidopsis, which lost the CRTL-B2 orthologue. The Asterids: wild carrot, Seep monkeyflower, and wild white petunia exhibit a single CRTL-B1 orthologue, while the transition between petunia and tobacco is accompanied with duplication of this gene) red asterisk), which was further diversified into two separate clades in tomato, pepper, eggplant, and potato (CRTL-B1/3 –red/blue).

**Supplemental Figure 10: Evolution of lycopene accumulating fruit is accompanied by silencing of *Sl*CRTL-B3 transcription during fruit ripening.** While maintaining stable expression in both leaf, flower, and ripe fruit of *S. lycopersicoides*, expression of *CRTL-B3* in orange stage fruit of 397 lycopene accumulating accessions is downregulated. SP -*S. pimpinellifolium* (27 accessions), SLC - S. *lycopersicum var cerasiforme* (110 accessions), SLL - *S. lycopersicum L.* (258 accessions). In addition (not in the graph) *S. cheesmaniae* (LA0429): 0.04 RPKM, *S. galapagense* (LA0528): 0.23 RPKM.

**Supplemental Figure 11: IL population polyphenol compounds mQTLs** in chromosome 5 (A), and chromosome 10 (B). By each compound: on the right LOD scores throughout the specified chromosome, on the left is compound accumulation in the corresponding bin, specified by allelic variation: AA (homozygous VF36), AB (heterozygous), BB (homozygous *S. lycopersicoides*).

**Supplemental Figure 12: Analysis of LA4276 fruit methanol extracts.** A,C: LC-MS total chromatogram spectra. In black are biological repetitions of control LA4276 segregated out from *S. lycopersicoides* insertion, in red are LA4276 harboring homozygous insertion to Ch.10. A-B: fruit peel, C-D: fruit flesh. B, D: significantly upregulated compounds associated with *S. lycopersicoides* introgression. E – total phenolic content estimation performed by O.D.300 in comparison to ferulic acid standard. On the right peel, on the left fruit flesh of independent fruit out of segregating IL7276 population. In blue VF36 cultivated allele. In green are fruit derived from heterozygous plants, and in yellow are homozygous to *S. lycopersicoides* insertion.

**Supplemental Figure 13: An IL containing the *Pto/Prf* region of LA2951 is susceptible to DC3000mut5**. Tomato plants were vacuum infiltrated with $10^5$ cfu/mL of *Pseudomonas syringae* pv. *tomato* strain DC3000mut5 and representative leaves were photographed 10 days later. DC3000mut5 carries a version of AvrPtoB lacking its C-terminal E3 ligase domain. The truncated AvrPtoB is recognized by Pto and by Fen. In RG-pto11 *Pto* is nonfunctional and the resistance in this line to DC3000mut5 is due to Fen. RG-prf3 lacks a functional *Prf* gene which disables both Pto and Fen. IL4284, carrying the *Pto/Prf* region of LA2951, is susceptible to DC3000mut5 which was expected since its *Prf* gene has multiple mutations indicating a loss of function.

**Supplemental Tables**

**Supplemental Table 1: Gene model statistics**

**Supplemental Table 2: Repetitive Elements in *S. lycopersicoides*, *S. lycopersicum,* and *S. pennellii***

**Supplemental Table 3: Insertion segregation in the ILs**

**Supplemental Table 4: Detailed IL map including coordinates**

**Supplemental Table 5: Introgression Bin Map created by SNPBinner, mapped to *S. lycopersicoides* genome.**

**Supplemental Table 6: Fruit carotenoid accumulation, Subsets i/iii and ii/iii - which was further used for QTL analysis.**

**Supplemental Table 7: Fruit carotenoid accumulation field 2016 iii/iii line subset.**

**Supplemental Table 8: IL BINs after RNAseq SNP calling**

**Supplemental Table 9: LC-MS peak area of polyphenol compounds accumulation** comparison between cultivated control ripe fruit (VF36, marked in red) and *S. lycopersicoides* ripe green fruit (Lyc, green).

**Supplemental Table 10: Comparison of polyphenol biosynthesis structural genes expression** between VF36 ripe fruit and *S. lycopersicoides* ripe green fruit

**Supplemental Table 11: Polyphenols accumulation in ripe fruit across the IL population.**

**Supplemental Table 12: Immunity gene survey for NLR, RLK and RPL genes in *S. lycopersicoides*, *S. lycopersicum,* and *S. pennellii***

**Supplemental Table 13: Immunity-associated genes in the *S. lycopersicoides* genome.**

**Supplemental Table 14: Enrichment and Depletion Analysis for drought stress Time Points 1 and 2 using MapMan Bins as classes**

References

**Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G.B., Tanksley, S.D. and Giovannoni, J.J.** (2005) Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell*, **17**, 2954–2965.

**Albrecht, E. and Chetelat, R.T.** (2009) Comparative genetic linkage map of Solanum sect. Juglandifolia: evidence of chromosomal rearrangements and overall synteny with the tomatoes and related nightshades. *Theor. Appl. Genet.*, **118**, 831–847.

**Albrecht, E., Escobar, M. and Chetelat, R.T.** (2010) Genetic diversity and population structure in the tomato-like nightshades Solanum lycopersicoides and S. sitiens. *Ann. Bot.*, **105**, 535–554.

**Ali, M.Y., Sina, A.A.I., Khandker, S.S., Neesa, L., Tanvir, E.M., Kabir, A., Khalil, M.I. and Gan, S.H.** (2020) Nutritional Composition and Bioactive Compounds in Tomatoes and Their Impact on Human Health and Disease: A Review. *Foods*, **10**.

**Alonge, M., Wang, X., Benoit, M., et al.** (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145–161.e23.

**Alseekh, S., Aharoni, A., Brotman, Y., et al.** (2021) Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods*, **18**, 747–756.

**Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J.** (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

**Bao, W., Kojima, K.K. and Kohany, O.** (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

**Bauchet, G. and Causse, M.** (2012) Genetic diversity in tomato (Solanum lycopersicum) and its wild relatives. In *Genetic Diversity in Plants*. InTech.

**Beddows, I., Reddy, A., Kloesges, T. and Rose, L.E.** (2017) Population Genomics in Wild Tomatoes-The Interplay of Divergence and Admixture. *Genome Biol. Evol.*, **9**, 3023–3038.

**Bedinger, P.A., Chetelat, R.T., McClure, B., et al.** (2011) Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. *Sex. Plant Reprod.*, **24**, 171–187.

**Belser, C., Istace, B., Denis, E., et al.** (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants*, **4**, 879–887.

**Benjamini, Y. and Hochberg, Y.** (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

**Blanca, J., Montero-Pau, J., Sauvage, C., et al.** (2015) Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics*, **16**, 257.

**Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W.** (2011) Scaffolding pre-

assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.

**Bolger, A.M., Lohse, M. and Usadel, B.** (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

**Bolger, A.M., Poorter, H., Dumschott, K., et al.** (2019) Computational aspects underlying genome to phenome analysis in plants. *Plant J.*, **97**, 182–198.

**Bolger, A., Scossa, F., Bolger, M.E., et al.** (2014) The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nat. Genet.*, **46**, 1034–1038.

**Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L.** (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

**Brog, Y.M., Osorio, S., Yichie, Y., Alseekh, S., Bensal, E., Kochevenko, A., Zamir, D. and Fernie, A.R.** (2019) A Solanum neorickii introgression population providing a powerful complement to the extensively characterized Solanum pennellii population. *Plant J.*, **97**, 391–403.

**Buchfink, B., Xie, C. and Huson, D.H.** (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

**Campbell, M.S., Holt, C., Moore, B. and Yandell, M.** (2014) Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics*, **48**, 4.11.1–39.

**Campbell, M.S., Law, M., Holt, C., et al.** (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–524.

**Canady, M.A., Ji, Y. and Chetelat, R.T.** (2006) Homeologous recombination in Solanum lycopersicoides introgression lines of cultivated tomato. *Genetics*, **174**, 1775–1788.

**Canady, M.A., Meglic, V. and Chetelat, R.T.** (2005) A library of Solanum lycopersicoides introgression lines in cultivated tomato. *Genome*, **48**, 685–697.

**Chayut, N., Yuan, H., Ohali, S., et al.** (2017) Distinct Mechanisms of the ORANGE Protein in Controlling Carotenoid Flux. *Plant Physiol.*, **173**, 376–389.

**Chin, C.-S., Peluso, P., Sedlazeck, F.J., et al.** (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, **13**, 1050–1054.

**Chitwood, D.H., Kumar, R., Headland, L.R., et al.** (2013) A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *Plant Cell*, **25**, 2465–2481.

**Cipollini, M.L. and Levey, D.J.** (1991) Why some fruits are green when they are ripe: carbon balance in fleshy fruits. *Oecologia*, **88**, 371–377.

**Colanero, S., Perata, P. and Gonzali, S.** (2020) What's behind Purple Tomatoes? Insight into the Mechanisms of Anthocyanin Synthesis in Tomato Fruits. *Plant Physiol.*, **182**, 1841–1853.

**Colanero, S., Tagliani, A., Perata, P. and Gonzali, S.** (2020) Alternative Splicing in the

Anthocyanin Fruit Gene Encoding an R2R3 MYB Transcription Factor Affects Anthocyanin Biosynthesis in Tomato Fruits. *Plant Communications*, **1**, 100006.

**Davis, J., Yu, D., Evans, W., Gokirmak, T., Chetelat, R.T. and Stotz, H.U.** (2009) Mapping of loci from Solanum lycopersicoides conferring resistance or susceptibility to Botrytis cinerea in tomato. *Theor. Appl. Genet.*, **119**, 305–314.

**Egashira, H., Kuwashima, A., Ishiguro, H., Fukushima, K., Kaya, T. and Imanishi, S.** (2000) Screening of wild accessions resistant to gray mold (Botrytis cinerea Pers.) in Lycopersicon. *Acta Physiol. Plant*, **22**, 324–326.

**Ellinghaus, D., Kurtz, S. and Willhoeft, U.** (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

**English, A.C., Richards, S., Han, Y., et al.** (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, **7**, e47768.

**Espley, R.V., Brendolise, C., Chagné, D., et al.** (2009) Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples. *Plant Cell*, **21**, 168–183.

**Fantini, E., Falcone, G., Frusciante, S., Giliberto, L. and Giuliano, G.** (2013) Dissection of tomato lycopene biosynthesis through virus-induced gene silencing. *Plant Physiol.*, **163**, 986–998.

**Fray, R.G. and Grierson, D.** (1993) Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. *Plant Mol. Biol.*, **22**, 589–602.

**Gao, L., Gonda, I., Sun, H., et al.** (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.*, **51**, 1044–1051.

**Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B. and Willmitzer, L.** (2009) 13C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.*, **81**, 6546–6551.

**Gonda, I., Ashrafi, H., Lyon, D.A., et al.** (2019) Sequencing-Based Bin Map Construction of a Tomato Mapping Population, Facilitating High-Resolution Quantitative Trait Loci Detection. *Plant Genome*, **12**. Available at: http://dx.doi.org/10.3835/plantgenome2018.02.0010.

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

**Guimarães, R.L., Chetelat, R.T. and Stotz, H.U.** (2004) Resistance to Botrytis cinerea in Solanum lycopersicoides is Dominant in Hybrids with Tomato, and Involves induced Hyphal Death. *Eur. J. Plant Pathol.*, **110**, 13–23.

**Han, Y. and Wessler, S.R.** (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.*, **38**, e199.

**Hekman, J.P., Johnson, J.L. and Kukekova, A.V.** (2015) Transcriptome Analysis in Domesticated Species: Challenges and Strategies. *Bioinform. Biol. Insights*, **9**, 21–31.

**Hoff, K.J. and Stanke, M.** (2019) Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics*, **65**, e57.

**Hosmani, P.S., Flores-Gonzalez, M., Geest, H. van de, et al.** (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. Available at: https://www.biorxiv.org/content/10.1101/767764v1 [Accessed June 2, 2021].

**Ijaz, R., Ejaz, J., Gao, S., Liu, T., Imtiaz, M., Ye, Z. and Wang, T.** (2017) Overexpression of annexin gene AnnSp2, enhances drought and salt tolerance through modulation of ABA synthesis and scavenging ROS in tomato. *Sci. Rep.*, **7**, 12087.

**Ji, Y., Pertuzé, R. and Chetelat, R.T.** (2004) Genome differentiation by GISH in interspecific and intergeneric hybrids of tomato and related nightshades. *Chromosome Res.*, **12**, 107–116.

**Käll, L., Krogh, A. and Sonnhammer, E.L.L.** (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.*, **35**, W429–32.

**Karakülah, G., Arslan, N., Yandım, C. and Suner, A.** (2019) TEffectR: an R package for studying the potential effects of transposable elements on gene expression with linear regression model. *PeerJ*, **7**, e8192.

**Katoh, K. and Standley, D.M.** (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

**Kilambi, H.V., Manda, K., Rai, A., Charakana, C., Bagri, J., Sharma, R. and Sreelakshmi, Y.** (2017) Green-fruited Solanum habrochaites lacks fruit-specific carotenogenesis due to metabolic and structural blocks. *J. Exp. Bot.*, **68**, 4803–4819.

**Kim, D., Langmead, B. and Salzberg, S.L.** (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

**Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L.** (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.

**Klee, H.J. and Tieman, D.M.** (2018) The genetics of fruit flavour preferences. *Nat. Rev. Genet.*, **19**, 347–356.

**Knapp, S. and Peralta, I.E.** (2016) The Tomato (Solanum lycopersicum L., Solanaceae) and Its Botanical Relatives. In M. Causse, J. Giovannoni, M. Bouzayen, and M. Zouine, eds. *The Tomato Genome*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 7–21.

**Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M.** (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

**Korf, I.** (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics.

*Genome Res.*, **19**, 1639–1645.

**Kumar, D., Kumar, R., Baek, D., Hyun, T.-K., Chung, W.S., Yun, D.-J. and Kim, J.-Y.** (2017) Arabidopsis thaliana RECEPTOR DEAD KINASE1 Functions as a Positive Regulator in Plant Responses to ABA. *Mol. Plant*, **10**, 223–243.

**Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L.** (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.

**Langmead, B. and Salzberg, S.L.** (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

**Lee, J.M., Joung, J.-G., McQuinn, R., Chung, M.-Y., Fei, Z., Tieman, D., Klee, H. and Giovannoni, J.** (2012) Combined transcriptome, genetic diversity and metabolite profiling in tomato fruit reveals that the ethylene response factor SlERF6 plays an important role in ripening and carotenoid accumulation. *Plant J.*, **70**, 191–204.

**Li, G., Wang, L., Yang, J., et al.** (2021) A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nat. Genet.*, **53**, 574–584.

**Li, H., Handsaker, B., Wysoker, A., et al.** (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

**Li, J., Liu, L., Bai, Y., Zhang, P., Finkers, R., Du, Y., Visser, R.G.F. and Heusden, A.W. van** (2011) Seedling salt tolerance in tomato. *Euphytica*, **178**, 403–414.

**Lin, T., Zhu, G., Zhang, J., et al.** (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.*, **46**, 1220–1226.

**Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A.R.** (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.*, **1**, 387–396.

**Love, M.I., Huber, W. and Anders, S.** (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

**Mapleson, D., Venturini, L., Kaithakottil, G. and Swarbreck, D.** (2018) Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience*, **7,** giy131

**Martin, C., Butelli, E., Petroni, K. and Tonelli, C.** (2011) How can research on plants contribute to promoting human health? *Plant Cell*, **23**, 1685–1699.

**Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D. and Tanksley, S.D.** (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science*, **262**, 1432–1436.

**Mathieu, J., Schwizer, S. and Martin, G.B.** (2014) Pto kinase binds two domains of AvrPtoB and its proximity to the effector E3 ligase determines if it evades degradation and activates plant immunity. *PLoS Pathog.*, **10**, e1004227.

**Mazo-Molina, C., Mainiero, S., Haefner, B.J., Bednarek, R., Zhang, J., Feder, A., Shi, K., Strickler, S.R. and Martin, G.B.** (2020) Ptr1 evolved convergently with RPS2 and Mr5 to mediate recognition of AvrRpt2 in diverse solanaceous species. *Plant J.*, **103**, 1433–1445.

**Mazo-Molina, C., Mainiero, S., Hind, S.R., et al.** (2019) The Ptr1 locus of Solanum lycopersicoides confers resistance to race 1 strains of Pseudomonas syringae pv. tomato and to Ralstonia pseudosolanacearum by recognizing the type III effectors AvrRpt2/RipBN. *Mol. Plant. Microbe. Interact.,* **32**, 949-960

**McCarthy, D.J., Chen, Y. and Smyth, G.K.** (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

**McKenna, A., Hanna, M., Banks, E., et al.** (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

**McQuinn, R.P., Wong, B. and Giovannoni, J.J.** (2018) AtPDS overexpression in tomato: exposing unique patterns of carotenoid self-regulation and an alternative strategy for the enhancement of fruit carotenoid content. *Plant Biotechnol. J.*, **16**, 482–494.

**Menzel, M.Y.** (1962) PACHYTENE CHROMOSOMES OF THE INTERGENERIC HYBRID LYCOPERSICON ESCULENTUM x SOLANUM LYCOPERSICOIDES. *Am. J. Bot.*, **49**, 605–615.

**Mistry, J., Chuguransky, S., Williams, L., et al.** (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

**Molitor, C., Kurowski, T.J., Fidalgo de Almeida, P.M., et al.** (2021) De Novo Genome Assembly Of Solanum Sitiens Reveals Structural Variation Associated With Drought And Salinity Tolerance. *Bioinformatics*. **14** 1941–1945.

**Muñoz-Gómez, S., Suárez-Baron, H., Alzate, J.F., González, F. and Pabón-Mora, N.** (2021) Evolution of the Subgroup 6 R2R3-MYB Genes and Their Contribution to Floral Color in the Perianth-Bearing Piperales. *Front. Plant Sci.*, **12**, 633227.

**Panno, S., Davino, S., Caruso, A.G., Bertacca, S., Crnogorac, A., Mandić, A., Noris, E. and Matić, S.** (2021) A Review of the Most Common and Economically Important Diseases That Undermine the Cultivation of Tomato Crop in the Mediterranean Basin. *Agronomy*, **11**, 2188.

**Pecker, I., Gabbay, R., Cunningham, F.X., Jr. and Hirschberg, J.** (1996) Cloning and characterization of the cDNA for lycopene beta-cyclase from tomato reveals decrease in its expression during fruit ripening. Plant Mol Biol, 30, 807-819.

**Pedley, K.F. and Martin, G.B.** (2003) Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato. *Annu. Rev. Phytopathol.*, **41**, 215–243.

**Peralta, I.E., Spooner, D.M. and Knapp, S.** (2008) Taxonomy of wild tomatoes and their relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae). *Syst. Bot. Monogr.,* **84**, 1-186

**Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. and Salzberg, S.L.** (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.*, **11**, 1650–1667.

**Pertuzé, R.A., Ji, Y. and Chetelat, R.T.** (2002) Comparative linkage map of the Solanum lycopersicoides and S. sitiens genomes and their differentiation from tomato. *Genome*, **45**, 1003–1012.

**Phills, B.R., Provvidenti, R. and Robinson, R.W.** (1977) Reaction of Solanum lycopersicoides to viral diseases of tomatoes in New York. *Rep. Tomato Genet. Coop. Tomato Genet. Coop.*, **27**.

**Phills, B.R., Robinson, R.W. and Shail, J.W** (1977) Evaluation of Solanum lycopersicoides for resistance to fungal disease and nematodes. *Rep. Tomato Genet. Coop. Tomato Genet. Coop.*, **27**.

**Phills, B.R., Robinson, R.W. and Shail, J.W.** (1977) Evaluation of Solanum lycopersicoides for resistance to fungal disease and nematodes. *Rep. Tomato Genet. Coop. Tomato Genet. Coop.*, **27**, 18–19.

**Pimentel, H., Bray, N.L., Puente, S., Melsted, P. and Pachter, L.** (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods*, **14**, 687–690.

**Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O.** (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–6.

**Quinlan, A.R.** (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–34.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

**Razali, R., Bougouffa, S., Morton, M.J.L., et al.** (2017) The genome sequence of the wild tomato Solanum pimpinellifolium provides insights into salinity tolerance. *bioRxiv*, 215517. Available at: https://www.biorxiv.org/content/10.1101/215517v1 [Accessed April 3, 2019].

**Rick, C.M., Cisneros, P., Chetelat, R.T. and DeVerna, J.W.** (1994) Abg—a gene on chromosome 10 for purple fruit derived from S. lycopersicoides. *Rep. Tomato Genet. Coop. Tomato Genet. Coop.*, **44**, 29–30.

**Rick, C.M., De Verna, J.W., Chetelat, R.T. and Stevens, M.A.** (1986) Meiosis in sesquidiploid hybrids of Lycopersicon esculentum and Solanum lycopersicoides. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 3580–3583.

**Riely, B.K. and Martin, G.B.** (2001) Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene Pto. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 2059–2064.

**Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K.** (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

**Roberts, R., Liu, A.E., Wan, L., Geiger, A.M., Hind, S.R., Rosli, H.G. and Martin, G.B.** (2020) Molecular Characterization of Differences between the Tomato Immune Receptors Flagellin Sensing 3 and Flagellin Sensing 2. *Plant Physiol.*, **183**, 1825–1837.

**Robinson, M.D., McCarthy, D.J. and Smyth, G.K.** (2010) edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

**Ronen, G., Carmel-Goren, L., Zamir, D. and Hirschberg, J.** (2000) An alternative pathway to beta -carotene formation in plant chromoplasts discovered by map-based cloning of beta and old-gold color mutations in tomato. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 11102–11107.

**Ronen, G., Cohen, M., Zamir, D. and Hirschberg, J.** (1999) Regulation of carotenoid biosynthesis during tomato fruit development: expression of the gene for lycopene epsilon-cyclase is down-regulated during ripening and is elevated in the mutant Delta. *Plant J.*, **17**, 341–351.

**Rosebrock, T.R., Zeng, L., Brady, J.J., Abramovitch, R.B., Xiao, F. and Martin, G.B.** (2007) A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. *Nature*, **448**, 370–374.

**Schmidt, M.H.-W., Vogel, A., Denton, A.K., et al.** (2017) De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell*, **29**, 2336–2348.

**Schreiber, M., Stein, N. and Mascher, M.** (2018) Genomic approaches for studying crop evolution. *Genome Biol.*, **19**, 140.

**Schwacke, R., Ponce-Soto, G.Y., Krause, K., et al.** (2019) MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant*, **12**, 879–892.

**Schwizer, S., Kraus, C.M., Dunham, D.M., Zheng, Y., Fernandez-Pozo, N., Pombo, M.A., Fei, Z., Chakravarthy, S. and Martin, G.B.** (2017) The Tomato Kinase Pti1 Contributes to Production of Reactive Oxygen Species in Response to Two Flagellin-Derived Peptides and Promotes Resistance to Pseudomonas syringae Infection. *Mol. Plant. Microbe. Interact.*, **30**, 725–738.

**Smit AFA, H.R.** (2008-2015) RepeatModeler Open-1.0. Available at: http://www.repeatmasker.org.

**Smith, J.E., Mengesha, B., Tang, H., Mengiste, T. and Bluhm, B.H.** (2014) Resistance to Botrytis cinerea in Solanum lycopersicoides involves widespread transcriptional reprogramming. *BMC Genomics*, **15**, 334.

**Soneson, C., Love, M.I. and Robinson, M.D.** (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.*, **4**, 1521.

**Steuernagel, B., Witek, K., Krattinger, S.G., et al.** (2020) The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. *Plant Physiol.*, **183**, 468–482.

**Strickler, S.R., Bombarely, A., Munkvold, J.D., York, T., Menda, N., Martin, G.B. and Mueller, L.A.** (2015) Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. *PeerJ*, **3**, e793.

**Sun, X., Jiao, C., Schwaninger, H., et al.** (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.*, **52**, 1423–1432.

**Szinay, D., Wijnker, E., Berg, R. van den, Visser, R.G.F., Jong, H. de and Bai, Y.** (2012) Chromosome evolution in Solanum traced by cross-species BAC-FISH. *New Phytol.*, **195**, 688–698.

**Takei, H., Shirasawa, K., Kuwabara, K., Toyoda, A., Matsuzawa, Y., Iioka, S. and Ariizumi, T.** (2021) De novo genome assembly of two tomato ancestors, Solanum pimpinellifolium and Solanum lycopersicum var. cerasiforme, by long-read sequencing. *DNA Res.*, **28**. Available at: http://dx.doi.org/10.1093/dnares/dsaa029.

**The 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., et al.** (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.*, **80**, 136–148.

**Tohge, T., Scossa, F., Wendenburg, R., et al.** (2020) Exploiting Natural Variation in Tomato to Define Pathway Structure and Metabolic Regulation of Fruit Polyphenolics in the Lycopersicum Complex. *Mol. Plant*, **13**, 1027–1046.

**Tomato Genome Consortium** (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

**Venturini, L., Caim, S., Kaithakottil, G.G., Mapleson, D.L. and Swarbreck, D.** (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience*, **7**, giy093

**Vijay, N., Poelstra, J.W., Künstner, A. and Wolf, J.B.W.** (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–634.

**Walker, B.J., Abeel, T., Shea, T., et al.** (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

**Wang, X., Gao, L., Jiao, C., et al.** (2020) Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nat. Commun.*, **11**, 5817.

**Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M.** (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543-548

**Willits, M.G., Kramer, C.M., Prata, R.T.N., De Luca, V., Potter, B.G., Steffens, J.C. and Graser, G.** (2005) Utilization of the genetic resources of wild species to create a nontransgenic high flavonoid tomato. *J. Agric. Food Chem.*, **53**, 1231–1236.

**Wu, S., Lau, K.H., Cao, Q., et al.** (2018) Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nat. Commun.*, **9**, 4580.

**Xu, Z. and Wang, H.** (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, **35**, W265–8.

**Yan, S., Chen, N., Huang, Z., Li, D., Zhi, J., Yu, B., Liu, X., Cao, B. and Qiu, Z.** (2020) Anthocyanin Fruit encodes an R2R3-MYB transcription factor, SlAN2-like, activating the transcription of SlMYBATV to fine-tune anthocyanin content in tomato fruit. *New Phytol.*, **225**, 2048–2063.

**Zeng, L., Velásquez, A.C., Munkvold, K.R., Zhang, J. and Martin, G.B.** (2012) A tomato LysM receptor-like kinase promotes immunity and its kinase activity is inhibited by AvrPtoB. *Plant J.*, **69**, 92–103.

**Zhao, L., Qiu, C., Li, J., Chai, Y., Kai, G., Li, Z., Sun, X. and Tang, K.X.** (2005) Investigation of disease resistance and cold tolerance of Solanum lycopersicoides for tomato improvement. *HortScience*, **40**, 43–46.

**Zheng, Y., Gao, S., Padmanabhan, C., et al.** (2017) VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology*, **500**, 130–138.

**Zhong, S., Joung, J.-G., Zheng, Y., Chen, Y.-R., Liu, B., Shao, Y., Xiang, J.Z., Fei, Z. and Giovannoni, J.J.** (2011) High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.*, **2011**, 940–949.

**Zhou, Y., Zhu, J., Tong, T., Wang, J., Lin, B. and Zhang, J.** (2019) A statistical normalization method and differential expression analysis for RNA-seq data between different species. *BMC Bioinformatics*, **20**, 163.

Figures

**Figure 1 Circos diagram of the 12 *Solanum lycopersicoides* chromosomes.**

(A) Gene densities. (B) Repeat densities. Densities are calculated for 1-Mb non-overlapping windows and range from 0% to 100%. Different colors are used for values falling within 0% to 25%, 25% to 50%, 50% to 75%, and 75% to 100% ranges.

**Figure 2 Introgressed segments in the *Solanum lycopersicoides* introgression lines on the genome of *Solanum lycopersicoides***. Green indicates homozygous introgressed region and blue indicates heterozygous region. Detailed maps for individual chromosomes are depicted in Supplemental Figure 4

**Figure 3 Transcriptional regulation of carotenoid flux in cultivated tomato and non-lycopene accumulating wild tomato fruit. A**. Carotenoid biosynthesis pathway with pathway genes indicated in blue. **B**. Gene expression (RPKM) in ripe fruit of lycopene accumulating *S. lycopersicum* (cultivated) and *S. pimpinellifolium* (wild progenitor of cultivated tomato) and non-lycopene accumulating wild species *S. lycopersicoides* and *S. pennellii.* Fold changes in fruit versus leaves are shown in parenthesis. Background color represents differences of log ratio of fold change. The gene list in B) follows the pathway as represented in A, except CRTL-B three homologs which are marked by asterisks.

**Figure 4 Analysis of *Aubergine* Locus.**

A) Phylogenetic tree of subgroup 6 R2R3MYB genes from *S. lycopersicum* (Sl), *S. lycopersicoides* (Slyd), *S. pennellii* (Sp), *S. tuberosum* (St), *P. axillaris* (Pa), *P. hybrida* (Ph) and *P. inflata* (Pi). The maximum-likelihood method was applied with 1000 replicate bootstrap support, following MUSCLE alignment of corresponding genomic sequences. B) Syntenic analysis of the R2R3MYB subgroup 6 gene clusters in the genus Solanum. C) Identity between the *AN2-like2* genes from *S. tuberosum, S. lycopersicum, S. pennellii* and *S. lycopersicoides.* Black indicates identical nucleotides, white indicates difference. D) Transcript levels of *SlydAN2*, *SlydAN2-like* and *SlydAN2-like2* in fruit of *S. lycopersicoides.* E) Comparison of promoter sequences of *SlAN2-like* in WT tomato, *Aft* locus in tomato, *ScAN2-like* in *Solanum chilense* and *AN2-like* in *S. lycopersicoides* candidate for *Abg* locus. Red boxes show sequences common to fruit-making anthocyanin but absent in WT tomato gene.

Figure 5 **Polyphenol accumulation across the IL individuals underlying the mQTLs**.
Data are shown as a heatmap as a $\log_2$ fold change relative to the VF36 control.

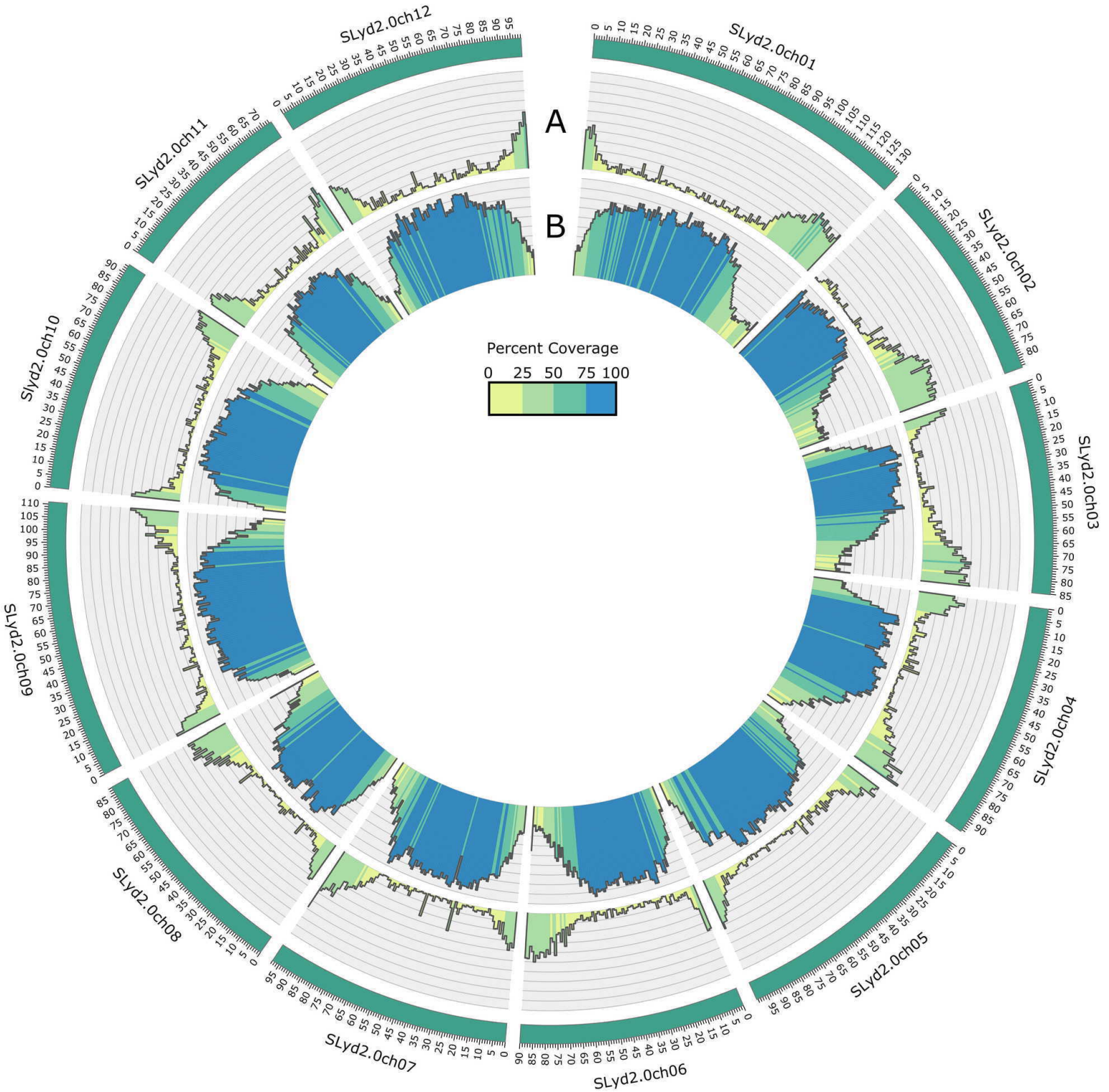
**Figure 6 LA2951 does not recognize AvrPto or AvrPtoB in *Pseudomonas syringae* pv. *tomato*. A)** Responses of introgression lines (ILs) having the *Pto/Prf* region from LA2951 to *Pseudomonas syringae* pv. *tomato* (*Pst*) DC3000 and DC3000Δ*avrPto*Δ*avrPtoB* (DC3000ΔΔ). Plants were vacuum infiltrated with $10^5$ cfu/mL of the *Pst* strains and photographs of representative leaves were taken 7 days post-inoculation. Areas in the red boxes are shown in close-ups to better visualize disease symptoms. Controls are Rio Grande-PtoR (RG-PtoR) which expresses *Pto/Prf* and is resistant to DC3000 but susceptible to DC3000ΔΔ, and RG-prf3 and

RG-pto11 which have mutations in the *Prf* and *Pto* genes, respectively, and are susceptible to both *Pst* strains. The three IL lines having the *Pto/Prf* region from LA2951 were susceptible to DC3000 indicating that LA2951 lacks a functional *Pto* and/or *Prf* gene. **B)** Genome organization of the *Pto/Prf* region in LA2951, RG-PtoR and Heinz 1706. The open reading frames and orientation of the *Pto* gene family members and *Prf* are shown. The coordinates correspond to the genome sequence of each accession with the first nucleotide of *PtoA* set to 0 in each case. Note that *Fen* is also referred to as *PtoB* and *Pto* is also referred to as *PtoE(Riely and Martin, 2001)*. *In LA2951 *Pto* is not present and *Prf* and *PtoD* each contain multiple mutations including one in each that leads to a premature stop codon.

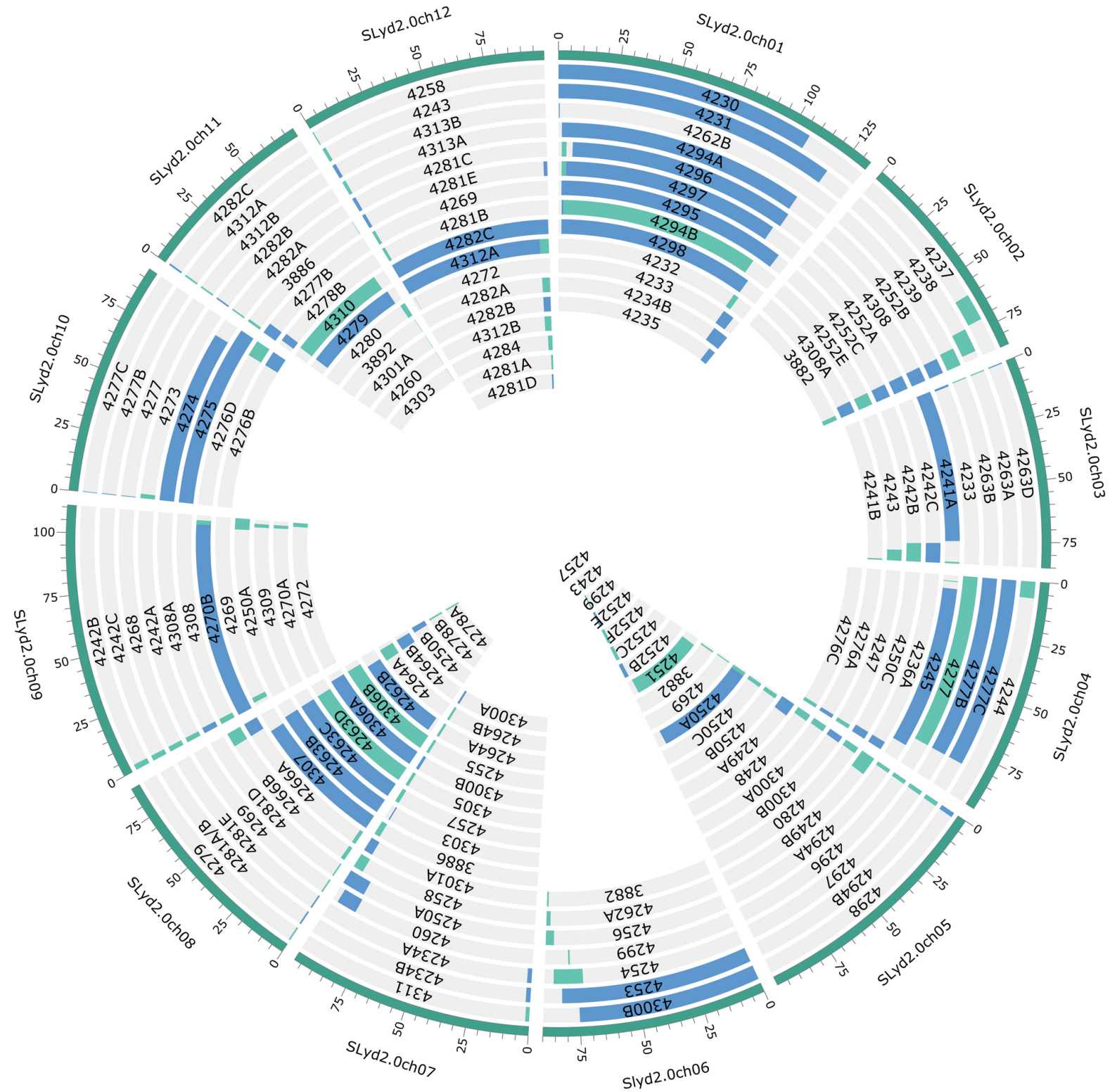
Tables

Table 1 shows the summary statistics for the *S. lycopersicoides* assembly, as well as numbers for 'Heinz 1706' SL4.0 and *S. pennellii* 'LA0716'.

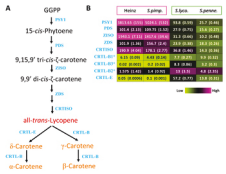
| | ***S. lycopersicoides*** | ***S. pennellii* v2** | ***S. lycopersicum* Heinz v 4.0** |
|---|---|---|---|
| No. of pseudomolecules | 12 | 12 | 12 |
| longest sequence (Mbp) | 133.5 | 109.3 | 90.9 |
| Contig N50 (bp) | 253,764 | 60,347 | 6,007,830 |
| total length (Mbp) | 1,152 | 926 | 782.5 |
| expected genome size (Mbp) | 1,200 | 942 | 781 |
| Total size (bp) of unanchored contigs (% of assembly) | 135,089,793 (10.5) | 63,101,713 (6.4%) | 9,643,250 (1.2%) |



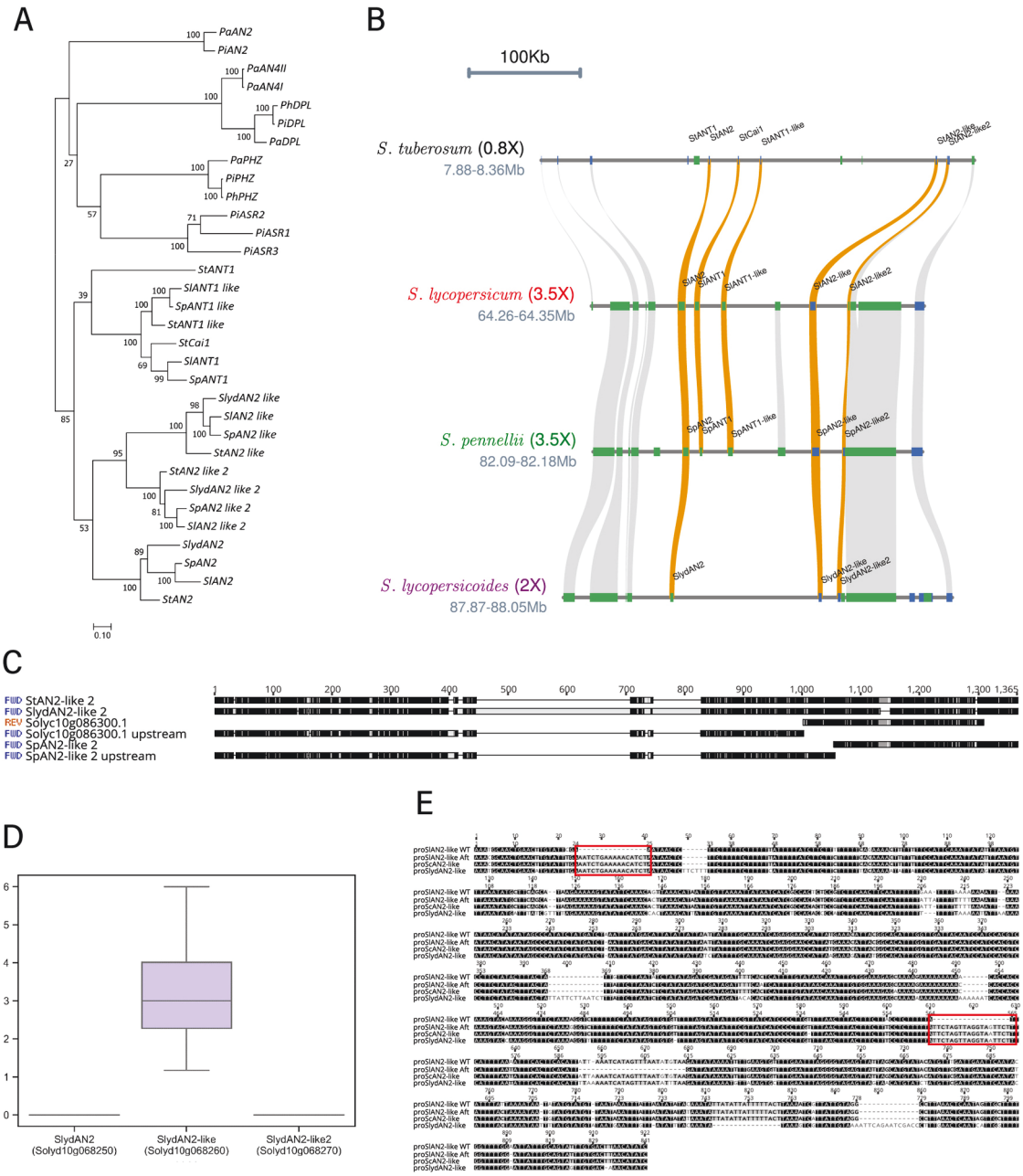

TPJ_15770_Figure1_NEW_Slyd_circos_ColourMod_Final-1-1.jpg



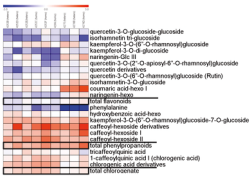

TPJ_15770_Figure_2v2-DOWN.jpg



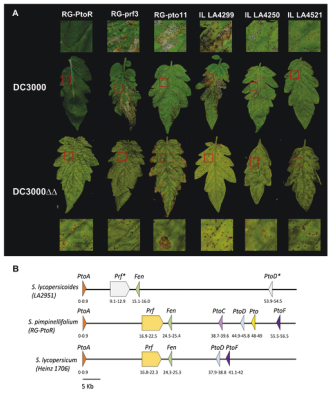

TPJ_15770_Figure_3DOWN.jpg




TPJ_15770_Figure_4.jpg




TPJ_15770_Figure_5DOWN.jpg




TPJ_15770_Figure_6DOWN.jpg