The genome of *Solanum pennellii* and M82

Contents

Nature Genetics: doi:10.1038/ng.3046

Supplementary Note

1 Genome sequencing and assembly

1.1 Plant material

Solanum pennellii LA0716 seeds were obtained from the Tomato Genetics Resource Center, UC Davis, CA, USA. Seeds were sown directly on soil and grown in a greenhouse. For paired-end libraries, young leaves were collected and directly frozen in liquid nitrogen to perform DNA extraction. For mate-pair libraries, young leaves were harvested directly into ice-cold TE (10 mM Tris-HCl, 1 mM ethylenediaminetetraacetic acid (EDTA)) buffer and were processed for DNA extraction immediately afterwards.

1.2 Nucleic DNA extraction for paired-end libraries

0.5 g of young leaves were ground to a fine powder in liquid nitrogen using a mortar and pestle and transferred to a 15 mL polyethylene centrifuge tube containing 10 mL of ice-cold nuclei extraction buffer (10 mM TRIS-HCl pH 9.5, 10 mM EDTA pH 8.0, 100 mM KCl, 500 mM sucrose, 4 mM spermidine, 1 mM spermine, 0.1% beta-mercaptoethanol). The suspended tissue was mixed thoroughly with a wide-bore pipette and filtered through two layers of Miracloth (CalBiochem) into an ice-cold 50 mL polyethylene centrifuge tube. 2 mL Lysis Buffer (10% Triton X-100 in Nuclei extraction buffer) was added to the filtered suspension and mixed gently for 2 minutes on ice. The nuclei were pelleted by centrifugation at 2000 x g for 10 minutes. 500 µL CTAB extraction buffer (100 mM Tris pH 7.5, 0.7M NaCl, 10 mM EDTA, 1% 2-Mercaptoethanol, 1% Cetyltrimethyl Ammonium Bromide (CTAB)) was added to the nuclei pellet, mixed and incubated for 30 min at 60°C. The mixture was extracted with 350 µL Chloroform/Isoamyl Alcohol (24:1). Finally DNA was precipitated by Isopropanol and washed with 75% EtOH several times.

1.3 High molecular weight nucleic DNA extraction for mate-pair libraries

High molecular weight DNA extraction was performed using a modified version of a protocol used to isolate DNA from tomato[1]. Mature plants were stored in darkness for ca. 48 hours before harvesting to reduce starch levels. Between 20 and 100 g of young leaves and flower buds were harvested directly into ice cold TE Buffer (pH 7.0). Plant material was removed from the TE buffer and placed into a pre-chilled Waring blender together with 600 mL of freshly made pre-chilled extraction buffer

(1 M 2-methyl-2,4-pentanediol, 10 mM PIPES, 10 mM $MgCl_2$, 4% (w/v) PVP-10, 10 mM sodium metabisulfite, 25 mM 2-mercaptoethanol, 0.5% (w/v) sodium diethyldithiocarbamate, 200 mM L-lysine and 6 mM ethylene glycol tetraacetic acid (EGTA), pH 6.0) and homogenized for 30 seconds at full speed. The homogenized plant material was squeezed through 4 layers of Miracloth and then further gravity filtered through 8 layers of Miracloth. 10% (v/v) Triton X-100 was added to the homogenate to a final concentration of 0.5% (v/v) and the mixture incubated on an ice bath with gentle rocking for 30 minutes.

The homogenate was then centrifuged at 800 x g for 20 minutes at 4 °C to pellet the nuclei. The supernatant was discarded and the pellet gently resuspended using a brush soaked in freshly made pre-chilled nuclear buffer (0.5 M 2-methly-2,4-pentanediol, 10 mM PIPES, 10 mM $MgCl_2$, 0.5% (v/v) Triton X-100, 10 mM sodium metabisulfite, 60 mM 2-mercaptoethanol, 200 mM L-lysine, and 6 mM EGTA, pH 7.0). Since the pellet contains large amounts of cellular debris, the following washing process was applied three times. The volume was brought up to 10 mL with the same nuclear buffer and the resuspended pellet gently mixed. The mixture was centrifuged at 600 x g for 20 minutes at 4 °C and the supernatant discarded. After three washing cycles, the pellet became grey-white and no traces of green were visible. After the final centrifugation step, the pellet was resuspended using a small paint brush soaked in nuclear buffer and brought up to 5 mL using the same nuclear buffer. 20% (w/v) SDS was added to a final concentration of 2% (w/v) to lyse the nuclei. The contents were gently mixed by inverting the tube.

The tube was then heated in a water bath at 60 °C for 10 minutes and allowed to cool to room temperature. 5 M sodium perchlorate was added to a final concentration of 1 M and the tube gently inverted. The mixture was centrifuged at 400 x g for 20 minutes at room temperature. The supernatant was gently removed to a new tube using wide bore pipette tips. These tips were used for the remainder of the extraction procedure to minimize the shearing of the DNA.

DNA extraction was performed by adding an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) to the tube. The tube was then placed horizontally on a rocker and gently rocked for 30 minutes. The mixture was centrifuged at 3000 x g for 10 minutes at room temperature and the upper phase transferred to a new tube. A second phenol/chloroform/isoamyl alcohol extraction was performed followed by two extractions using only chloroform. The upper phase was transferred to a new tube and 3 M sodium acetate added to a final concentration of 0.3 M. The tube was gently inverted several times to mix. The DNA was precipitated by adding 2 volumes of ice cold ethanol and the tube inverted several times. The precipitated DNA was transferred to a clean tube using a glass rod and the ethanol was allowed to evaporate. The DNA was dissolved in either TE buffer or DNAase free water and the concentration measured using Nanodrop (Thermo Fischer, Schwerte, Germany) and Qubit 2.0 (Invitrogen, Karlsruhe Germany).

1.4 Illumina sequencing of paired end libraries

For the small insert libraries, the genomic DNA was sheared to 200-500 b fragments using the Covaris S2 instrument (Covaris Inc. Massachusetts, USA). The DNA was re-suspended in 1x low TE buffer (Applied Biosystems P/N 4389764). End repair of sheared fragments, addition of an A residue to the 3’ end of blunted fragments, and ligation of adaptors was according to Illumina’s instructions. The entire adaptor-modified DNA was resolved on a 2% agarose gel (including 400 ng/mL Ethidium bromide) run in TAE buffer for 90 minutes at 120 Volts. Fragments of 260, 330 and 600 bp were excised under illumination from a Dark Reader (Clare Chemical Research, Dolores, CO, USA). The DNA was then isolated with a Gel Extraction Kit (Qiagen, Hilden, Germany) and amplified by PCR for 12 cycles with the supplied PCR primers 1.1 and 1.2 (Illumina) and quantified with a Qubit 2.0 fluorometer (Invitrogen, Karlsruhe, Germany). The DNA was diluted and stored at -20˚C as a 10 nM stock in EB buffer (Qiagen, Hilden, Germany) supplemented with 0.1% Tween-20. Validation of the libraries was performed using an Agilent Bioanalyzer by running 1 µL of sample on a DNA 1000 chip (Agilent Technologies, Germany).

1.5 Illumina / hybrid mate-pair libraries

Long mate-pair/paired-end libraries, ranging between 3 kb and 40 kb were prepared from high molecular weight DNA (prepared as described in Section 1.3). In cases where DNA was dissolved in TE buffer, an additional purification step with Agencourt Genfind v2 or Amicon Ultra-0.5mL (Millipore, Schwalbach, Germany) was included.

One 5 kb library was prepared using the Illumina mate pair library preparation kit (Cat. No: 1004876, Illumina) and following the recommendations of Illumina mate pair sample preparation guide (Cat. No: PE-930-1003 REV.A, Nov. 2009).

For longer jumping libraries, a hybrid Roche / Illumina approach was adopted. The DNA was fragmented into the appropriate fragment sizes (3, 5, 8, 20 and 40 kb) using the HydroShear PlusTM DNA Shearing Device (Digilab, USA). The end repaired and cleaned long span fragments were size selected on a 0.5% agarose gel running 17 hours at 35 V overnight. The ends were circularized using the Cre-loxP approach described in the 454 PE Roche protocols (“GS FLX Titanium Paired End Library Preparation Method Manual”, Oct. 2009) with the exception that the circular DNA was sheared using a Covaris S2 instead of nebulization. After the immobilization of the sheared fragments, the samples were processed according to the Illumina protocol for mate pair library sample preparation using the Illumina PE sample preparation kit (Cat#: PE-102-1001). All washing procedures were done three times with 500 µL TE and 1x with Tris-HCL (Qiagen, Hilden, Germany). The fragments were amplified 20 cycles. For final library size selection, Agencourt AMPure beads XP (Beckman Coulter, Krefeld, Germany) were used.

The final mate-pair library was prepared using a fosmid end approach. Genomic DNA was sheared (HydroShear Plus, Digilab) and end-repaired, then ligated into the pNGS FOS vector (NxSeq^{TM} 40 kb Cloning kit, cat. No. 42028-1, Lucigen) and packaged *in vitro* using bacteriophage lambda extract (GigapackIII Gold, Agilent Technologies) and transfected into the *E. coli* Fos strain (Lucigen). The bacteria were plated on agar plates (245mm x 245mm, Corning) and colonies were harvested *en masse.* From the bacteria pool (approximately 2 million colonies) the fosmid DNA was purified and digested with CviQI (New England Biolabs). DNA of 8-9 kb, containing vector and ends of the gDNA inserts, was gel purified, religated and amplified with Illumina PE primers.

1.6 Illumina DNA sequencing

All but the final library preparation were sequenced as follows. Cluster generation was performed on a Cluster Station (cat no. SY-301-2001; Illumina), according to the manufacturer’s instructions. DNA Sequencing was performed on a Genome Analyzer GAIIx (Illumina) using the TruSeq SBS Sequencing Kit v5 according to the manufacturer’s instructions; Sequencing control software was SCSversion2.8 and RTA1.8.7.

For the final library (40 kb fosmid end based), DNA sequencing was performed on a HiSeq2000, equipped with on-instrument HCS version 1.5.15 and real time analysis (RTA) version 1.13. Cluster generation was performed on a cBot (recipe: PE_Amp_Lin_Block_Hyb_v8.0, Illumina) using a flow cell v3 and reagents from TruSeq PE Cluster Kits v3 (Illumina) according to the manufacturer's instructions. Sequencing was performed in paired end mode with 100 bp read length and 25% PhiX spiked-in.

1.7 BAC-end sequencing

A BAC genomic library of *S. pennellii* (LA716) was built in pBeloBACII (52,992 clones) using HindIII partially digested genomic DNA. Average insert size was approximately 120 kb (range 40-320 kb) based on PFGE analyses of 50 random clones [2].

For BAC-end sequencing, 0.2-0.5 μg of purified DNA was used and sequencing reactions were performed with ABI Big Dye Terminator v3.1. Samples were read in an ABI 3730x1 sequencer. Vector and low quality sequences were trimmed out, and only BAC-end sequences larger than 100 b on both ends were considered further.

Since only 10,615 BAC-end pairs passed this threshold, which would provide less than 2x physical coverage of the estimated 1.2 Gb genome (assuming 200 kb insert sizes), this was considered insufficient for inclusion in an automated assembly pipeline. It was decided instead to use them for anchoring and as an independent scaffold validation dataset.

1.8 Filtering and contamination

The paired end libraries were trimmed using Trimmomatic (v0.13)[3], to remove Illumina adapter sequences and low quality bases. For adapter trimming, the TruSeq2 paired-end adapter sequences supplied with Trimmomatic were used with the suggested thresholds. Low quality bases (quality score below 3) were removed from both ends of the reads, then the sliding window trimmer was used to remove low-quality sequence on the 3' end of the reads, using a required average quality score of 15 over 4 bases. Reads shorter than 36 b were dropped. More precisely, the list of trimming steps was as follows:

- ILLUMINACLIP:TruSeq2-PE.fa:2:40:15
- LEADING:3
- TRAILING:3
- SLIDINGWINDOW:4:15
- MINLEN:36

The filtering process for the mate-pair libraries was customized depending on the library preparation protocol. For the first 5Kb library, which was created with the Illumina mate-pair protocol, filtering was as above but with a less stringent sliding window (SLIDINGWINDOW:4:10), and shorter minimum length of 24.

For the hybrid Roche / Illumina libraries, the filtering process was further modified by adding the LoxP linker sequence (TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACG) plus its reverse complement sequence to the adapter file. The presence of this sequence within the reads is an artifact of library preparation which would otherwise prevent the alignment of many valid reads.

For the NxSeq fosmid library, the TruSeq3-PE adapters were used, but the other filtering steps were the same as the other mate pair libraries. An additional step, which removed the first 3 bases of each read (HEADCROP:3) was also added, since these bases come from the fosmid vector, rather than the DNA insert.

Chloroplast and mitochondrial contamination were estimated by alignment of the filtered paired-end libraries using Bowtie[4] (V0.12.7) with the publicly available *S. lycopersicum* chloroplast genome (genbank GI:113531108) and mitochondrial sequences (genbank GI:378405840 – GI:378406034) used as baits.

The paired library statistics are presented as Supplementary Table 1- 3.

1.9 Library insert size estimation

The insert sizes of all libraries were estimated by aligning reads using Bowtie[4] (v0.12.7) against appropriate bootstrap assemblies and compared to the gel based estimations. For the paired-end libraries, the bootstrap assembly used no pairing information and the results are shown in Supplementary Table 4.

For estimating the size of the mate-pair libraries, the paired-end information in the short paired end libraries was used within the bootstrap assembly, to create sufficiently long contigs to allow an unbiased size estimate, and the results are presented in Supplementary Table 5.

To prevent biases, the reads from each pair were independently aligned and only reads with a single alignment were used. Pairs were considered to have valid alignments if they both hit the same contig or scaffold in the appropriate relative orientation and were separated by a distance up to twice the gel based estimate. For mate-pair libraries, pairs with distance estimates below 500bp were also removed.

After the size estimation, the mate-pair libraries were then combined into appropriate groupings (referred to as MP04-MP10), based on length, for scaffolding of the final genome assembly.

1.10 Genome size and coverage estimation

The genome size was estimated using a k-mer counting approach, described by Li et al.[5] All 19-mers in the filtered PE libraries, comprising 1.7 billion reads and 165 billion bases were extracted, yielding 132.5 billion k-mers.

These were combined into a histogram, which revealed the major coverage peak at 107, as shown in Supplementary Figure 1. The 3.4 billion k-mers with coverage below 35 were considered as likely errors. The remaining 129.8 billion k-mers, divided by the coverage peak of 107, indicated a genome size of approximately 1,207 Mb.

Based on this genome size estimate, the raw genome coverage provided by the 229.36Gb of paired-end libraries was estimated to be approximately 190x, while the filtered paired-end data has a coverage of approximately 137x. Since mate-pair libraries often contain sequence artifacts from the library preparation process, and thus should not be used to create the initial contig sequences, the 183.57 Gb of mate-paired sequence data was not included in the coverage estimates.

1.11 Assembly pipeline

The filtered PE data was error corrected using the SOAP error corrector (V1.00), using 8 threads, but otherwise default parameters.

Assembly and initial scaffolding was performed with SOAPdenovo[6] (V1.05), using both the paired-end libraries and the shorter (<10kb) mate pair libraries. The K-mer size was set to 63 (-K 63), and read-repeat resolution (-R), and low coverage graph trimming (-D -d) were enabled. For improved performance, 48 threads (-p 48) were used.

During assembly, some issues were found with the bubble popping algorithm as implemented in SOAPdenovo, which resulted in many short, poorly supported contigs in the output. This issue was diagnosed and corrected, although at the cost of significantly increasing computation time. The code for the bug fix is available on our website (usadellab.org).

The longer (10+ kb) mate-pair libraries were used for scaffolding using SSpace[7] (V1.0) on the output of SOAPdenovo, without contig extension (-x 0), but otherwise default parameters.

After scaffolding, the SOAP GapCloser (V1.12) was used with the most conservative settings available, requiring a 31 base overlap (-p 31), and also longer read support (-l 151). Afterwards short sequences, of less than 2kb, were removed.

1.12 Contaminant removal

After completing the assembly, scaffolds were aligned against the NCBI Non-redundant Nucleotide (NT) database (downloaded on 18th May 2013) using BLASTN (Blast+ Version 2.2.28) with default parameters, except for a minimum e-value cut-off of 1e-10. Multiple hits from each scaffold were combined and the scaffold 'assigned' to the closest reference sequence based on the best combined hit score. The taxonomy data corresponding to this closest reference sequence was used to determine if each scaffold was likely to be of non-plant origin, and thus should be removed. This resulted in the removal of 12 small scaffolds comprising 29,742 bases, leaving 4579 scaffolds in the draft *S. pennellii* genome sequence.

1.13 Sequence validation and correction

After the genome had been assembled the resulting assembly was corrected as follows. Firstly a sample from 12 high-quality libraries with insert size of ~515 b (fragment size average: [502 – 531 bp]; Fragment size standard deviation: [33–64 b]) were mapped to the *S. pennellii* genome assembly. Based on the alignments about 21,279 inconsistency candidates (homozygous SNPs) between the reads and the assembly was detected. As this suggested that the data could be improved, all paired-end reads were mapped back to the assembled scaffolds using BWA, with 97.77% of the reads aligning successfully. Based on this alignment, genomic variants were called using Samtools[8], resulting in 275,260 putative variants.

These variants were filtered using a custom script, eliminating those which had insufficient or very high coverage, as the latter likely arise from repetitive regions, those which were not well supported by both forward and reverse reads, and those for which the existing sequence had adequate support, indicating probable heterozygosity or variation within the sequenced population.

This resulted in 22,019 well supported SNPs between the original read data and the genome assembly, which were applied to correct the assembled genome.

1.13.1 Gap filling validation

In order to assess the correctness of the regions filled by the SOAP GapCloser, the alignment of the paired-end libraries was also determined on the "pre-GapCloser" assembly (also filtered to 2Kbp minimum length). The gap-closed assembly showed a higher overall alignment rate (97.77% vs 95.86%) as well as a higher valid pairing rate (90.83% vs 84.68%), indicating that the gap-filled assembly contains a larger portion of the original read dataset. By plotting the depth of coverage of each non-'N' assembly base from this alignment, ~7.5% more bases can be seen with coverage near the 135x peak, which would be expected as a result of correctly filling gaps (Supplementary Figure 2). Fewer bases are seen with moderately low coverage (5-50 fold), consistent with aligning additional reads near and within the gap filled regions.

However, there are more bases with very low coverage in the gap filled assembly, e.g. 0 to 3 supporting reads. These comprise an additional 550 kb which can be considered as an indication that not all of the gaps are filled correctly, since an incorrect base within the assembly could cause the aligner to place many reads in an alternative location, especially if many very similar alternatives exist (as expected for repetitive regions).

Furthermore, in addition to filling almost 50 Mb worth of gaps, the GapCloser reduced the overall assembly size by approximately 8.8 Mb. This suggests that GapCloser may be biased in favor of filling gaps with shorter sequences when possible, which may result in collapsed tandem repeats in some cases. This fits with the assessment of Boetzer et al[9] using an earlier version of the GapCloser (V1.10) on a human test dataset.

Overall, the evidence indicates that the gap closing process results in a substantially more complete assembly, but it also introduces a modest number of additional errors.

1.14 Final assembly summary and comparison

The final assembly comprised 4579 scaffolds representing 942.6 Mb, which is consistent with the estimated 1,207 Mb genome size determined above. The largest scaffold was slightly larger than 10 Mb and the weighted median size (N50) was 1.7 Mb in 156 scaffolds. The N90 was 437 kb indicating that 90% of the assembly was in units of almost half a Megabase. The whole assembly contained gaps of approximately 67.2 Mb represented by Ns (Supplementary Table 6).

Comparison of the *S. pennellii* assembly against the published *S. lycopersicum* cv. Heinz and *S. tuberosum* genomes reveals that the genome assembly is, despite being assembled only from NGS data, statistically comparable to the *S. tuberosum* genome. *S. pennellii* has a larger size (942Mb vs 715Mb) and moderately higher N50 (1.74 Mb vs 1.35Mb), although it contains a higher percentage of Ns (7.1% vs 6.2%). The *S. lycopersicum* cv. Heinz genome, which applied both clone-based and whole genome shotgun based techniques using a mix of Sanger and NGS technologies, is clearly the most complete assembly, with an N50 of 16.5Mb. The final contig size for each genome, achieved by splitting the assemblies on any N base, again ranks *S. pennellii*, with a contig N50 of 45.7 kb between *S. tuberosum* (31.4 kb) and Heinz (86.9 kb)

1.15 Anchoring and building of pseudomolecules

Out of 13,763 marker sequences available, 6,328 were used as anchors to assign the position of the *S. pennellii* scaffolds onto the EXPEN2000[10,11] genetic map. Markers were gathered from different sources: 4,103 sequences correspond to mapped markers (RFLPs, SSRs, PCRs and COS) downloaded from public ftp sites (www.solgenomics.net); 543 sequences correspond to BAC-end sequences previously mapped[2]; 333 sequences correspond to DarT markers mapped by[12] and 8,784 sequences correspond to SNP markers mapped by[13].

During the initial phase of anchoring using all these available marker sequences, it was possible to localize around 50% of the scaffolds to an approximate location. The addition of BAC-ends to create superscaffolds plus some manual curation of ambiguous marker alignments improved this to ca. 57%, but this was still relatively low. The primary issue was low density of markers in the more central regions of the chromosomes, due to low rates of recombination, but unfortunately this is also where the scaffolds were most difficult to assemble, and thus shortest. Some inconsistent markers where also found, but the data was not sufficient to reliably determine in each case whether the marker or the scaffold was incorrect or where exactly to cut problematic scaffolds.

To overcome these issues, '*de-novo*' markers were generated based on the IL RAD-Seq data from[14] by aligning (with strict settings) all reads against both the *S. lycopersicum* cv. Heinz and cv. M82 genomes, keeping the unaligned reads from these, and then aligning these against the *S. pennellii* scaffolds (also using strict settings). These hits were considered as *S. pennellii* specific sequences, and given that they came from known ILs, this information could be used to indicate the rough origin of each scaffold within the *S. pennellii* genome. Although this signal is moderately noisy, the large number of new markers for each scaffold and the known relationships between the ILs, made it possible to anchor considerably more of the genome to chromosomal regions (although at only the resolution of the ILs).

As part of the anchoring process, it was also possible to identify 118 scaffolds which were likely mis-assembled, which were then split during the anchoring process. A total of 147 breaks were introduced into these 118 scaffolds, equivalent to one mis-assembly per 6.4 Mb of sequence.

For many large scaffolds, and those near the end of chromosomes, it was possible to anchor and orient them with a combination of the traditional and RAD-Seq derived markers. For the remainder, assignment to the chromosome region was done using the traditional and RAD-Seq markers, but the final ordering and orientation of scaffolds within the region was done using synteny to the published *S. lycopersicum* cv. Heinz sequence. However, only locations within the determined chromosome region were considered.

2 Independent genome validation

In addition to the validation based on the paired-end DNA libraries described above, independent BAC, EST, Unigene and RNA-Seq data was used to further validate the correctness and completeness of the genome.

2.1 Base error rate assessment using BACs

2.1.1 Initial genome assessment

Only nine sequenced BACs were available for comparison against the assembled genome (Accession numbers FJ809740.1 to FJ809747.1 and FJ812349.1). Alignment using BLASTN (blast 2.2.28+) could cover 99.07% of these BACs on average, with a combined mismatch plus INDEL rate of less than 0.09%. (See Supplementary Table 9)

Manual inspection of the alignments revealed that many of the incomplete regions were due to gaps (i.e. N filled regions) in the scaffolds, and that a substantial number of the errors were due to the low-quality sequence at the end of the BACs or INDELs in homopolymer containing regions. This might be explained by the fact that most of the BACs were sequenced using 454 technology which is known to be problematic in homopolymeric regions. With this in mind, the numbers given above can be considered as pessimistic estimates of the true genome quality.

For eight of the BACs, the large scale structure of scaffolds and BACs agreed. In one case however (CP020G005), the alignment covered the entire BAC but only ~26% of the scaffold region. Comparison against the *S. lycopersicum* genome indicated that the scaffold was most likely correct, and that the BAC sequence most likely had a large dropout, bordered by a repetitive region.

2.1.2 BAC resequencing

To clarify if the differences were due to BAC sequencing errors, the BACs were resequenced using 2x300bp paired-end libraries using an Illumina MiSeq sequencer as follows. The nine BAC lines were incubated at 37°C in 40 mL of LB medium (with 12.5 µg/mL Chloramphenicol or 11 µg/mL Tetracycline) with shaking overnight. One clone (CP034K014) failed to grow and DNA was extracted from the other eight clones using a modified alkaline lysis protocol. In brief, clones were pelleted at 3000 x g for 45 minutes at 4°C, and subsequently resuspended in 2 mL buffer P1 (50 mM Glucose, 10 mM EDTA pH 8.0, 25 mM Tris-Cl pH 8.0) on ice. After completely resuspension the cells were lysed with 2 mL of solution P2 (1% (w/v) SDS, 0.2 M NaOH) and incubated for 10 minutes in ice. Thereafter the whole solution was neutralized by adding 2 mL of solution P3 (3 M Potassium acetate). The resulting slurry was centrifuged for 1h at 3000 x g and the cleared residue (ca 5 mL) transferred to a new tube. The solution was then extracted with an equal volume of Chlorofom:Isoamylalcohol (24:1) twice before 2 volumes of ice cold absolute Ethanol was added to the aqueous phase. After incubation for 1h at -20°C the DNA was pelleted at 20000 x g for 30 minutes and washed twice with ice cold 70% (v/v) Ethanol and finally resuspended in ultra-pure, nuclease free H_2O. The resulting DNA was used to create an individual library for each line using the TruSeq3 DNA LT kit in accordance with the manufacturer's instructions and sequenced on a MiSeq (Illumina) using a 2x300 bp kit.

2.1.3 BAC correction and updated genome assessment

The new BAC sequence data were filtered using Trimmomatic (v0.32) using the included TruSeq3 PE adapters, and using the Maximum Information quality trimmed (MAXINFO:40.0.3) with all other parameters were chosen exactly as in Section 1.8. The filtered sequences were aligned against the existing BAC sequences (combined with the *E. coli* K-12 Mg1655 genome sequence, genbank id 49175990, as additional bait) with BWA[15] (v0.5.9r16).

Analysis of these alignments indicated that the CP020G005 clone, which was not in structural agreement with the *S. pennellii* assembly, indeed contained large amounts of sequence not represented in the previous BAC assembly. Furthermore, it was established that the BP029K005 line contained no significant sequence commonality to the BAC with the same identifier (nor to any of the other previously sequenced BACs), suggesting mis-identification had occurred at some point.

For the remaining lines, the assembled BAC sequence was corrected based on these alignments. The Short Read Micro Aligner[16] (v0.1.15) was used to resolve local ambiguity caused by conflicting alignments. The resulting resolved alignment was used to call variants using Samtools (v0.1.18) and these were then used to correct the BAC assemblies. These were then compared to the *S. pennellii* assembly using BLASTN (blast 2.2.28+) exactly as above (section 2.1.1). The correction resulted in a substantial improvement in the sequence comparison, with mismatches reduced by 41.8% (from 395 to 165) and INDELs reduced by 73.7% (from 224 to 59), giving a combined error rate of 0.037%, suggesting that indeed a lot of sequence differences were caused by errors in the BAC sequences rather than the *S. pennellii* assembly.

The remaining BAC sequence data, for which no independent reference was available (CP020G005 and BP029K005), were then de-novo assembled (using SPAdes v2.5.1, with default parameters), but this resulted in a fragmented assemblies which would not provide a useful basis of comparison. As an alternative, this BAC sequence data was aligned against the *S. pennellii* assembly (as above), and the corresponding regions identified. The "BP029K005" line was found to map against a ~103 kb segment of scaffold253.1, while the CP020G005 was found to align against a ~147 kb region of scaffold23.1 plus a ~196 kb region of scaffold70.1 (neither region alone was sufficient to align this data).

Although the lack of a suitable independent reference limited the testing potential of this data, the successful identification of large regions of the expected size is still positive evidence in favor of the *S. pennellii* assembly. Furthermore, only a modest number of variants (90 mismatches, 33 INDELs) were detected by re-aligning the reads with the 3 corresponding regions, a combined error rate of approximately 0.028%, which is comparable to the other BAC sequences. (Supplementary Table 9)

2.2 Structural assessment using BAC-end sequence data

The 10,615 BAC-end pairs sequenced as described in Section 1.7 were combined with 72 and 242 additional sequences[2] and aligned against the scaffold sequences using BLASTN (from blast+ 2.2.28), using an e-value cutoff of 1e-10 but otherwise default parameters. Reads which had only one hit, or had a bit-score margin above 500 between the best and second best hits were considered unambiguously aligned. Of the 6,015 BAC-end pairs which were unambiguously aligned on both ends within one scaffold, 5,996 (99.68%) had the correct relative end orientation and expected distance (<300 kb), while 19 (0.32%) BAC-end pairs were suggestive of potential misassembly.

Since the evidence for mis-assembly was weak, and could easily be an artifact of collapsed/missing repeats (making an ambiguous alignment appear unambiguous), it was decided not to make scaffold modifications on the basis of BAC-end evidence. Indeed, the scaffold splitting performed on the basis of the markers in section 1.15 did not resolve any of the 19 problematic BAC-end sequences.

2.3 Completeness assessment using RNA data

Publicly available EST data, from both *S. pennellii* and *S. lycopersicum* was downloaded from NCBI.

BLASTN (blast+ 2.2.28) was used to align the EST sequences against the *S. pennellii* assembly using an e-value cutoff of 1e-05 and with 'dust' filtering disabled and reporting only one hit per query (-max-target-seqs 1), but otherwise default parameters. Compatible alignments, such as those caused by intron splicing, were merged using a custom script.

For *S. pennellii* 7812 sequences could be retrieved, but not all of these were from the same cultivar LA716 as the genome sequence, which could have some impact on the alignment rates. From this dataset, 88.8% (6940 of 7812) of the ESTs could be aligned when requiring 95% accuracy and 95% completeness. This number increased to 96% (7503 of 7812) when the completeness was set to 80% at 95% identity.

In the case of *S. lycopersicum* 307,350 sequences could be retrieved from the NCBI, and these were aligned to the genome as above. For this dataset, 77.9% of the ESTs could be aligned with 95%

completeness with a minimal identity of 95% (239,445 of 307,350) and 93.2% aligned when the completeness was set to 80% at 90% identity (286,556 of 307,350).

Furthermore, the combined tomato unigene dataset which includes sequences from many tomato species was downloaded from the SOL website (*solgenomics.net*). This was aligned using the same approach, and resulted in 83.2% (35,154 of 42,257) alignment with 90% accuracy and 80% completeness. This relatively low alignment could suggest assembly incompleteness, so the alignment process was repeated against the *S. lycopersicum* cv. Heinz assembly for comparison. This resulted in 85.9% (36,317 of 42,257) alignment, a relatively small improvement. This indicated that the most likely explanation is that some of the unigene sequences are incorrect. (Supplementary Table 10)

Finally, the *S. pennellii* RNA-Seq datasets, which are described below, were aligned against the genome using TopHat2[17] allowing up to 4 mismatches (-N 4 –read-edit-dist 4). Overall, 91.17% of the RNA reads were successfully aligned. To determine if the unaligned reads were due to assembly incompleteness, the unmapped reads were then aligned against the *S. lycopersicum* cv. Heinz assembly, as before. Only a small number of reads, corresponding to 0.33% of the dataset, were uniquely aligned to the *S. lycopersicum* assembly, suggesting issues other than assembly incompleteness were likely responsible for most non-aligning reads. In any case, this number compares well to a study by Engstrom[18] who showed that between 83% and almost 94% of reads stemming from real experiments could be mapped using TopHat for mouse and human data.

In summary, independent RNA data, both from public sources and sequenced specifically for this project, indicate that the gene space is mostly covered by the *S. pennellii* genome sequence. The datasets which produced lower than expected alignment rates, i.e. the Unigene and to a lesser extent, the *S. pennellii* RNA-Seq data, were also tested against the *S. lycopersicum* cv. Heinz assembly, where they performed only slightly better, indicating that the low alignment rates reflected issues with the datasets rather than incompleteness in the *S. pennellii* assembly.

3 Chloroplast assembly

Direct *in-silico* assembly of a chloroplast genome from a whole-genome shotgun dataset is challenging, primarily due to the presence of copies or near copies of the chloroplast genome within the nuclear genome. Although the true chloroplast sequence should have much higher coverage, despite the use of nuclear-enrichment DNA extraction protocols, the estimated post-filtering nuclear coverage of about 137x would still be sufficient to ensure the ‘nuclear’ versions of the chloroplast sequences were considered valid alternative sequences by most assembly tools. These alternative sequences would presumably occur quite frequently within the chloroplast sequence, since they can occur independently for each nuclear copy of any part of the chloroplast genome.

In the best case, such alternatives typically result in a fragmented assembly, since the assembler cannot decide on a single ‘path’ through the assembly graph, and must instead produce short contigs for each linear graph portion. In the worst case, the extremely high coverage of the chloroplast sequences would cause them to be mis-classified as repeat sequences, and they would be masked during assembly. In either scenario, a relatively poor sequence assembly can be expected.

As a result, a pre-processing strategy to enable assembly of very high coverage sequences was adopted. The published *S. lycopersicum* chloroplast genome[19] was used as bait to estimate the expected coverage of the chloroplast. This indicated an average coverage of approximately 58,400 fold.

Next, the k-mer coverage level was calculated for 19-mers within the paired-end libraries. Each read was then classified based on the median coverage of its constituent k-mers. The libraries were then filtered, and read pairs for which both reads had an estimated coverage of 10,000 or above were retained. This target coverage level represented a considerable margin above the nuclear 19-mer coverage level of 107, yet still far below the estimated chloroplast coverage determined above.

Given the extremely high coverage, it was decided to use only reads from libraries with 100bp or longer, which should in general be more informative. This reduced the data volume by approximately 60%. Since this dataset still had coverage far above the typical levels used for sequence assembly, these sequences were further reduced by creating 4 data subsets, which retained 1%, 2%, 5% and 10% of the read pairs, selected at random.

These 4 data subsets were assembled using SPAdes [20] (v2.5.1), using k-mer sizes of 21, 33, 55, 77, 95, 111 and 127 (taking advantage of the atypical feature of SPAdes to use multiple 'k' values per run). Although the longest reads were only 150 b long, which would normally mean using the suggested kmers of 21, 33, 55 and 77, the high coverage available made it possible to use higher values, up to the maximum supported (127). The use of 3 additional 'k' values (95,111,127), rather than the two suggested in the SPAdes manual (99,127) was due to the large fraction of 100 b reads in the dataset, which would be difficult to exploit at a k-mer size of 99.

The final assembly, which merges the results of the various k-mer sizes, was compared across the 4 datasets. In each data subset, the 3 longest contigs were identical, and could be confirmed using BLASTN (blast 2.2.28+) as corresponding to the large-single copy, inverted repeat and small single-copy regions of *S. lycopersicum* chloroplast.

Due to a technical limitation of SPAdes, which apparently requires mate pair reads that are at least a long as the longest k-mer used, it was not possible to scaffold the chloroplast contigs into a single molecule automatically. Instead, alignments of the mate pair data were manually inspected and this verified the structural arrangement of these elements is as in *S. lycopersicum*. The final chloroplast sequence was manually arranged, as indicated by the mate pair data. Finally the circular chromosome was split to best align with the *S. lycopersicum* sequence, for ease of comparison.

4 Gene identification

The genes on the *S. pennellii* genome were predicted using AUGUSTUS version 2.7[21], using the parameters trained for *S. lycopersicum*. To aid gene finding, RNA-Seq data from *S. pennellii* was sequenced as described below, and combined with publically available RNA-Seq and EST data from various *Solanaceae* species to generate hints for AUGUSTUS about gene position and structure. The individual data sets are described in the following sub-sections.

4.1 Generation of a varied *S. pennellii* RNA-Seq data set

In order to provide a diverse set of extrinsic evidence data for the identification of genes, *S. pennellii* was grown under a diverse set of environmental challenges (Supplementary Table 28). In addition, different plant organs were harvested under controlled conditions resembling a sub-set of the plant organs of *S. lycopersicum* investigated previously[22].

4.2 Plant growth

The plants for the diurnal time series were grown in a phytochamber under 14 hour day conditions at 400 µE/m^2/s, 22°C, 50% relative humidity during the day, alternated with 20°C and 50% relative humidity during the night for eight weeks. Each plant was harvested at the specified time during the day (4 samples, A01-A04), night (3 samples A05-A07) or under extended night conditions (3 samples A08-A10). (See Supplementary Table 28, where D+X, N+X specifies X hours after daybreak or nightfall, respectively; ED and EN the end of the day and the end of the night and XN+X specifies X hours into extended night).

Additional soil-grown plants were germinated on soil and transferred post-germination to individual 6 cm pots, and grown in standard conditions in a greenhouse for 6 weeks. These plants were treated and sampled as follows:

- 4 tissue-specific samples pools, each consisting of material from 6 untreated 6-week old plants were taken from the above-ground tissues. These 4 sample pools consisted of small leaves (1-2 days growth), mature leaves, meristem/node tissue from the stem, and inter-node stem (samples A11-A14).
- 3 plants were subjected to *Pseudomonas syringae* pv. tomato DC3000 injections. Samples were taken 24 hours after inoculation from infected leaves, uninfected existing leaves, and small new leaves which had formed since infection. Equivalent tissues from each plant were pooled (samples A15-A17).
- 2 plants were moved to a cold (4 °C) growth chamber for 24 hours, and pooled samples taken from small new leaves and mature leaves (A18/A19)
- 2 plants were subjected to high UV levels (2 W/m^2) for 72 hours, and 2 plants were subjected to moderate UV (1 W/m^2) for 120 hours, in appropriate specialized growth chambers. Samples were taken from each condition for small new leaves and mature leaves, and pooled (A20/A21/A25/A26).
- 4 plants were left without added water for 4 days, followed by 3 days with limited watering. Samples were taken and pooled from small new leaves, mature leaves and root (A22-A24)
- 1 plant was wrapped in a transparent fine net, and a Colarado beetle (*Leptinotarsa decemlineata*) placed inside for 72 hours. Samples were taken from both visibly damaged and non-visibly damaged leaves (A27/A28).
- 2 plants were moved to a high-light chamber (1500 µE/m^2/s) for 96 hours. Mature leaves, showing high levels of anthocyanin accumulation were harvested and pooled (A29)

- 4 plants were transferred to field conditions for 10 days. New small leaves and mature leaves were sampled and pooled (A31, A32).

For hydroponic cultures, seeds were germinated on filter paper soaked in water before they were transferred to “full nutrition” liquid media (800 µM $Ca(NO_3)_2$, 330 µM FeEDTA, 550 µM K_2HPO_4, 1 mM KNO_3, 500 µM $MgSO_4$, 7.6 µM H_3BO_3, 70 nM $CuSO_4$, 1.6 µM $MnSO_4$, 70 nM Na_2MoO_4, 130 nM $ZnSO_4$) and grown in greenhouse conditions for 5 weeks.

Plants were then transferred to separate hydroponic cultures for stress treatment. The stresses were applied as follows:

- To apply salt-stress, plants were transferred to fresh media, as above, with 50mM/100mM NaCl added for the low salt (A33, A41) and high salt stress (A34, A42) experiments respectively. An additional 50mM/100mM (using a 4 M NaCl solution to minimize dilution) was added each day for 3 additional days until a target concentration of 200mM/400mM NaCl was reached.
- To apply nitrogen starvation, plants were transferred to fresh media, as above, but without $Ca(NO_3)_2$ or KNO_3, which were substituted by 500 mM K_2SO_4, and 800 µM $CaCl_2$ (A35, A43).
- To apply iron starvation, plants were transferred to fresh media, as above, but without FeEDTA (A36, A44).
- To starve plants for magnesium, plants were transferred to media, as above, but without $MgSO_4$, which was substituted with 500 mM K_2SO_4 (A37, A45).
- For calcium depletion, plants were transferred to media, as above, without $Ca(NO_3)_2$, but with 3 mM KNO_3 (A38, A46).
- Control plants were given fresh media, as above.

Plants were grown under these conditions for one additional week. For each stress condition, 4 plants were pooled with shoot and root material harvested separately. For control plants, 8 plants were combined into two pools with shoot (A39, A40) and root material (A47, A48) harvested separately.

For seedling tissue harvesting, 12 plants were germinated on sandy soil and grown for 3 weeks conditions, with shoot and root tissues harvested separately into 6 pools of 2 plants each (D1-D12). All other tissue samples were taken from mature plants grown under greenhouse conditions, when appropriate tissues became available. Unopened buds and fully opened flowers were harvested from multiple plants and combined into 3 pools per tissue (D13-D18). Pollen samples were also taken from flowers (A30). Immature and mature fruits were harvested 35/70 days after anthesis, respectively, and also combined into 3 pools per tissue (D19-D24).

All harvested tissues were immediately frozen in liquid nitrogen and stored at -80°C for RNA extraction.

4.3 RNA extraction

Total RNA isolation was performed with a phenol-chloroform based method as previously described [23].

4.4 RNA sequencing

2 of the 72 RNA samples failed quality controls (D11, D12), but the remaining 70 were used to create RNA-Seq libraries using the Illumina TruSeq RNA kit in accordance with the manufacturer's instructions. In total, 400 M reads, comprising 20.4 Gb were sequenced. These libraries were combined with existing publically available RNA-Seq data to create extrinsic evidence for gene finding, and some were also used for differential gene expression analysis.

4.5 Public RNA-Seq and EST data

Publically available RNA-Seq data, comprising 40 runs from *S. pennellii* (SRR027939, SRR088750, SRR088752, SRX252076-SRX252078, SRX252024-SRX252052, SRX251982-SRX251985), 38 from *S. lycopersicum* (SRR363116-SRR363124, SRR404309-SRR404329, SRR404331, SRR404333, SRR404334, SRR404336, SRR404338, SRR404339, SRR412747, SRR412748 and SRR507782), and 53 from *S. tuberosum* (SRR122108-SRR122140, SRR124121, SRR124126, SRR124127, SRR124130-SRR124132, SRR124138, ERR029909-ERR029917, ERR029920, ERR029921, ERR029924) were downloaded from the SRA (http://www.ncbi.nlm.nih.gov/sra). These archives were extracted, yielding approximately 130Gbp of sequence data. Species breakdown of this data was approximately 40 Gb (30.5%) from *S. lycopersicum,* 11 Gb (8.7%) from *S. pennellii*, and 79 Gb (60.8%) from *S. tuberosum*.

In addition to the RNA-Seq data, 737250 ESTs comprising almost 500Mb of sequence data, were downloaded from GenBank (September 2011). These ESTs include sequences from 13 Solanaceae or closely related species (*S. pennellii, S. lycopersicum, S. pimpinellifolium, S. tuberosum, S. melogena, S. torvum, S. habrochaites, S. chacoense, S. lycopersicum var. carsiforme, S chilense, S. peruvianum, Nicotiana benthamiana, Coffea Arabica*).

This EST data was combined with the public RNA-Seq data and RNA data sequenced within the project, and processed to create 'hints' for gene finding as described below.

4.6 Generation of extrinsic evidence for gene prediction

The available transcript data was mapped to the *S. pennellii* genome, using BLAT[24] as recommended by the AUGUSTUS manual, to generate hints from ESTs and RNA-Seq (available on the webpage: augustus.gobics.de). The hint data set comprised 77 GB and provided information about position of exons and introns in the *S. pennellii* genome. Since a relatively large hint set was available, it was decided to increase the extrinsic evidence weighting used by Augustus.

4.7 *S. pennellii* gene annotation and filtering

Augustus yielded 55,147 transcripts consisting of 51,110 genes and 4,037 additional splicing variants. These genes were filtered using two approaches, homology with known proteins and RNA evidence.

In the first approach, the 51,110 genes, minus 608 which are sequence duplicates, were compared to protein sequences from other species. 41,860 genes were assessed as sufficiently similar to known protein sequences, and thus retained.

In the second approach, the 70 libraries of RNA-Seq data were aligned against the transcripts. All genes which had coverage of at least 3 for 150 nucleotides or 75% of the mRNA were considered sufficiently supported and retained. In total, 35,160 transcripts passed this threshold, of which 32,086 overlapped with those already retained due to homology. These 44,828 genes, plus their 138 duplicates, form the final 44,966 primary gene models. Their 3,958 alternative splicing forms bring the total to 48,924 protein-coding transcripts.

A second high-confidence gene set was created by retaining only those genes which had both homology (as above) and also coverage of at least 2 for 150 nucleotides or 75% of the mRNA. This created a set of 31,643 genes, which was supplemented by 630 additional genes from orthologous pairs identified in section 5.2 below. This resulted in a high-confidence gene set with 32,273 genes, plus their 3,600 alternative splicing forms, bringing the total to 35,873 protein-coding transcripts.

4.8 Gene statistics and validation

The current primary genome annotation consists of 44,966 genes with 48,924 protein-coding transcripts including alternative splicing forms. The average protein length is 518 amino acids, while the N50 length is 726 amino acids. There are, on average, 5.7 exons per gene model.

To verify the completeness of these gene models, we compared the number of RNA-Seq reads which aligned against the unfiltered and filtered gene models. In both cases, the alignment rate was 87.06%, with only 7538 additional reads out of almost 342M reads aligning against the unfiltered models. This indicated that there was very little direct evidence for the models that had been removed by filtering.

Using BLASTP (blast 2.2.28+), we have found that our gene models could align against 21,704 proteins (61.3%) of the *Arabidopsis* proteome (TAIR10) with at least 80% coverage and a minimum 50% identity. Similarly, when aligning our gene models against the ITAG2.3 release of *S. lycopersicum* proteins, BLASTP produced 24,523 (70.6%) alignments of minimum 80% identity and minimum 80% coverage, while 36,113 (64.2%) of the proteins from the potato genome could be found at the same 80% identity / coverage threshold. When comparing the codon usage of *S. pennellii* to *S. lycopersicum* no gross difference could be identified, although there was a minor increase in codons with higher GC content (Supplementary Table 11).

5 Functional gene annotation

5.1 Functional classification using mercator

All potential 44,966 *S. pennellii* protein coding transcripts were submitted to Mercator[25] for a draft annotation as well as for a functional annotation using MapMan classes. Mercator grouped 20,076 (ca. 44.6 %) proteins into various meaningful MapMan bins (Supplementary Dataset 2 and Supplementary Table 13). This is below the annotation rate of the well-studied model species *Arabidopsis thaliana* which has 60% meaningful annotations[26]. However given that *Arabidopsis thaliana* is still the most well annotated plant genome, this number was judged acceptable. The results from the Mercator pipeline are presented as a pie chart in Supplementary Figure 3. The resulting corresponding mapping file is provided in Supplementary Dataset 1. This file includes i) a short annotation for each gene and ii) the MapMan annotation. This file can directly be used in MapMan to visualize data.

5.2 Identification of simple orthologous pairs

In order to identify simple orthologous one to one pairs, a reciprocal best blast search strategy was used. To this aim, the protein models for *S. lycopersicum* were downloaded (ITAG2.3_proteins.fasta) and aligned against the ones from *S. pennellii* using BLASTP with an e-value threshold of 10^{-10}. Subsequently, only the best matching pairs were kept. This resulted in 21,662 putative simple orthologous pairs. The list of pairs is available in Supplementary Dataset 3. This list was generated for direct one to one comparisons, but a full list of all orthologous sets (including n to m relationships) was assembled using Ortho MCL (Section 5.4) which comprised 33,741 protein coding transcripts.

5.3 Comparison between the two tomato species and potato

To allow an unbiased comparison between the three sequenced *Solanum* species, *S. lycopersicum* and *S. tuberosum* were also subjected to the exact same Mercator pipeline. As the MapMan ontology (unlike GO) strives to reduce redundancy, a total comparison between proportions is more easily realized. As can be seen in Supplementary Table 12, 13 and Supplementary dataset 1, the informative classifications were largely similar with 18,591 informative assignments for *S. lycopersicum* (54.0%) and 17,877 for *S. tuberosum (45.8%)*. This data was further analyzed by checking each first sublevel BIN for a change of at least 10% and testing it with an approximate z-test for difference in proportions.

Interestingly, the test revealed that the draft *S. pennellii* genome contained significantly less annotated genes in photosynthesis-lightreaction (p<1.0e-10), and mitochondrial electron transport chain (NAD-dehydrogenases and cytochrome c oxidases, p<0.02 in both cases). This seemed intriguing, as these were genes sets specifically related to organelles. It has previously been observed that genes involved in photosynthetic processes are enriched in differentially expressed genes amongst tomato species[27] and that photosynthesis remains active longer in *S. pennellii* fruit (also see Supplementary Note Section 5.9 and discussion therein). When comparing the classifications to the potato genome, the number of photosystem genes was almost identical to *S. pennellii*, and in terms of mitochondrial genes, the potato genome, often featured even less genes than *S. pennellii* (Supplementary Figure 4).

Given that a similar situation occurred for chloroplastidic and mitochondrial genes, a simple explanation might of course be the integration of organellar genomes into the nuclear genome or different gene calling. To explore this, we used the genome browser (http://phytonetworks.ucdavis.edu/gb2/gbrowse) established for the comparison of *S. pennellii*, *S. lycopersicum*, *S. pimpinellifolium* and *S. habrochaites*[27]. This resource shows the reads mapping from these genomes to the *S. lycopersicum* reference. We manually analyzed cytochrome c oxidase, photosystem I subunits and photosystem NADH dehydrogenases as these were amongst the strongest drivers for the changes in the Bin counts (extracted from Supplementary Table 13). We analyzed the expression in these four species, but also if the alignments were in accordance to each other and to the reference gene models. Interestingly, we observed, that both for cytochrome c oxidase and for the NADH dehydrogenases there was very often no read support for the reference gene model in any species. It is possible that many of these genes might not be expressed and thus might not be called in *S. pennellii.* One very interesting exception was Solyc11g042420 which showed expression in *S. lycopersicum* and *S. pimpinellifolium* but not in the two green fruited species. The situation was more complex in the case of the photosystem I polypeptide subunits, where frequently the RNA-Seq evidence was not in support of the gene model structure (The whole analysis is given in Supplementary Dataset 16). To finally analyze these data and to distinguish non expressed genes (likely derived from organellar introgressions in the nuclear genome) it will be necessary to obtain very deep transcriptomic datasets and even more additional tomato reference genome sequences.

Further categories of gene families that were less abundant in the *S. pennellii* genome included pectin esterases (p=0.08), cytokinin genes (metabolism and signaling, p<0.01) and auxin related genes. Furthermore, there were fewer genes implicated in biotic stress (p=0.013), for lipid transfer proteins (p=0.03) and genes involved in transcription or its regulation (p=0.02). Investigating transcriptional regulators for changes in individual transcription factor families revealed that this was not due to a simple loss of all transcription factor families, but showed a strong decrease in the number of ARR (atypical response regulators), C3H, AS2, ATrich and AtSr type transcriptional regulators as well as B3, chromatin remodeling factors, histone deacetylases, methyl binding proteins and LUG type transcription factors. Once again the values for potato were more similar to the *S. pennellii* values and thus also in most cases lower than the ones for *S. lycopersicum*. Furthermore more genes were classified as protein synthesis (p<0.01) as well as assembly/co-factor ligation (p<0.001) protein modification and degradation (p<0.001 in both cases) and unspecified development for *S. lycopersicum* (0.017)

Potentially the increased in auxin-related and development genes might be due to the domestication of tomato which might have selected for additional ripening related factors. This would also explain the increase of pectin esterase genes, some of which are following the fruit ripening process in their expression[28,29]. We further investigate fruit ripening and maturation in Section 5.9.

Despite some general trends the correlation between counts on 2nd level categories was higher for *S. lycopersicum* at 0.997 than for *S. tuberosum* at 0.987. These very high correlation values, however, reflect the relatedness of these different species.

5.4 Comparison of gene family clusters

The predicted protein sequences of *S. pennellii* were compared with the proteomes of *S. lycopersicum* (ITAG v2.3), *Solanum tuberosum* (PGSC v3.4), *Arabidopsis thaliana* (TAIR v10) and *Oryza sativa* (MSU v7)[30,31], after filtering of alternative splicing and transposable elements. Protein sequence clusters were identified by an all-against-all comparison using BlastP (max. E-value 1 x 10^{-6}) followed by a clustering using OrthoMCL (inflation parameter 1.5) resulting in 33,741 protein coding transcripts from *S. pennellii* being grouped into clusters. The shared and distinct clusters are shown by the Venn diagram (Supplementary Figure 5) and are listed in Supplementary Dataset 4.

5.5 Analysis of protein families

The *S. pennellii*, *S. lycopersicum* and *S. tuberosum* genomes were analyzed for their domain content using Interproscan and Pfam separately. In the case of Pfam A (i.e. manually curated families) the results were ordered by the occurrence of a specific domain in the *S. lycopersicum* genome to suppress transposon related protein domains in *S. pennellii* (Supplementary Dataset 5). Plotting the thirty most abundant *S. lycopersicum* domains (Supplementary Figure 6), showed that the most abundant families were Protein kinases and LRR type domains. In general a strong similarity between the different species could be observed, which is in agreement with the automated Mercator classifications. The strongest differences were found for many LRR type domains, as well as for the Fbox and Fbox-like Pfam domains, where the potato genome harbored many more protein coding genes than the two tomato species. Proteins containing LRR domains are known to function in both defense-related and developmental processes[32]. However, the majority of LRR domain proteins appear to be involved in plant immunity. Prominent examples include plasma membrane-localized receptor-like kinases (RLKs), which play a decisive role in basal defense triggered by microbe/pathogen-associated molecular patterns (M/PAMPs[33]), and cytoplasmic nucleotide-binding leucine-rich repeat proteins (NB-LRRs), which generally confer isolate-specific resistance against diverse types of pathogens[34]. The presence of considerably more LRR domain-containing proteins in *S. tuberosum* compared to *S. lycopersicum* (nearly twice as many) has been previously noted[35] . This difference has been interpreted as the possible result many factors in *S. tuberosum* including polyploidization, genome size variation, natural selection, artificial selection including domestication history, breeding and cultivation, and gene family interactions[35]. In particular, the loss of duplicated gene copies from the tomato genome following whole genome duplication events has been attributed to the lower number of LRR domain proteins in tomato[35]. Alternatively, or in addition, differential pathogen pressures in natural habitats or during the course of domestication may account for the striking differences in gene numbers encoding LRR domain proteins between potato and tomato.

Whilst, in most cases there were slightly more *S. lycopersicum* proteins harboring a given domain, this situation was reversed for the P450 domain. Similar to the Mercator classification there were many more potato genes having a P450 domain, which was followed by *S. pennellii* and then by *S. lycopersicum* (Supplementary Figure 6). The data for the pectinesterase domain also corroborated the Mercator results, where 64 proteins showed this domain in *S. pennellii,* 80 in *S. lycopersicum* and 71 in *S. tuberosum* (Supplementary Datasets 5, 6).

5.6 Gene family analysis

In order to further explore the significance of the P450 and pectin esterases, phylogentic trees were built and individual sequences compared.

5.6.1 Pectin Esterases

5.6.1.1 Pectin Methyl Esterases

The Interproscan results were searched for matches to "Pectinesterase". This resulted in the identification of 79 and 66 transcripts for *S. lycopersicum* and *S. pennellii*, respectively. For *S. pennellii* four transcripts were removed as they represented splice isoforms, leaving 62 loci for *S. pennellii.* The encoded proteins were then aligned to the PFAM (PF01095) model using hmmalign from the hmmer3 package[36]. This was used to build a multiple sequence alignment. Subsequently, all proteins having lost more than 50% of the sequence aligning to the conserved PFAM domain were removed and the alignment corrected. This resulted in a total of 65 and 60 proteins in *S. lycopersicum* and *S. pennellii*, respectively. A protein sequence from *Selaginella moellendorffii* was used as an outgroup.

The resulting multiple sequence alignment was then subjected to the 'proml' program in the PHYLIP 3.695 package[37] and the maximum likelihood tree created using Figtree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/), and shown in Supplementary Figure 18.

5.6.1.2 Pectin Acetyl Esterases

Pectin acetylesterases were investigated as described for pectin methyl esterases, but the keyword "Pectinacetylesterase" was used. This resulted in 17 proteins to be included in the family tree for *S. pennellii* and 19 for *S. lycopersicum* (Supplementary Figure 19).

5.6.2 P450

The interproscan annotation was searched for P450 genes, resulting in the identification of 255 and 323 transcripts for *S. lycopersicum* and *S. pennellii* respectively. To further categorize the genes, the hidden markov models (HMMs) for the superfamilies CYP51 CYP74 CYP97 CYP710 CYP71 CYP82

CYP85 and CYP86 were obtained from CYPED (http://www.cyped.uni-stuttgart.de/cgi-bin/CYPED5/index.pl). The full list of *S. lycopersicum* and *S. pennellii* transcripts were scanned using hmmscan with these HMMs. Each gene was then assigned to a single CYP family accordingly.

For superfamily CYP71, this resulted in 349 hits, 201 for *S. pennellii* and 148 for *S. lycopersicum*. A multiple sequence alignment was obtained by using hmmalign with the CYP71 HMM as a reference. A maximum likelihood phylogenetic tree was generated using the 'proml' tool from the phylip package. The tree was rooted using CYP711 as outgroup, which only comprised a single gene in both *S. pennellii* and *S. lycopersicum* (Supplementary Figure 20, Supplementary Figure 21).

The transcripts count for the remainder of the CYP superfamilies were 107 and 122 for *S. lycopersicum* and *S. pennellii*, respectively. These transcripts were grouped together and a multiple sequence alignment was generated using the P450 HMM. The phylogenetic tree was generated using 'proml' and the tree was mid-point rooted.

S. lycopersicum has previously been show to lack the CYP727 superfamily. We also saw no evidence that *S. pennellii* contains any genes from this family. Though *S. pennellii* showed a greater number of genes overall, some of the difference can be explained by difference in gene annotation between *S. pennellii* and *S. lycopersicum*. However in many cases *S. pennellii* genes were clustered with probable pseudogenes from *S. lycopersicum* explaining the difference in numbers. In some of these instances, the *S. pennellii* gene would appear to have retained its functionality. One notable example is the case of scaffold348.1.g197, which clusters with a previously described P450 pseudo gene (*CYP71AT20P*). However, the *S. pennellii* version does not display an early stop codon seen in the *S. lycopersicum* version and from the expression data; this gene is found predominantly in mature fruit, but also at low levels in all other tissues.

Interestingly, *S. pennellii* lacks one CYP gene in a tandem gene region found in both *S. tuberosum* and *S. lycopersicum*. On closer examination of this region in the three genomes, it was revealed that both *S. tuberosum* and *S. lycopersicum* contain three genes in common with *S. pennellii* in this region but additionally one extra gene (Supplementary Figure 22), potentially suggesting a gene loss in *S. pennellii*. However due to the tandem repeat nature of this region, a missassembly cannot be entirely excluded.

An initial inspection of the CYP97 superfamily showed that *S. pennellii* contained 3 extra genes when compared to *S. lycopersicum*. When these genes were aligned together with *Solyc05g016330*, it was observed that each of the four *S. pennellii* genes covered half of the *Solyc05g016330* gene, with the four copies giving 2 times full coverage. Examination of the genome region revealed that each pair of *S. pennellii* genes were separated by a retrotransposon. This likely has consequences for the carotenoid biosynthesis pathways, in which the CYP97 superfamily are known to be involved.

5.7 Evolutionary analysis between *S. pennellii* and *S. lycopersicum*

The 21,662 orthologs between *S. pennellii* and *S. lycopersicum* described earlier were used as the basis for a Ka/Ks evolutionary analysis. For each orthologous pair, the *S. pennellii* and *S. lycopersicum* sequences were aligned using Muscle[38] (Version 3.8.31). The resulting amino acid alignments were converted to nucleotide alignments using Pal2Nal (Version 14).

In order to be used with KaKsCalculator (Version 1.2), the nucleotide multiple alignments were first converted into 'AXT' format, using the conversion tool supplied. These alignments were then grouped into 500 per file, and processed by KaKsCalculator using default parameters. The output from each group was merged to form the complete Ka/Ks analysis. (Supplementary Figure 14, Supplementary Dataset 7)

Of the 21,662 gene pairs assessed, 100 showed no nucleotide changes and thus are not considered further. The median Ka/Ks value was 0.23 consistent with our previous estimates based on several tomato species[27].

A large set of gene pairs were found to be under strong conservation pressure, with 4,818 pairs giving a Ka/Ks ratio below 0.1, and 4,684 of these had a p-value below 0.01. Unsurprisingly the resulting set mainly was enriched for housekeeping functions such as ribosomal proteins, TCA cycle and glycolysis genes, cell wall synthesis, photosystem, amino acid metabolism and histones, cytoskeleton and others. However surprisingly this set also comprised abiotic stress genes, SNF7 transcription factors and developmental genes. A complete list can be found in Supplementary Dataset 7.

With a more stringent 0.01 Ka/Ks cut-off, 1,438 pairs were found, 1061 of which were significant at a p-value of 0.01. However the identified categories were largely similar, potentially stressing more histone related processes. As this data set was more stringent, but still large enough for analysis, we also investigated underrepresented categories, showing at least a gene in this class. Apart from unknown and “misc” (i.e. large gene families), we found biotic stress, receptor kinases and P450 genes to be underrepresented amongst the highly conserved genes (Supplementary Table 15).

Perhaps unsurprisingly, considering the high similarity of the species, only a relatively small number of diversifying genes were found, and the majority of those did not show significant p-values. In total, 1,009 genes were found with a KaKs>1, but only 16 of these were significant at the 0.05 p-value threshold (Supplementary Table 16). At a threshold of KaKs>2, only 373 genes were found, with 12 having a p-value below 0.05 (Supplementary Table 17).

When analyzing the significant genes for an overrepresentation using the PageMan online tool [39] with FDR correction and all 21662 genes as background no meaningful process could be identified as only unknown genes were enriched. However when taking into account all proteins showing a Ka/Ks value>1, in addition to the overrepresentation of unknown genes, ACP protein candidates were overrepresented. These are potentially related to lipid and wax synthesis. This impression was strengthened by investigating protein categories having a KaKs>2 where once again ACP proteins were identified and its parent class, fatty acid synthesis and elongation was the next most enriched class, although this was not significant after FDR control (Supplementary Table 17).

5.8 Selected interspecific variation

5.8.1 Interspecific sequence variation of coding sequences

For the analysis of coding regions, the respective deduced amino acid sequences from *S. lycopersicum* cv. M82 and *S. pennellii* were downloaded from the genome browser set up for *S. pennellii* (described above) and aligned using the default settings of the pairwise alignment tool NEEDLE[40]. The amino acid alignments were then submitted to SIFT[41] in order to predict the impact of amino acid substitutions on protein function. SIFT scores were based on the degree of amino acid conservation in a set of related sequences retrieved through a BLASTP-based similarity search[42]. Amino acid sequences were also analyzed with InterPro[43] and CD-search[44] to locate the position of polymorphic amino acid residues with respect to domain organizations and active sites. Manual curation of the intron/exon structure was subsequently performed for those gene models that were differentially predicted between M82 and *S. pennellii*.

5.8.2 Interspecific sequence variation of promoter regions

For the analysis of promoter regions, the sequences of *S. lycopersicum* cv. M82 and *S. pennellii* up to 1000 bp upstream of the predicted ATG were aligned using Blast2seq[30] or the default settings of NEEDLE. Prediction of conserved promoter elements was performed with PlantProm[45] (http://linux1.softberry.com), while InDels in the aligned sequences were analyzed with PLACE[46] https://sogo.dna.affrc.go.jp) in order to detect variation in putative *cis*-acting regulatory elements.

5.9 Selected investigation of gene classes

In order to investigate differences between *S. lycopersicum* and *S. pennellii*, amino acid changes in coding regions and gene expression level for selected genes were checked (Supplementary Tables 21-24). In the following sections we discuss these functional classes one by one, paying attention to maturation events, the underlying endogenous metabolic pathways and current knowledge of the genes involved in fruit ripening and metabolism. Furthermore, since considerable information is available from mQTL analysis using *S. lycopersicum* and *S. pennellii* introgression lines[47-53], description of the result of mQTL is added in the Supplementary Table 20. The source of all published expression data is given in the table with the exception of the mature fruit expression data for *S. lycopersicum* and *S. pennellii* which has not previously been published but has been deposited in the NCBI data base: SRP041499.

5.9.1 Fruit development and Ripening related genes

We explored the *S. pennellii* genome for a number of well characterized and important genes involved in tomato fruit development and especially ripening including: ripening-related ethylene synthesis enzymes; *ACO1* (*Solyc12g005940*)[54], *ACS2* (*Solyc01g095080*)[55], *ACS4* (*Solyc05g050010*)[56], *PG* (*Solyc10g080210*)[57], ethylene receptors: *ETR1* (*Solyc12g011330*)[58], *Never-ripe* (*Nr/ETR3, Solyc09g075440*), *ETR4* (*Solyc06g053710*)[59] and well described but still not fully understood markers for ethylene response such as *E4* (*Solyc03g111720*) and *E8* (*Solyc09g089580*)[60,61]. A subset of important and well-described transcription factors (TFs) regulating overall ripening or aspects of this or related fruit development processes that were examined include: *GLK2* (*Solyc10g008160*), *GLK1* (*Solyc07g053630*), *TAGL1* (*Solyc07g055920*)[62], *FUL1* (*Solyc06g069430*)[63], *FUL2* (*Solyc03g114830*)[63], *RIN-MADS* (*Solyc05g012020*), *CNR* (*Solyc02g077920*), *HB1* (*Solyc02g086930*)[64], *AP2A* (*Solyc03g044300*)[65,66], *NOR* (*Solyc10g006880*)[54,67], *E4/E8 binding protein 1* (*Solyc04g070990*); genes related to fruit size, *FW2.2* (*Solyc02g090730*) *SUN* (*Solyc10g079240*) and the major ripening-related pectinase *PG* (*Solyc10g080210*) (see references[54,68] and Supplementary Dataset 13).

Ripening related enzymatic genes

ACS2 (1-aminocyclopropane-1-carboxylic acid synthase 2) is one of the key genes involved in autocatalytic ethylene production during fruit ripening[55,56], and it showed very low expression in *S. pennellii* compared to *S. lycopersicum* (10 times lower) consistent with the low ethylene evolution of maturing *S. pennellii* fruit[69]. In addition the *S. pennellii* gene has three additional nucleotides coding for an additional amino acid within its coding region, however, the functional consequence of this additional amino acid is not currently known.

There are three *E8* (1-aminocyclopropane-1-carboxylic acid like) proteins in both *S. lycopersicum* and *S. pennellii* with the most highly expressed one (*Solyc09g089580*) has a considerable smaller coding region than its ortholog in *S. pennellii* and correspondingly lower expression (2 times lower than *S. lycopersicum*). However, LeACO1 and ACS4 showed few amino acids changes and similar gene expression between the species. Moreover, analysis of polygalacturonase (PG), a key enzyme involved in the large changes in pectin structure that accompany fruit ripening[57], revealed 19 amino acids differences between *S. pennellii* and *S. lycopersicum* sequences, and its gene expression was 2.7 fold higher in *S. pennellii*.

PSYI (phytone synthase I) is the rate limiting enzyme responsible for the synthesis of fruit carotenoids, and the enzyme is not expressed in *S. pennellii*. Sequence alignments revealed only one amino acid difference, whereas PSY2 and 3 were more different. *PSY2* has similar expression in both species while *PSY3* didn't show any expression in any developmental stages in either species.

In contrast to the genes above, genes encoding β-lycopene cyclase 1, chromoplast-specific lycopene β -cyclase, lycopene β -cyclase 2 (*Lyc-B*) and *LycB_epsilon* are highly expressed in *S. pennellii* (4, 8, 463, 1345-fold higher compared to *S. lycopersicum* respectively), many amino acids changes were additionally found between *S. pennellii* and *S. lycopersicum* especially in LycB_epsilon. QTL analysis using the *S. pennellii* introgression lines showed that lycopene levels were below the level of detection in IL3-2, IL6-2, IL6-3, IL10-2 and IL12-2[70,71]. IL3-2 harbors the low expression *PSY1* gene consistent with this phenotype while 6-2 and 6-3 overlap in the region of the lycopene β –cyclase Beta allele which results in over-expression driving lycopene to β –carotene. In addition, IL10-2 and IL12-2 harbor *Beta-lycopene cyclase* (*Solyc10g079480*) and *LycB_epsilon* (*Solyc12g008980*), from *S. pennellii*, respectively. Therefore, all loci span strong candidate biosynthetic genes that could shift metabolic flux away from lycopene accumulation. It has been reported that the *PSY1* gene is under strong positive ethylene control during ripening, while *lycopene β-cyclase* is repressed[54,72]. The elevated expression of the downstream carotenoid synthesis genes in *S. pennellii* is consistent with the presumed reduced ethylene resulting from low *ACS2* expression described above and may additionally represent altered pathway feedback regulation resulting from reduced pathway flux.

Ripening-associated regulatory genes

The FUL1_FRUITFULL-like MADS-box[63] is considerably smaller due to amino acid deletions in the carboxy terminus in *S. pennellii* with one additional amino acid change (S121N in the K domain; also observed in other alleles), furthermore, the expression of this gene is very high in *S. lycopersicum* compared to *S. pennellii* (more than 30 times higher). It is noteworthy that the C-terminus of MADS-box proteins is associated with protein-protein interactions and this change could be reflective of reduced or lost function. The functionally redundant *FUL2* gene is intact and lower in expression in *S. pennellii* although not to the same extent (only 2 fold). The NOR transcription factor[54,67], has 11 amino acid differences mostly outside the conserved NAC domain (only two AA substitutions in this region at I30V, A41T and one is observed in other plant alleles) in the coding region between *S. lycopersicum* and *S. pennellii* and its mRNA abundance is three times higher in *S. lycopersicum*. The AP2A transcription factor[65,66] has a three nucleotide deletion toward the 3' end resulting in a premature stop codon in *S. pennellii.* There are six additional amino acid substitutions compared to *S. lycopersicum* as well as considerably lower expression in *S. pennellii*. AP2A is a negative regulator of ethylene synthesis and reduced activity is associated with accelerated ripening[65]. Reduced AP2A activity or expression in *S. pennellii* might compensate for the reduced ethylene synthesis of these fruit during ripening to effectively accelerate this process in the lower ethylene environment of

S. pennellii fruit maturation. The ethylene receptor genes (*ETR1*, *ETR3*, *ETR4*) have several amino acids changes between *S. lycopersicum* and *S. pennellii*, however, no significant difference in expression of these genes could be observed between *S. lycopersicum* and *S. pennellii*. GOLDEN2-LIKE (GLK) transcription factors regulate plastid, chlorophyll levels and fruit ripening[73], GLK1 and GLK2 showed eleven and nine amino acid difference between *S. lycopersicum* and *S. pennellii*, respectively. Consistent with their green ripe fruits, *GLK* expression levels are higher in mature fruits of *S. pennellii* compared to *S. lycopersicum*, however, in both species *GLK2* has higher expression than *GLK1*. Whilst there are several amino acid differences between the RIN-MADS[74], FUL2[63], LeHB1[64] and TAGL1[62] transcription factors of *S. lycopersicum* and *S. pennellii* they were expressed at similar levels in both species.

Fruit size

Analysis of FW2.2 revealed two amino acids differences between *S. lycopersicum* and *S. pennellii* consistent with two of the three differences previously reported[75]. Also consistent with this earlier study was the fact that this gene was generally lowly expressed but was more highly expressed in the mature fruit of *S. pennellii*. *Solyc10g079240*, one of the genes in the SUN locus[76], is more highly expressed in *S. lycopersicum* compared to *S. pennellii*. In *S. pennellii* the protein has a two amino acid deletions at the end, in addition to 16 amino acid differences overall, compared to *S. lycopersicum*.

Genes significant differently expressed between mature fruits of *S. lycopersicum* and *S. pennellii*

In order to investigate major differences in mature fruits, expression profiling of genes found in both the *S. lycopersicum* and *S. pennellii* genomes have been compared using mature fruit material. Among the top 100 highly expressed genes in *S. pennellii* mature fruit, photosynthetic related genes (21 genes), secondary metabolism (seven genes), development (six genes), DNA synthesis (five genes), hormone metabolism (five genes) and lipid metabolism (four genes). In addition, three genes associated with amino acid metabolism, cell wall metabolism, lipid transfer proteins and protein synthesis were in this list (Supplementary Tables 26, 27). Furthermore among the 21 photosynthetic related genes, many photosystem related genes; photosystem I light harvesting complex genes (LHCAs, three genes), photosystem II light harvesting complex related genes, *LHB*, *PSA*, *PSB*, totaling 11 genes); Calvin-Benson cycle related genes (*glyceraldehyde-3-phosphate dehydrogenase A*, *RuBisCO small subunit 3B*, totaling two genes) were observed. The difference in photosynthesis associated genes is highly interesting for two reasons. First, *S. pennellii* plastids do not undergo a chloroplast to chromoplast transition and hence it could be anticipated that they are more photosynthetically active when fruit mature. Secondly, as stated above *S. pennellii* maintains expression of GLK2 longer than *S. lycopersicum* and overexpression of this transcription factor has previously been shown to result in the upregulation of many photosynthetic genes[73]. By contrast, of the conserved genes showing lower expression levels in *S. pennellii,* 65% of the top 100 genes were of unknown protein function, whilst regulatory proteins included transcription factors (six genes) and post-translation modification proteins (six genes). These transcription factors include (*VRN1*, *Solyc04g015500*; *C2H2 zinc finger*, *Solyc04g015500*; *MADS-box Solyc07g052700*; *AT-hook, Solyc09g089620*; *PHD finger, Solyc01g087330*; *KH domain, Solyc03g034200*) and may represent additional regulatory mediators compensating for maturation in a reduced ethylene environment.

5.9.2 Primary metabolism

To extend this analysis, we next considered a set of 76 primary metabolism-related genes, which are involved in sugar, organic acids and amino acids metabolism from the literature (Supplementary Dataset 13)[77,78]. Alignment of their protein sequences indicated that a large number of these genes displayed no or only minor sequence differences between *S. lycopersicum* and *S. pennellii* (for example succinate dehydrogenase (*SDH 2-2, Solyc02g093680*), mitochondrial malate dehydrogenase (*mMDH, Solyc07g062650* and *fructokinase 1* and *2 Solyc03g006860* and *Solyc06c073190*, respectively). Moreover, these genes showed similar expression profiles when comparing different tissues of *S. lycopersicum* cv. M82 and *S. pennellii*. However, it is important to note that examples exist in the literature between these species whereby small changes in sequence results in a large change in function. One example is the *FW2.2* gene described above. Additionally, the introgression of a cell wall invertase isoform (*lin5,* S*olyc09g010080*) from *S. pennellii* into *S. lycopersicum* resulted in a higher fruit yield and higher soluble sugar content[79] despite the fact that only the 348D residue was uniquely associated with *S. pennellii* [79]. Recently, a similar study using *S. pennellii* introgression lines allowed the characterization of two cytosolically associated aconitase genes as important genes in controlling citrate and malate in tomato fruits (*ACO3a*, *Solyc07g052340* and *ACO3b*, *Solyc12g005860*)[80]. Furthermore, the lower abundance of malate in *S. pennellii* fruits [81] correlates with the higher expression of two phospho*enol*pyruvate carboxylase genes (*Solyc04g006970*, *Solyc09g015490*) and lower expression of one phospho*enol*pyruvate carboxykinase gene (*Solyc04g076880*) (Supplementary Dataset 13).

Sucrose transport and metabolism are well known determinants of crop yield and quality as they affect both growth and composition of harvestable sinks[82]. It has been suggested that the inhibition of the insoluble acid invertase and the use of the accumulated sucrose by other sucrolytic activities could be, respectively, considered as a limiting step and an adaptive responsive in the control of sucrose import and fruit growth under saline conditions[83,84]. Among the genes with sucrolytic activities, we observed differences in the sequences in two of the four cell wall invertases, the abovementioned *Lin 5* and *Lin 8*. The *Lin 8* gene (*Solyc10g083300*) showed considerable amino acid diversity in the first exon between *S. lycopersicum* and *S. pennellii*, however, in most tissues the expression of this gene was essentially the same between species. In contrast, sucrose synthase genes (*Solyc12g009300*, *SuSy2* and *Solyc07g042550*, *SuSy3*) whose activity seems to play an important function in the control of sucrose import[85,86], displayed only minor changes when comparing *S. lycopersicum* and *S. pennellii* sequences. Although it is important to note that there is a consistent significantly higher expression of *SuSy2* and *SuSy3* in *S. pennellii* fruits.

Given that primary metabolism is largely regulated post-translationally the lack of major changes in gene sequence is consistent with the minor differences in sugar and organic acid content between the species [81] irrespective of changes in gene expression (which were nevertheless also relatively, mild). Changes in the levels of amino acids were considerably higher [81]. On one hand, this could be the consequence of *S. pennellii*´s longer period of photosynthetic competence. However, it may also reflect variance in amino acid metabolism *per se*. In this vein it is important to note that two glutamate decarboxylase genes (*Solyc01g005000*, *Solyc03g098240*) have three and six amino acids changed between *S. lycopersicum* and *S. pennellii*, respectively; and that the levels of their expression are significantly higher in *S. pennellii* fruit. On the other hand, an aminotransferase gene (*Solyc07g043310*) showed relatively few amino acids changes between *S. lycopersicum* and *S. pennellii*, while higher expression was found in *S. lycopersicum* compared to *S. pennellii* in a range of tissues. Despite the fact that *S. pennellii* and *S. lycopersicum* do not display a great degree of variance at the level of primary metabolites it is important to note that many primary metabolite QTL have been reported for the *S. pennellii* x *S. lycopersicum* introgression line population[47,48]. Thus, variance in individual genes is clearly able to influence primary metabolism and the genome sequence should prove highly useful in cloning the genes responsible for many important traits which contribute to flavor and nutritional quality.

5.9.3 Secondary metabolism

In tomato species, secondary metabolites can be categorized into five families, acylsugars, polyphenols, glycoalkaloids, volatile organic compounds (VOCs) and terpenes[87]. Many genes of secondary metabolism have been well characterized in tomato previously; acylsugar biosynthesis, acetyl-CoA-dependent acyltransferase (*SlAT2*, *Solyc01g105580*)[53]; polyphenol biosynthetic genes, 4-coumarate:coenzyme A ligase (*4CL*, *Solyc07g008360*, *Solyc03g117870*, *Solyc06g068650*, *Solyc12g042460*, *Solyc03g097030*)[88], hydroxycinnamoyl CoA quinate transferase (*HQT*, *Solyc07g005760*)[89] and chlorogenate: glucarate caffeoyltransferase (*SlCGT*, *Solyc01g099020*)[90], chalcone synthase (*SlCHS1*, *Solyc05g053550*;SlCHS2, *Solyc09g091510*)[91], chalcone isomerase (*SlCHI*, *Solyc05g010320*), chalcone isomerase-like 1 (*SlCHL1*, *Solyc05g052240*)[92], trichome specific 3'/5'-myricetin-*O*-methyltransferases (*SlFOMT1*, *Solyc06g083450*) and 7/4'-myricetin-*O*-methyltransferases (*Solyc06g064500*, *SlFOMT2*)[93,94], dihydroflavonol 4-reductase (*SlDFR*, *Solyc02g085020*)[95], (*SlF3'5'H*, *Solyc11g066580*)[96], anthocyanin synthase (*SlAN1*, *Solyc08g080040*)[97]; glycoalkaloid biosynthesis, glycoalkaloid glycosyltransferase (*SlGAME1*, *Solyc07g043490*; *SlGAME17*, *Solyc07g043480*)[98,99], cytochrome P450 (*SlGAME4*, *Solyc12g006460*; *SlGAMER6,8*, *Solyc07g043460*)[99], 2-oxoglutarate-dependent dioxygenase (*SlGAME11*)[99] and aminotransferase-like protein (*SlGAME12*, *Solyc12g006470*)[99]; VOC related genes, aromatic amino acid decarboxylase (*LeAADC1A*, *Solyc08g068680*; *LeAADC1B*, *Solyc08g068610*; *LeAADC2*, *Solyc08g006740*) [100], carotenoid cleavage dioxygenase (*CCD*, *Solyc01g087250*)[101], *ADH2* (*Solyc06g059740*)[102], 13-lipoxygenase (*TOmLOXC*, *Solyc01g006540*)[103], catechol-*O*-methyltransferase (*CTOMT1*, *Solyc10g005060*)[52], neryl diphosphate synthas (*NDPS1*, *Solyc08g005680*)[49], phellandrene synthase 1, (*PHS1*, *Solyc08g005640*)[49]; terpenoid biosynthesis, terpene synthase (*TPS3*, *Solyc01g105870*), *TPS4* (*Solyc01g105880*), i (*Solyc01g105890*), *TPS9,12* (*Solyc06g059930*), *TPS14* (*Solyc09g092470*), *TPS17* (*Solyc12g006570*), *TPS32* (*Solyc01g101180*), *TPS36* (*Solyc06g060180*) and *TPS38* (*Solyc02g079840*)[104]. In the following sectionss a comparison at the sequence and expression level is performed. In cases where this has previously been reported litrature references are provided.

Acyl-sugars

SlAT2 (*Solyc01g105580*) which is involved in trichome acyl sugar metabolism is found in the mQTL region of IL1-3/1-4 for acyl-sugars[53]. As discussed previously, a large deletion in *S. pennellii* was observed (Supplementary Dataset 13). Furthermore, gene expression of SIAT2 was not observed in any *S. pennellii* tissues.

Polyphenols

The genes involved in core polyphenol biosynthesis do not show significant polymorphic differences in their coding region or the gene expression level between *S. lycopersicum* and *S. pennellii* in any tissues. However, the gene encoding *4CL*, which is a one of the key genes involving chemical and functional diversity in phenylpropanoid biosynthesis, shows some amino acid changes in their sequences albeit with a similar gene expression pattern.

On the other hand, the key gene for the production of chlorogenic acids, namely *HQT*[89] displayed a large deletion in the last exon of *S. lycopersicum*, although gene expression did not show a significant difference between *S. lycopersicum* and *S. pennellii.* These results suggested that the structural difference in this enzyme produces a much higher accumulation of CGA in *S. pennellii* fruits[81].

Common flavonoid biosynthetic genes including anthocyanin biosynthesis also display conserved sequences between *S. lycopersicum* and *S. pennellii*; however CHI showed clear difference in its last exon and its gene expression level in mature fruits (260 fold higher expression in *S. pennellii*). Since reduction of CHI activity during fruit ripening is thought to be a key factor for the production of naringeninchalcone which is known as a red fruit ripening marker metabolite, this polymorphic difference may cause the major difference of flavonoid biosynthesis between tomatoes of red ripe fruit and green ripe fruit. The genes encoding *FOMT* involved in trichome specific flavonoid biosynthesis[93,94] showed large difference in amino acid sequence and significant difference of their gene expression level (18 times higher in *S. pennellii* mature fruits). SlFOMT2 additionally has a large deletion in protein structure between *S. lycopersicum* and *S. pennellii* and is not highly expressed in *S. pennellii tissues,* although *S. pennellii* has much larger trichomes.

Glycoalkaloids

Steroidal triterpenoid biosynthesis is conserved between several Solanaceaeous species such as potato and tobacco. However they also display a large chemical diversity caused by the variety of modification types. Recent efforts to investigate steroidal glycoalkaloid biosynthesis in tomato and potato led to the identification of the GAME genes[98,99]. Some characterized GAME genes (*GAME1,2,3,4,6,8,11,12,17*) showed a significant different gene expression level between *S. lycopersicum* and *S. pennellii* in several tissues (13 to 1644 times higher in *S. pennellii* mature fruits). That said, GAME genes (*GAME4,6,8,11,12*) involved in core glycoalkaloid biosynthesis for aglycone construction did not show large difference in their protein sequence between *S. lycopersicum* and *S. pennellii*. It was only in the GAME genes (*GAME1,2,3,17*) involved in the reactions involving modification of glycoalkaloids that significant differences between *S. lycopersicum* and *S. pennellii* were observable. These results indicate that the core biosynthetic genes are conserved in the species, but differences in the genes involved in glycoalkaloid modifications may lead to the wide chemical diversity in the wild species.

VOCs

Volatile organic compounds (VOCs) provide a direct link between metabolism and tomato scent. The VOC contents elevate during the onset of ripening and peak either at or shortly before full ripening being under ethylene-dependent regulation in tomato[54,105]. Identification of genes that impact flavor is an important facet of the effort to improve the quality of tomato fruits[106]. Several genes involved in tomato volatile metabolism as well as the volatiles associated with their functions have been identified [107]. *LeAADC1* genes are aromatic amino acid decarboxylase involved in the synthesis of several volatiles such as phenylacetaldehyde, 2-phenylethanol, 1-nitro-2-phenethane, 2-phenylacetonitril and have been found by linkage to a QTL affecting a target volatile pathway. Although *LeAADC* genes did not show significant transcriptional differences in mature fruits between *S. lycopersicum* and *S. pennellii*, their protein sequences displayed large difference. The CCD[101] involved carotenoid derived volatile showed different expression (9 times higher in *S. pennellii* mature fruits) and largely different protein sequence (deletion of last 3 exons in *S. pennellii*).

ADH2, a gene involved in the production of Hexanol and *Z*-3-hexenol showed lower expression level in *S. pennellii* mature fruit than in *S. lycopersicum*, although their coding sequence was found to be similar. Furthermore, *TOmLOXC*[103] and *CTOMT1*[52] displayed higher expression in *S. pennellii* fruits, but no significant differences in their protein sequences. Such differences may suggest differences in the transcriptional regulatory system of VOC metabolism between *S. lycopersicum* and *S. pennellii.*

Terpenes

Genomic analysis of *S. lycopersicum* revealed 44 TPS genes which are key enzymes in terpenoids biosynthesis and play important roles in interactions to environments[104]. The *NDPS1* and *PHS1* genes involving monoterpene biosynthesis have been found by mQTL analysis targeting terpenoid profiling of glandular trichomes. As described previously[104], both genes showed significant difference in their coding sequences. In a recent study, several TPSs (*TPS3*, *TPS4*, *TPS5*, *TPS9*, *TPS14*, *TPS17*, *TPS32*, *TPS36*, and *TPS38*) have been characterized by gene expression profiling, in vitro array by recombinant protein. These known TPS did not show any gene expression in mature fruit in *S. lycopersicum* or *S. pennellii.* Exceptional to this were *TPS9* (sesquiterpene synthase), *TPS32* (farnesyl diphosphate synthase), *TPS17* (valencene synthase) and *TPS36* (diterpene-producing cembratrienol synthase) which were expressed in *S. pennellii* mature fruit. Although TPS17 had 21 amino acids differences between *S. lycopersicum* and *S. pennellii*, it did not have large deletion/insertion in the coding region. This may represent a different regulation mechanism resulting from diversity of terpenes between *S. lycopersicum* and *S. pennellii*. Furthermore, genes in the gene cluster of TPS in chromosome 1 (*TPS3,4,5, Solyc01g105870*, *Solyc01g105880*, *Solyc01g105890*) showed significant different gene expression and protein sequence. In spite of *TPS3* showing a higher expression in *S. lycopersicum* mature tissue, *TPS4* showed a higher expression level in *S. pennellii* tissues. On the other hand, TPS6 had large difference (1st, 3-7th exons were deleted) in *S. pennellii.* Such genomic and transcriptomic differences can explain the chemical composition variation of terpenes including VOC seen between *S. lycopersicum* and *S. pennellii*.

6 Expression analysis in different samples

6.1 Read mapping

Reads were filtered using Trimmomatic, as for the paired-end DNA libraries (Section 1.8), but with the TruSeq3 single-ended adapter file used when appropriate. The filtered reads were mapped to the genome using BWA (version 0.5.9-rc16), and combined into a single BAM file per sample using Samtools (version 0.1.18). Read counts for each gene were extracted using Samtools, and merged into a single count table using a custom script.

6.2 Statistical analysis

In order to assess differential expression in the different tissues, data was loaded into R and edgeR[108] was used to gauge differential expression. For all comparison sets, the process of library size normalization and dispersion calculation, as suggested by the edgeR manual, was followed.

For the tissue comparison, differential expression between the replicated groups (6 replicates for leaf, 4 replicates for root, 3 replicates each for bud, flower, immature fruit and mature fruit) was calculated using edgeR's exactTest function for all pairs of tissues.

For the hydroponic data set, each condition was tested against the 2 control samples from the same tissue (shoot or root), again using the exactTest function. It should be noted however, that only the control was replicated and p-values are thus estimated only.

In addition to the edgeR analysis, we calculated the RPKM values for the tissue specific samples, using a custom R script, to enable comparison with samples from *S. lycopersicum*.

6.3 Data analysis

In order to assess the data, data was loaded into MapMan and visualized. As expected, large differences in the photosystem could be determined for green versus non-green tissues. This also indicates that the Mercator annotation, at least for the plastidic processes, is largely correct, as these can be identified as being strongly down regulated in the root (Supplementary Figure 23).

6.4 Comparison to the *S. lycopersicum* dataset

The publicly available tissue specific expression dataset from *S. lycopersicum* and *S. pimpinellifolium*[109] was joined with the *S. pennellii* dataset on the gene level by using the one to one gene relationship established in Section 5.2.

In this joint data set, a comparison of the *S. pennellii* and *S. lycopersicum* as well as *S. pimpinellifolium* expression data was performed as follows: The RPKM values were clustered hierarchically using 1-r_s (i.e. 1 – Spearman correlation) as the distance measure and average linkage clustering. This showed that tissues generally aligned between the *S. pennellii* and *S. lycopersicum* (Supplementary Figure 24). However as can also be seen from the correlation between tissues, for fruits no exact stage can be applied. In addition, the *S. pennellii* flower bud correlates slightly better to *S. lycopersicum* flower data (Supplementary Table 29). This is probably due to slightly different flower stages sampled.

7 Annotation of repetitive sequences in *S. pennellii*

7.1 Description

The *S. pennellii* genome assembly was deconstructed into contigs and sequences > 80kb were selected (320 Mb). This sub-genome was used for *de novo* repeat identification with the TEdenovo pipeline[110] from the REPET package (parameters were set to consider repeats with at least 5 copies). The consensus sequences generated were then used as probes for whole genome annotation with the TEannot pipeline[111] from the REPET package. Repeat probes were classified using the REPET dedicated utility followed by semi-manual curation. A similar approach was applied to the *S. lycopersicum* genome. For the estimation of time of insertions, full length LTR retrotransposons were identified using the LTR_Finder program[112], each pair of LTRs was retrieved and aligned using MUSCLE[38], evolutionary distances were calculated with "Distmat" from the Emboss package[40] using the Kimura two-parameter model[113], and distances per site were transformed into ages by applying a rate of 1.3x10^{-8} substitution per site per generation as estimated previously[114].

7.2 Results

Overall, repetitive sequences contribute about 75% (715 Mb) and 80% (626 Mb) of the *S. pennellii* and *S. lycopersicum* genomes, respectively. Excluding assembly gaps (i.e. Ns), repeated elements represent almost 82% of the ATGC space in *S. pennellii*. It appears nonetheless that both repeated and unique sequences contribute to the greater size of the *S. pennellii* genome (Supplementary Figure 10).

Among transposable elements (TEs) LTR-retrotransposons (LTR-RT) are by far the most abundant repeats in both genomes. LTR-RTs constitute 45% of the *S. pennellii* genome. Gypsy-type elements (349 Mb) show predominance over Copia-type LTR-RTs (79 Mb), as observed in *S. lycopersicum* (Supplementary Figure 11). As anticipated, LTR-RTs appear to play a significant role in genome size variation in the *Solanaceae* by representing 355 Mb in *S. lycopersicum*, and 428 Mb in *S. pennellii*, i.e. at least an additional 70 Mb of LTR-RTs in the *S. pennellii* genome.

We assessed the recent activity of LTR-retrotransposons in the two tomato genomes as well as in *S. tuberosum* and found a significant difference in the last million years during which LTR-RT activity (and/or retention) appears considerably higher in *S. pennellii* as compared to *S. lycopersicum* (Supplementary Figure 12). More specifically, it appears that young Copia insertions are much more frequent in the *S. pennellii* genome as compared to *S. lycopersicum (*Supplementary Figure 13).

These results suggest the occurrence of different genome dynamics in these two species since separation from a common ancestor. *S. tuberosum* also appears to have recently undergone significant LTR-RT accumulation. In contrast to LTR-RTs, non-LTR retrotransposons contribute as little as 1.2-3.1% to the tomato genomes and class II TEs (DNA and Helitrons) as little as 3.3-3.5% (Supplementary Figure 11). Incidentally, each of these genomes also contains significant amounts of endogenous pararetrovirus (ca. 6 Mb in *S. pennellii* and 26 Mb in *S. lycopersicum*) suggesting pervasive acquisition of viral DNA through horizontal transfer, reminiscent to previous observation in the *Nicotiana tabacum*[115].

8 Co-location of stress genes with transposable elements

8.1 Assembly of a stress-enriched gene list

In order to assemble a stress gene set, we first made a survey of the tomato-specific literature to compile a first list of genes which were known to be responsive to salt and drought, at the transcriptional and/or at the post-transcriptional level (differential expression and/or differential protein amount/activity between *S. lycopersicum* and *S. pennellii*). The list thus contained a first set of genes encoding:

- ion transporters[116-118] ;
- LEA proteins[119];
- ASR proteins (ABA/water stress/ripening induced)[120];
- dehydrins[121,122];
- osmotins[123];
- glutaredoxins[124];
- WRKY, NAC, DREB, AREB/ABF transcription factors[125-130];
- annexins[131,132];
- c-repeat binding factors[133];
- mitogen-activated protein kinases[134,135];
- chloroplast antioxidative enzymes[136,137];
- histone protein variants[138];
- SAUR proteins (small auxin up-regulated RNAs);
- fibrillins[139];
- several salt/drought responsive genes from large-scale tomato transcriptomics studies[140-142] and from public repositories in curated databases (e.g. Genevestigator: www.genevestigator.com; SOL genomics: http://solgenomics.net).

This first list was then extended to include genes from biochemical/regulatory pathways underpinning metabolic or morphological phenotypes which are involved in the differential responses of *S. lycopersicum* and *S. pennellii* to salt/drought. The additional gene sets thus included:

- the synthesis and degradation of osmoprotectants (proline, sugar-alcohols, etc)[143,144];
- the synthesis and degradation of polyamines[145,146];
- genes related to cuticle metabolism and regulation[27,147].

Additional genes from *Arabidopsis thaliana* responsive to salt/drought/osmotic stress were downloaded directly from the Plant Metabolic Network (http://plantcyc.org) and the sequence of their respective tomato orthologs retrieved through reciprocal best TBLASTN hit (in this case, pathways/gene families included mannitol degradation, proline biosynthesis and degradation, putrescine biosynthesis and degradation, suberin biosynthesis).

The whole stress gene set was then filtered on the basis of the genomic locations of drought or salt-related QTL detected in a panel of *S. lycopersicum* x *S. pennellii* introgression lines[140,148-154] Supplementary Table 20). We further restricted our selection of candidate genes by focusing on four introgression lines showing the highest number of significant QTL for drought and salt stress (IL2-5, IL7-4-1, IL8-3 and IL9-1; see Supplementary Table 21). When multiple stress-related genes were located in the same genomic interval, the selection of the most likely candidate(s) was aided by the magnitude of their differential expression[27] and by the extent of the genetic differences found between *S. lycopersicum* cv. M82 and *S. pennellii*. This subset of genes was thus considered to contain candidates for QTL associated to drought or salt tolerance. Detailed analyses of coding and promoter sequences were conducted on this subset of genes according to the methods outlined in Section 5.8 ("Selected interspecific variation").

8.2 Creation of stress genes vs non-stress gene sets

The stress related gene set, described above, comprising of 389 genes (Supplementary Table 22) was extended to include genes in orthologous clusters of the stress sets as these are likely to play a role in stress as well. Non-stress related clusters were also compiled to form the set of non-stress related genes. To prevent biases caused by the large transposon-related gene clusters in *S. pennellii*, and since all members of large gene clusters are unlikely to be involved in a single biological process, it was decided not to include any cluster having more than 400 members between *S. pennellii* and *S. lycopersicum* in the non-stress gene sets (there were no stress related clusters above this threshold in any case). This procedure, although based on an arbitrary cutoff, resulted in gene sets with similar sizes and proportion of stress genes across the two species, with a stress gene set of 3,067 genes plus 25,344 non-stress genes for *S. lycopersicum* and a stress gene set of 2,844 for *S. pennellii* plus 24,848 non-stress genes. (Supplementary Table 25)

For analyses which required a 1:1 correspondence between the *S. pennellii* and *S. lycopersicum*, the set of 21,662 orthologous pairs described in Supplementary Note section 5.2 were used. This also resulted in a narrower stress gene set of 2,413 genes in both species.

8.3 Co-location of stress genes by transposable element type

The transposable elements, annotated in Supplementary Note Section 7, were tested for direct overlap and co-location within a 5 kb window (excluding overlap) of the stress gene sets in both the *S. pennellii* and *S. lycopersicum*. The ratio of stress genes within the overlapping or nearby genes was tested using Fisher's exact test, with p=0.05 taken as the threshold of significance.

Gypsy transposable elements showed a depletion of stress genes in *S. pennellii* overlapping (631 stress genes of 6,781 total overlapping, p=0.000491, odds ratio=0.849), but an enrichment of stress genes within a 5kbp window (1,075 of 9,602, p=0.00214, odds ratio=1.135). Copia transposable elements showed a stronger depletion of stress genes overlapping (442 of 5125, p=2.006e-6, odds ratio=0.777) and also a stronger enrichment within a 5 kb window (1,003 of 8,219, p=4.179e-10, odds ratio=1.301). (Supplementary Dataset 8)

In *S. lycopersicum*, Gypsy elements showed a slight, non-statistically significant, depletion of stress genes, both in overlap (420 of 4050, p=0.3668, odds ratio=0.950) and in a 5kbp window (1,057 of 10,055, p=0.280, odds ratio=0.957). Copia elements showed a signal similar to, though weaker than that in *S. pennellii,* with non-statistically significant depletion of stress genes overlapping (431 of 4,256, p=0.14, odds ratio=0.921) but statistically significant enrichment within a 5 kb window (1,152 of 9,916, p=0.0010, odds ratio=1.140).

These results indicated a potential difference between *S. pennellii* and *S. lycopersicum*, and since the results were consistently stronger for Copia compared to those for Gypsy, we focused exclusively on Copia elements in all later analysis.

8.4 Detailed analysis of co-location of stress genes with Copia elements

The 21,662 orthlogous gene pairs were used for a more detailed analysis of the distance relationship between Copia and the stress genes. Copia elements overlapping each gene (0) as well as in the regions between 100-0, 200-100,300-200,400-300,500-400, 1000-500, 1500-1000, 2000-1500, 2500-2000, 3000-2500, and 3500-3000 from the gene were then counted and overlapped with the stress set. This number was divided by all the genes carrying a Copia element in the respective regions. This was done separately for *S. pennellii* (A) and *S. lycopersicum* (B), and the ratio between these values (C) (Supplementary Figure 17). At the same time from all genes having a simple ortholog the same number of stress genes was randomly drawn 1000 times without replacement, and the distribution is shown as an underlying box plot. Interestingly for *S. pennellii* a strong enrichment in the 100bp region around the gene was observed whilst in *S. lycopersicum,* depletion in the 200-100bp region was found. This resulted in a significant difference in the ratio at the 200-100bp BIN (p-value for enrichment ~0.04). (Supplementary Dataset 9)

8.5 Enrichment of salt stress responsive genes near Copia elements

The hydroponic salt stress dataset, analyzed in Supplementary Note Section 6.2, was used to create sets of up- and down-regulated genes under moderate salt stress for both root and shoot. These lists were filtered for genes with FDR values of 0.01 or below, an absolute log fold change of at least 2, and limited to those in the orthologous gene pair set. The 5 kb window around Copia elements contained 6,273 of the 21,662 orthologous gene pairs, which were then tested for over-representation against the sets of up/down regulated genes from root/shoot dataset, using Fisher's exact test with 0.05 as the threshold of significance.

The Copia co-located genes showed significant enrichment for up-regulated genes in both root (111 of 324, p=0.0359, odds ratio=1.283) and shoot (204 of 580, p=0.000979, odds ratio=1.342). There was no enrichment of down-regulated genes in either root (44 of 144, p=0.7124, odds ratio=1.08) or shoot (45 of 166, p=0.668, odds ratio=0.912).

The orientation of the Copia element relative to the gene was also tested, but found that the proportions of responsive genes with upstream vs downstream Copia elements was almost identical, with 64 salt-responsive root genes having Copia upstream vs 66 downstream, and 114 salt-responsive shoot genes having Copia upstream vs 115 downstream. (Supplementary Dataset 10)

8.6 Validation of link between salt stress responsive genes and Copia elements

To reconfirm the relationship between Copia elements and salt-responsive treatments, an additional salt-stress experiment was carried out. Seeds of *S. pennellii* (LA0716) were germinated at 22°C in standard glass pots containing MS agar (5 seeds/pot). At the 4-leaf stage, seedlings were transferred to new MS agar pots containing 100 mM NaCl. Control seedlings were transferred to new MS agar pots containing no salt. After 72 hours from the application of salt stress, whole seedlings were harvested, immediately frozen in liquid nitrogen and reduced to a fine powder. Samples were then stored at -80°C.

RNA extraction and sequencing of 3 salt-stressed and 3 control samples was carried out as for previous RNA-Seq data in the project (Sections 4.3 and 4.4), yielding 40.6 million reads totaling 4.10 Gb of sequence data.

This RNA data was filtered and mapped as in Section 6.1, resulting in a post-filtered dataset of 40.1 million reads, totaling 4.01 Gb, with a 91.3% alignment rate against the *S. pennellii* transcriptome.

Differential expression between the stress and control groups was determined using EdgeR, as in Section 6.2, using a false discovery rate threshold of 0.01 as the threshold of significance. In total, 440 genes were found to be up-regulated, and 642 genes down-regulated. These were then filtered against the 21,662 orthlogous genes, resulting in 362 up and 540 down regulated genes respectively.

As before, Fisher's exact test was used to determine if the 6,273 Copia co-located of 21,662 orthologous gene pairs were enriched for these responsive genes. Of the 362 genes found to be up-regulated, 128 were found to within a 5K window of Copia elements (p=0.008, odds ratio=1.349), indicating a statistically significant enrichment. Down-regulated genes showed a minor, not statistically significant enrichment (171 of 540, p=0.1636, odds ratio=1.141). (Supplementary Dataset 10)

8.7 Non-stressed, species-specific expression of Copia co-located genes

The cross-species comparison from[27] was used to create two sets of the 1000 genes with relatively higher expression in *S. pennellii* and *S. lycopersicum* respectively, under non-stress conditions, from our list of 21,662 orthologous pairs. We tested these against genes which have non-overlapping Copia element located within 500bp either up or downstream of the gene, only in that species. There were 501 genes with an upstream Copia, 470 of which are only in *S. pennellii*, and 591 genes with a downstream Copia, 519 of which are only in *S. pennellii*. For *S. lycopersicum*, the corresponding counts were 534 genes with a Copia upstream, 503 of which are species specific, and 642 downstream, 570 of which are species specific.

We found a clear expression-reduction effect for genes with Copia elements within 500bp upstream, where 27 genes with Copia elements in *S. pennellii* only showed lower expression in *S. pennellii* vs 17 with higher expression. Correspondingly, 33 genes with Copia elements upstream only in *S. lycopersicum* showed lower expression in *S. lycopersicum*, while only 22 showed higher expression. Testing with Fisher's exact test indicated a p-value of 0.0439, and an odds ratio of 2.36.

The corresponding test of genes with species-specific Copia elements downstream did not reveal any species-specific effect. Furthermore, no additional effect could be seen using larger windows up to 5kbp. We concluded that a likely cause of the expression reduction in genes with nearby upstream Copia elements was loss or damage to the promoter sequence. (Supplementary Dataset 11)

8.8 Enhanced response of Copia co-located genes to stress

The *S. pennellii* and *S. lycopersicum* drought stress datasets from [155] were used to create sets of species specific up and down regulated genes, using Bowtie[4] and edgeR[108]. Due to lack of replication, the dispersion parameter was manually set to 0.1. 4 sets of responding genes were created, comprising of genes differentially responding on each species. Each list was ranked by false discovery rate (FDR), and the top 500 genes in each species in each direction was retained.. This selection method was used to compensate for the relatively higher sensitivity to differential expression in *S. lycopersicum*, which was primarily due to the limited number of reads in the *S. pennellii* control dataset (~2.9 million vs. 16/19/22 million the other datasets).

These lists of responding genes were then compared across species, and the common genes removed, resulting in a list of 293 species unique up-regulated genes in each species, and 299 unique down-regulated genes in each species.

For comparison, the sets of genes with non-overlapping Copia elements within 5 kb only in one species were created, with 3,862 such genes identified in *S. pennellii* and 4,456 in *S. lycopersicum*. The uniquely responding genes were then tested against the set of uniquely Copia co-located genes in each species using Fisher's exact test.

In *S. pennellii*, the Copia co-located set was found to be enriched for both up-regulated (66 of 293, p=0.0379, odds ratio=1.346) and down-regulated (69 of 299, p=0.0221, odds ratio=1.390) uniquely responsive genes. *S. lycopersicum* showed no enrichment in uniquely down-regulated (59 of 299, p=0.7733, odds ratio=0.949) genes, and a slight though non-significant depletion in uniquely up-regulated genes (49 of 293, p=0.1093, odds=0.773). (Supplementary Dataset 12).

We also tested the correlation pattern across all our RNA-Seq datasets (Supplementary Table 28) between genes specifically down-regulated in *S. pennellii*. Within those related to Copia elements we found an average correlation of r=0.138. Within those not associated to Copia we only observed an average correlation of r=0.112. Due the large number of correlations, this small difference was highly significant (p<0.001).

9 The M82 genome

9.1 Plant material

S. lycopersicum (LA3475, cv. M82) seeds were sourced from CM Rick Tomato Genetics Resource Center, Univ. of California at Davis. Seeds were grown on sterile media, propagated twice then transferred to standard tomato soil and grown in a greenhouse.

9.2 Nucleic DNA extraction and sequencing

Nucleic DNA extraction and paired-end Illumina sequencing was performed as for *S. pennellii.*

9.3 Sequencing yield and quality filtering

The 9 paired-end lanes of the M82 resequencing run yielded 574 million raw reads, totaling 47.4 billion bases. This was filtered using Trimmomatic (V0.13)[3], to remove Illumina adapter sequences and low quality bases, and resulted in 551 million reads comprising 44.4 billion bases.

The quality filtered libraries were aligned using BWA against the combined *S. lycopersicum* nuclear and organelle sequences as reference, and merged into a single alignment file using Samtools. The alignment revealed a mean insert size of 188.7, with a standard deviation of 20.6. The alignment rate was 96.52%, indicating the high quality of the sequenced data.

9.4 Variant calls

The merged alignment was filtered for duplicates and a 'pileup' created, both using Samtools. This was then used to call variants vs. the *S. lycopersicum* reference sequence. Manual examination of variants revealed a relatively large number of poorly supported and conflicting variants, with >2 alleles indicated in many cases. This was not considered realistic in a largely homozygous diploid organism.

To resolve such complex variants, which are likely the result of mapping artifacts, the alignment was then processed by the Short Read Micro Aligner (SRMA) tool, which applies a more complex local re-alignment and consensus calling strategy. The resulting variants were filtered for those with >60% support for alternate sequences and coverage by at least 5 reads.

The resulting 1,294,919 variants, consisting of 1,144,933 SNPs and 149,986 INDELs were applied to the reference to create a first-iteration M82 consensus genome sequence. The alignment process was repeated, to allow the additional variants lost due to poor mapping in more divergent regions to be found and incorporated. This resulted in an additional 43,591 SNPs which were applied to the first-iteration M82 genome to create the final genome.

Thus we could in total identify 1,188,524 SNPs and 149,986 INDELs, totaling 1338510 variant calls. We then analyzed the SNP rates between Heinz and M82, *S. pimpinellifolium* and Heinz and *S. pimpinellifolium* and M82. Interestingly in the regions where the SNP rate was high between M82 and Heinz, the SNP rate between M82 and *S. pimpinellifolium* dropped strongly, indicating also potential introgression of the *S. pimpinellifolium* into the M82 genome (Supplementary Figure 7-9). Whilst this has been previously suggested by some of us for Chromosomes 4, 5, 11 and 12 using RNA-Seq data [27], this more refined analysis showed that especially in Chromosome 11, there seems to be a very strong effect in the gene poor (heterochromatic region), which could previously not be identified. Furthermore we did not find strong evidence for an introgression on Chromosome 12. That there are introgressions of *S. pimpinellifolium* into the cultivated tomato has been observed also for the Heinz genome[22] and it was recently discussed that introgression breeding might have dragged in "wild species" regions into cultivated tomatoes through linkage drag[156].

9.5 Summary

As previously established, the SNPs/Indel Rate for M82 is not uniform across the genome. Whilst Chromosomes 4,5 and 11 showed an overall much higher diversity rate, the Chromosomes 1,6-10 and to a lesser extent 2, 3 and 12 showed peaks of difference mostly in gene rich regions (Supplementary Figure 8). Chromosome 10 showed an additional peak at around 50 Mb.

10 Wax and cutin analysis

10.1 Wax analysis

Two mature leaflets were collected from each of 5 plants of *S. lycopersicum* cv. M82 and *S. pennellii* (i.e. 10 leaves per species and 5 independent biological plants per species) and were left in the dark for 2 hours to ensure stomatal closure. The leaves were scanned to measure surface area. 40 µg of each of tetracosane and tetracontane, dissolved in chloroform, were applied directly to each pair of leaflets as internal standards. Leaflets were rinsed twice with 75% ethanol for 1 minute then allowed to dry. Waxes were extracted by swirling each leaflet in 40 mL of chloroform for 1 minute. The two leaflets from the same plant were pooled, using the same chloroform to extract both. The extracts were air dried, re-suspended in 6 mL of chloroform and filtered. 800 µL of each of the 10 wax samples (5 from *S. lycopersicum* and 5 from *S. pennellii)* was dried in a stream of nitrogen gas at 40°C and derivatized with 10 µL of bis-N,O-(trimethylsilyl)trifluoroacetamide mixed with 10 µL of pyridine, for 30 minutes at 70°C. The derivatized samples were dried in a stream of nitrogen gas at 50°C, re-suspended in 100 µL of chloroform and analyzed by gas chromatography (Agilent GC6850 with cool-on-column inlet, flame ionization detector, Agilent DB-1 30 ft column and helium carrier gas). The oven temperature was held at 50°C for 2 minutes, increased by 40°C/min to 200°C, increased by 4°C/min to 235°C, held for 15 min, increased at 10°C/min to 315°C and held for 15 minutes. Compounds were identified by GC-MS analysis using a model 6890 GC (Agilent) coupled to a GC Mate

II mass spectrometer (JEOL) operating in electron impact mode, comparing spectra with those in the AOCS Lipid Library (http://lipidlibrary.aocs.org/ms/masspec.html).

10.2 Cutin analysis

One mature leaflet from each of 5 plants of *S. lycopersicum* cv. M82 and *S. pennellii* was collected and scanned to measure surface area. The leaves were then delipidated and depolymerized as described in[157]. For the depolymerization, the base catalysis approach was taken, but using 3 times greater volumes than those listed and incubating the samples at 60°C for 3.5 hours. 50µg of each of heptadecanoate and pentadecalactone were added beforehand as internal standards. 1 mL of each of the depolymerized cutin samples was derivatized and analyzed by GC as described above, except using a different heating program: the oven temperature was held at 50°C for 2 minutes, increased by 40°C/min to 120°C, held for 2 minutes, increased by 10°C/min to 320°C, then held for 15 minutes.

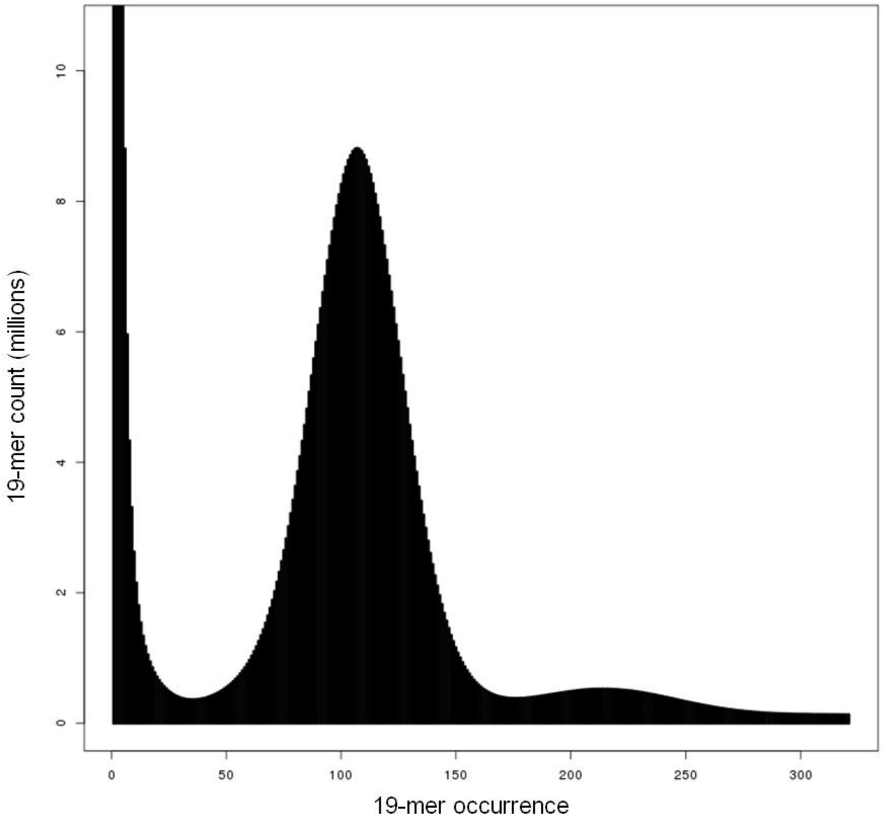
10.3 Quantitative PCR (qPCR) validation for cuticle related genes

Tissue from young expanding leaves was collected and immediately frozen in liquid nitrogen. Total RNA was extracted using the TRIzol® Reagent (Life Technologies, Cat. # 15596018, http://www.lifetechnologies.com) according to the manufacturer's instructions. 4 µg of RNA was treated with one unit of RQ1 DNase (Promega, Cat # M6101, http://www.promega.com) for 30 minutes at 37^{o}C. The DNase was then inactivated by addition of 1 uL RQ1 DNase Stop Solution and incubation at 65^{o}C for 10 minutes. Subsequently, cDNA was synthesized from 4 µg using the SuperScript®II Reverse Transcriptase (Life Technologies, Cat. # 18064014; www.lifetechnologies.com) and an oligo(dT) primer. cDNA concentrations were adjusted to 100 ng/µl and the HotStart-IT SYBR Green qPCR Master Mix (Affymetrix, Cat. # 75760, www.affymetrix.com) was used for the qPCR analysis. The 12 µl reaction volumes consisted of 6 µl qPCR mix, 0.25 µl ROX, 4.25 µl H_2O, 0.5 µl cDNA and 1 µl of a mix of forward and reverse primer (5 µM each), as suggested in the manufacturer's instructions. Reactions were performed using a ViiA™ 7 Real Time PCR System (www.lifetechnologies.com) with the following conditions: 95°C 10 min, followed by 40 *cycles* of 95°C 15 s, *60°C* 60 s, then one cycle each of 95°C 15 sec and 60°C 1 min, followed by a standard melt curve. All reactions were performed using three technical replicates and four biological replicates. Expression values were normalized to values for an actin gene and statistics performed using a two tailed t-test. Primers used are listed in Supplementary Table 19.

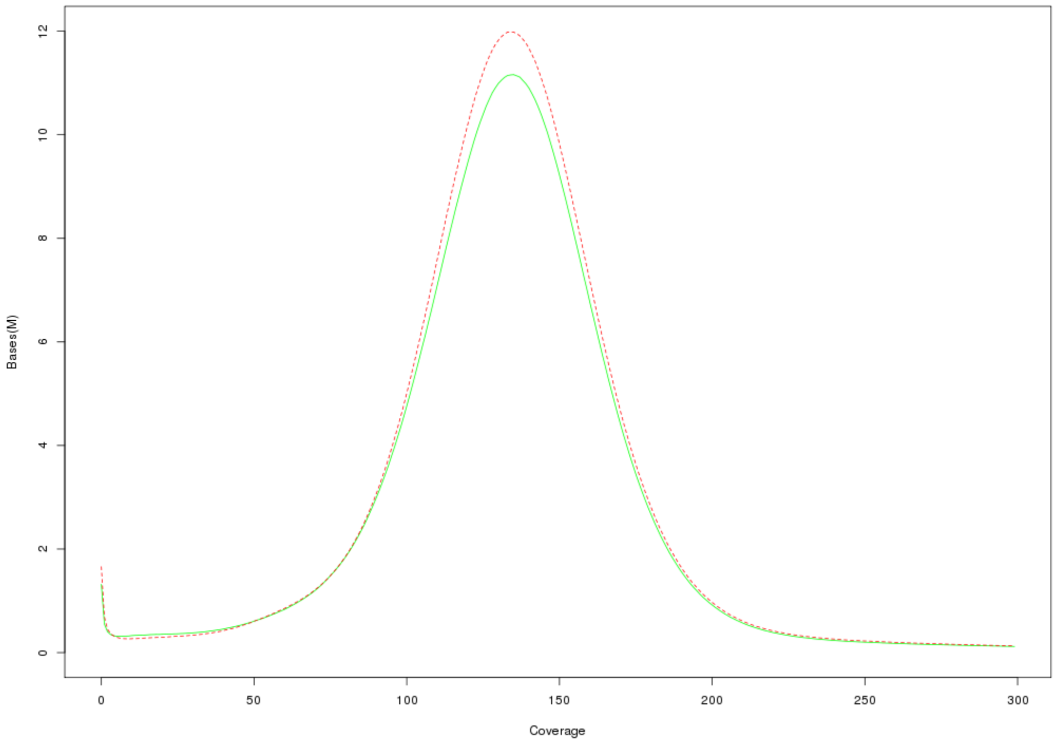
11 Sequencing data summary

In total all new sequencing data was generated for the assembly of *S. pennellii* (Sections 1.4, 1.5, 1.6) and provided all new Illumina data. BAC end sequences had already been deposited with SOL previously (Section 1.7). Full BAC sequence data came from published data (mainly Kamenetzky[2]) but was now corrected by one MiSeq run at very high coverage for 8 BACs. In addition new RNAseq data has been generated for the identification of genes (Section 4.2, 4.4) and for the expression of tissues specific data. Furthermore a specific stress treatment was performed in *S. pennellii* for the analysis of TE effects 8.6 to add to the public data set by Filipps and colleagues[155] and to have statistically replicated data. For the anchoring of Scaffolds to the genome public markers and the published RAD-Seq data set Chitwood[14] and colleagues was used. Also an additional available mature fruit dataset was used.

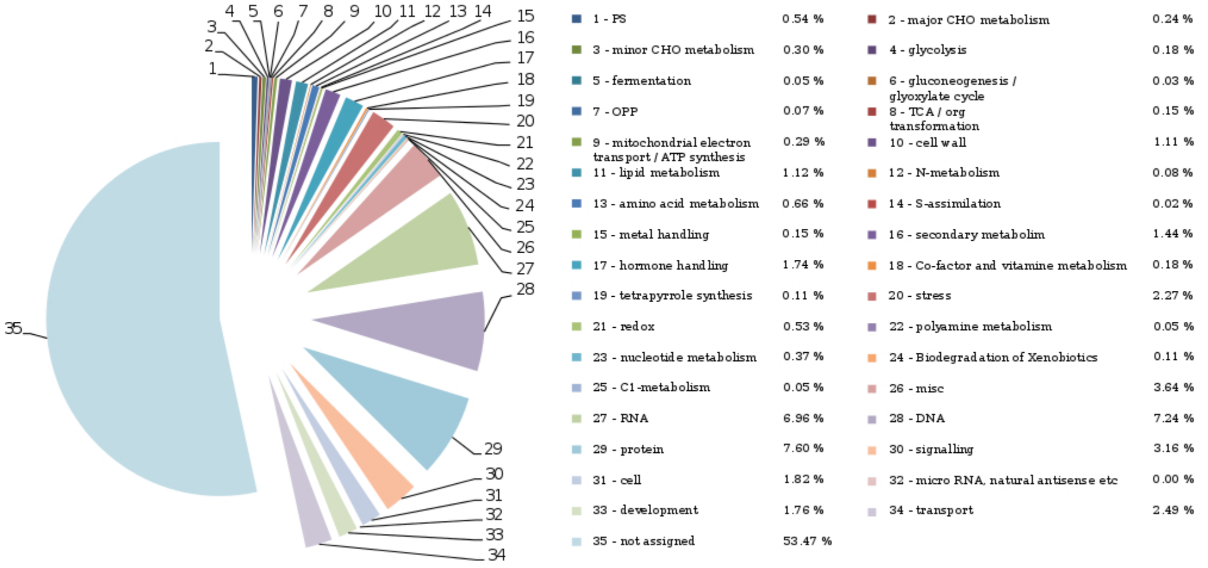
Supplementary Figures



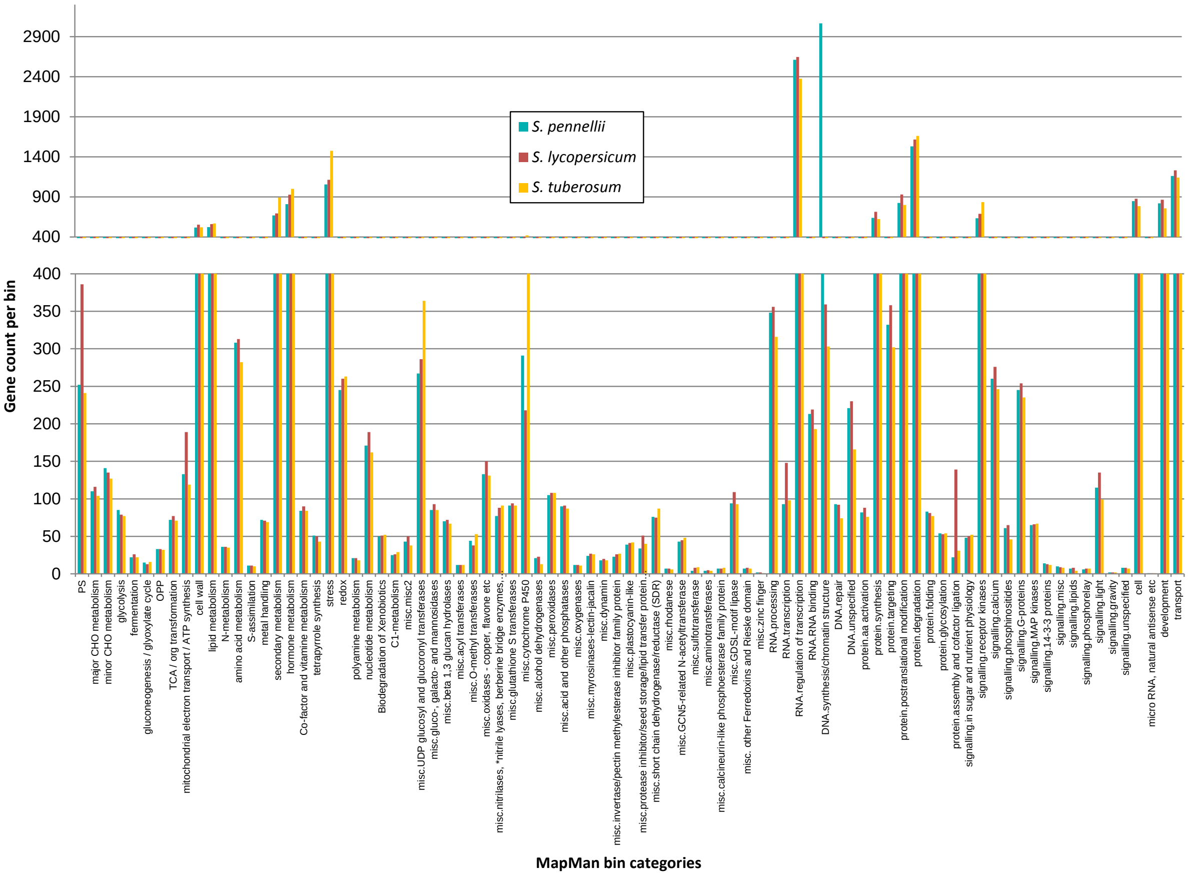
Supplementary Figure 1: Genome size estimation using 19-mer frequency distribution. The 19-mer frequency (x-axis) is plotted against the occurrence of the k-mer (y-axis). The main Poisson - shaped distribution peak represents the unique 19-mers found in the *S. pennellii* genome. The smaller second peak represents the number of 19-mers that are found twice in the genome. The initial peak (cut at 11 million) represents unique/low occurrence 19-mers and is believed to be erroneous 19-mers (due to sequencing errors).



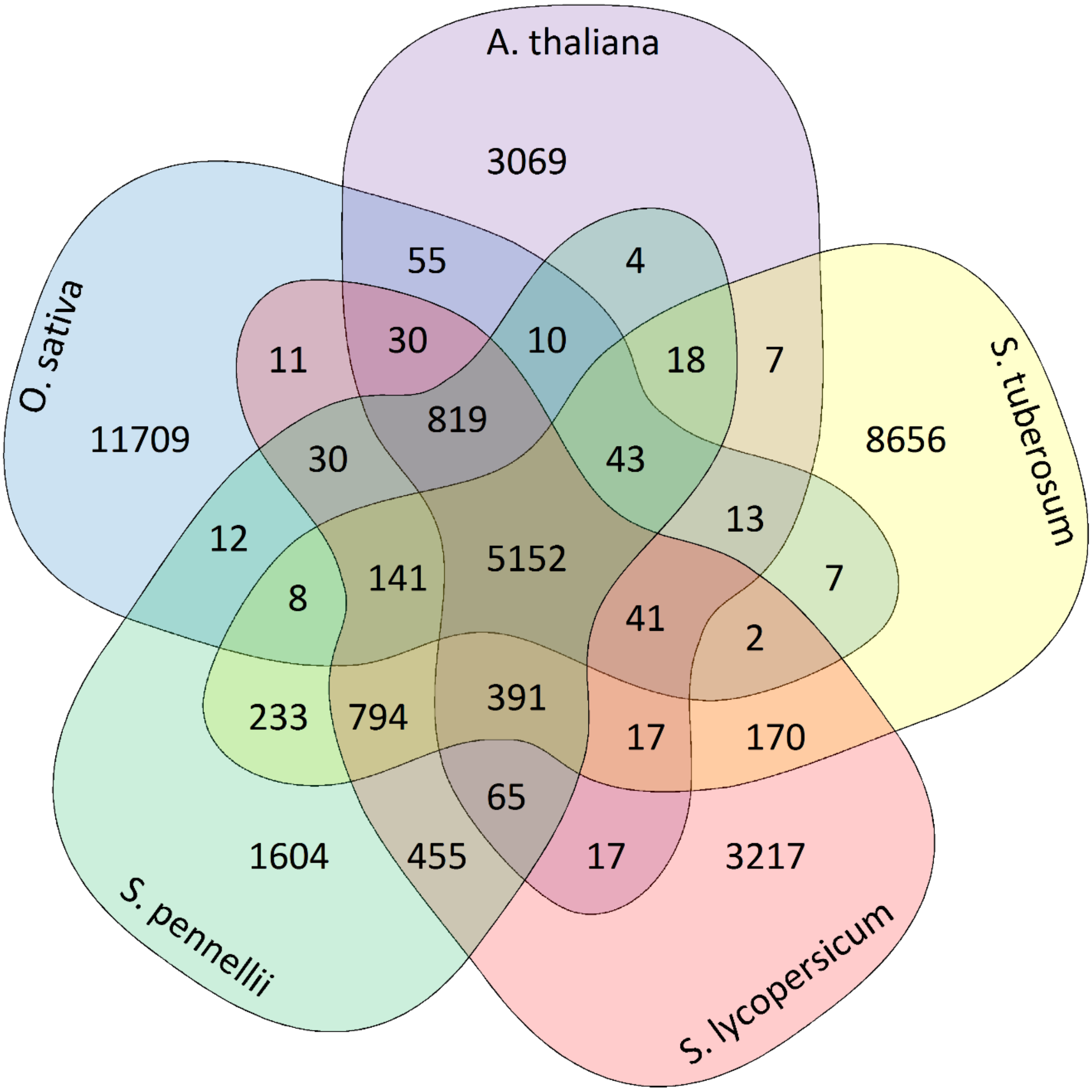
Supplementary Figure 2: Histogram of base coverage by realigning reads for unfilled (green solid line) and gap filled (red dashed line) assemblies. The filled assembly shows higher base count around the peak coverage (~135). However, a small increase in erroneous bases (average <5) can be seen in the gap filled assembly.



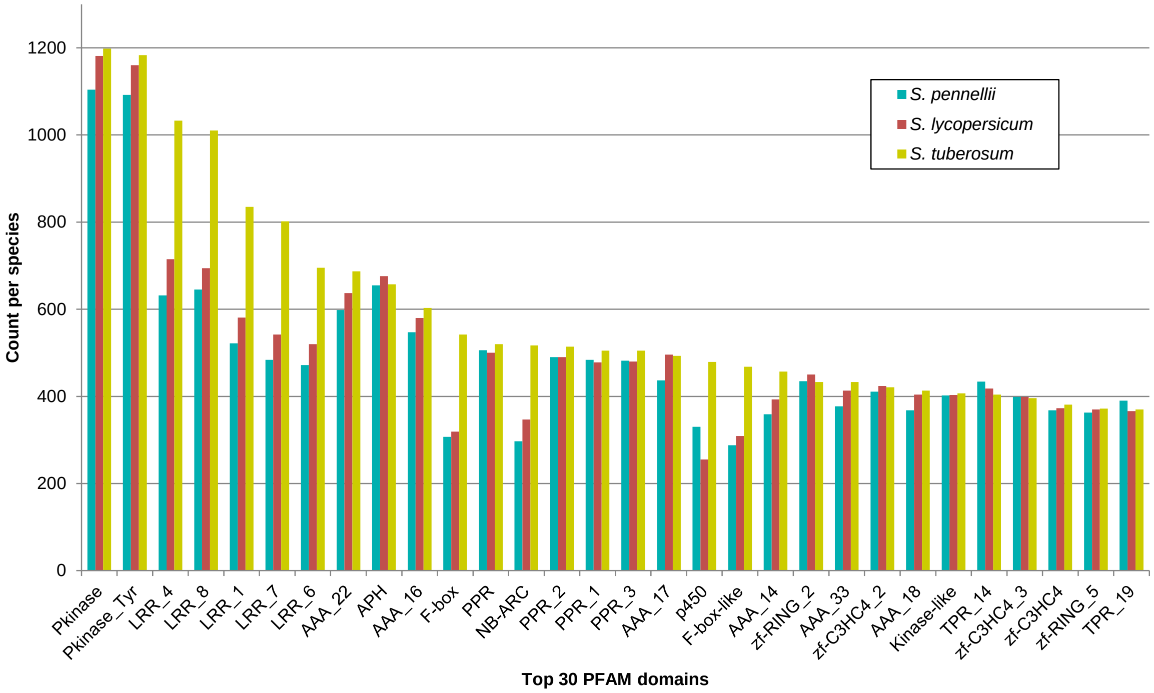
Supplementary Figure 3: Functional classes assigned to the protein coding genes of *S. pennellii* by the Mercator pipeline. The protein coding genes from the *S. pennellii* Augustus gene models were subjected to the Mercator functional prediction pipeline. The resulting protein predictions were classified into MapMan bins according to function. The top level MapMan bins are depicted above.



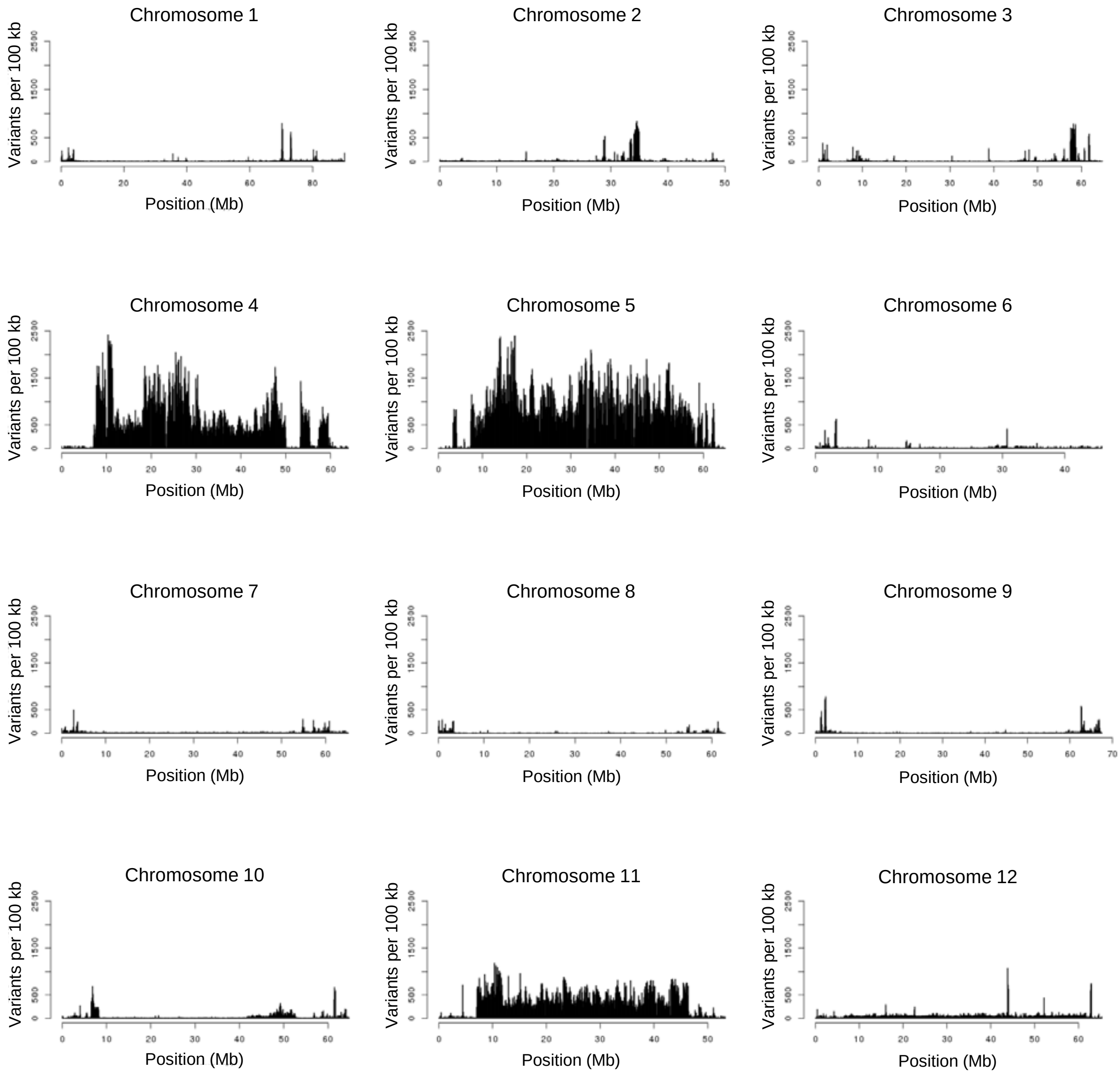
Supplementary Figure 4: Distribution of genes into MapMan functional categories for *S. pennellii*, *S. lycopersicum* and *S. tuberosum*. The genes from *S. lycopersicum* and *S. tuberosum* were subjected to the Mercator pipeline and the resulting classifications plotted together with the *S. pennellii* classifications. Six high-occurrence top levels MapMan functional bins were further sub-classified to provide greater granularity. The y-axis is split at 400 with a broader scale used for the upper part to allow better visualization of the data.



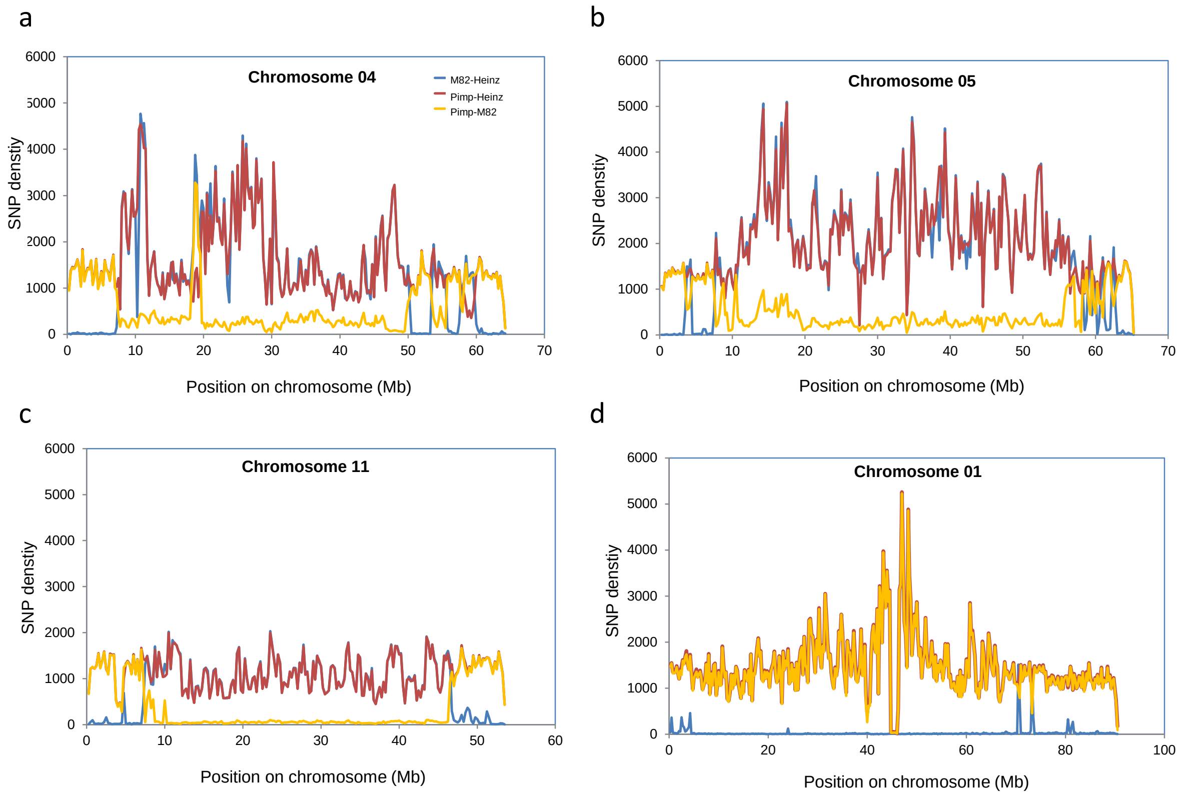
Supplementary Figure 5: Venn diagram of the protein sequence distribution among 5 species: *S. pennellii*, *S. lycopersicum* (ITAG v2.3), *S. tuberosum* (PGGSC v3.4), *Arabidopsis thaliana* (TAIR v10) and *Oryza sativa* (MSU v7). Protein sequences from all 5 species were clustered using BlastP and OrthoMCL and the results depicted on the 5-way Venn diagram**.**



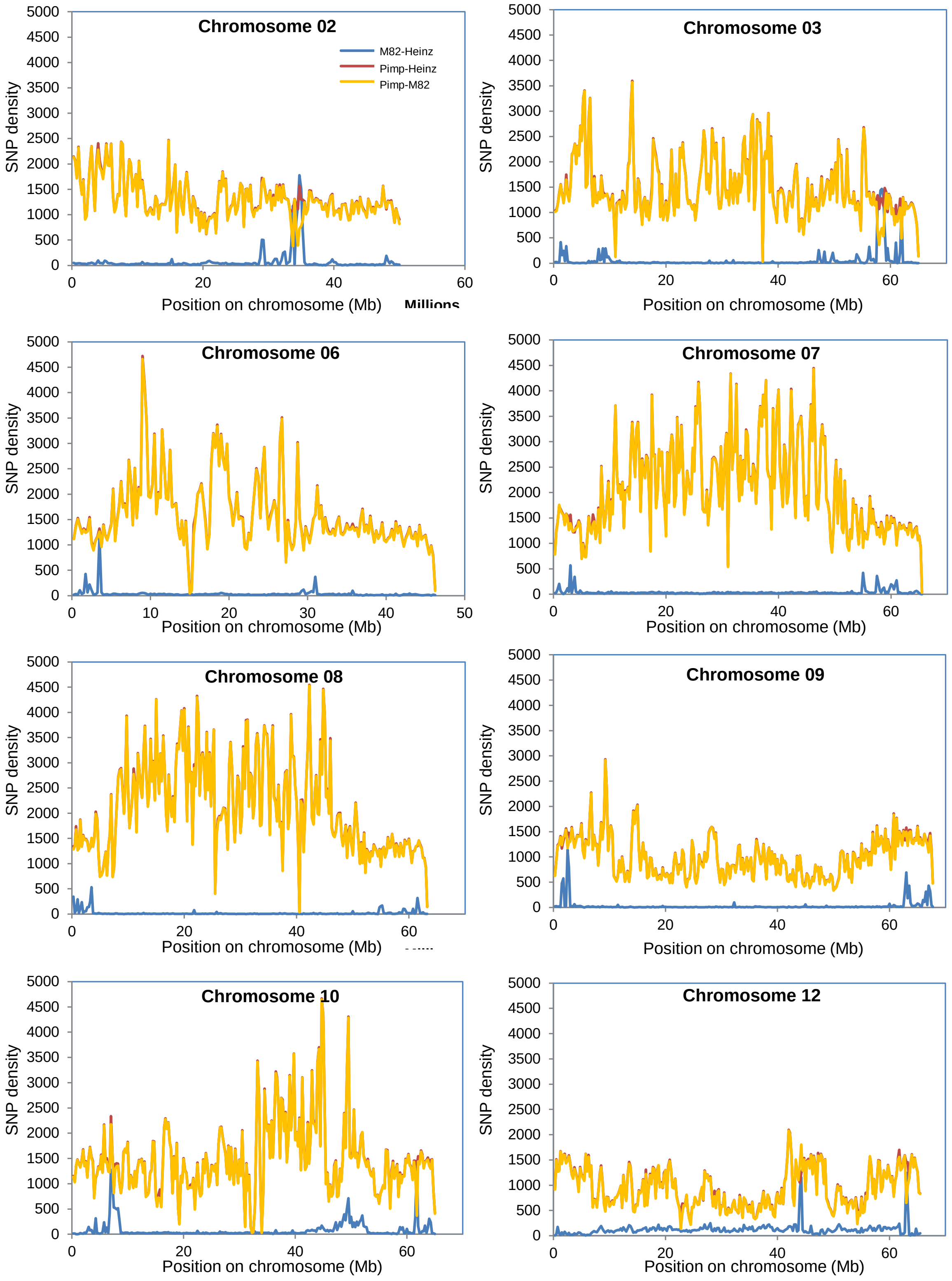
Supplementary Figure 6: Comparison of 30 protein domains representation between *S. pennellii*, *S. lycopersicum* and *S. tuberosum*. The Top 30 represented protein domains from *S. lycopersicum* were selected as a reference. The number of predicted genes which are included in these domains was plotted together with the corresponding numbers from the other two species.



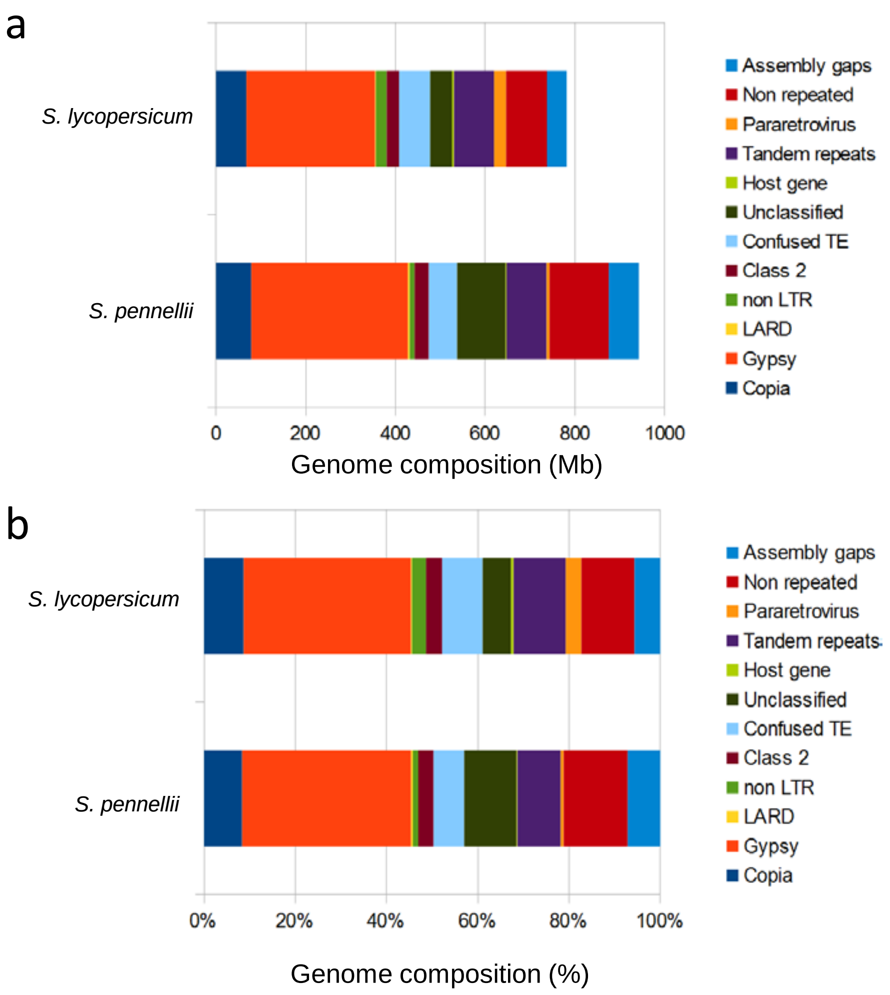
Supplementary Figure 7: Chromosomal variant density distribution between *S. lycopersicum* cv. M82 and *S. lycopersicum* cv. Heinz. A high variant density can be seen in chromosomes 4, 5 and 11 while only localized areas of the remaining chromosomes display such high variant density.



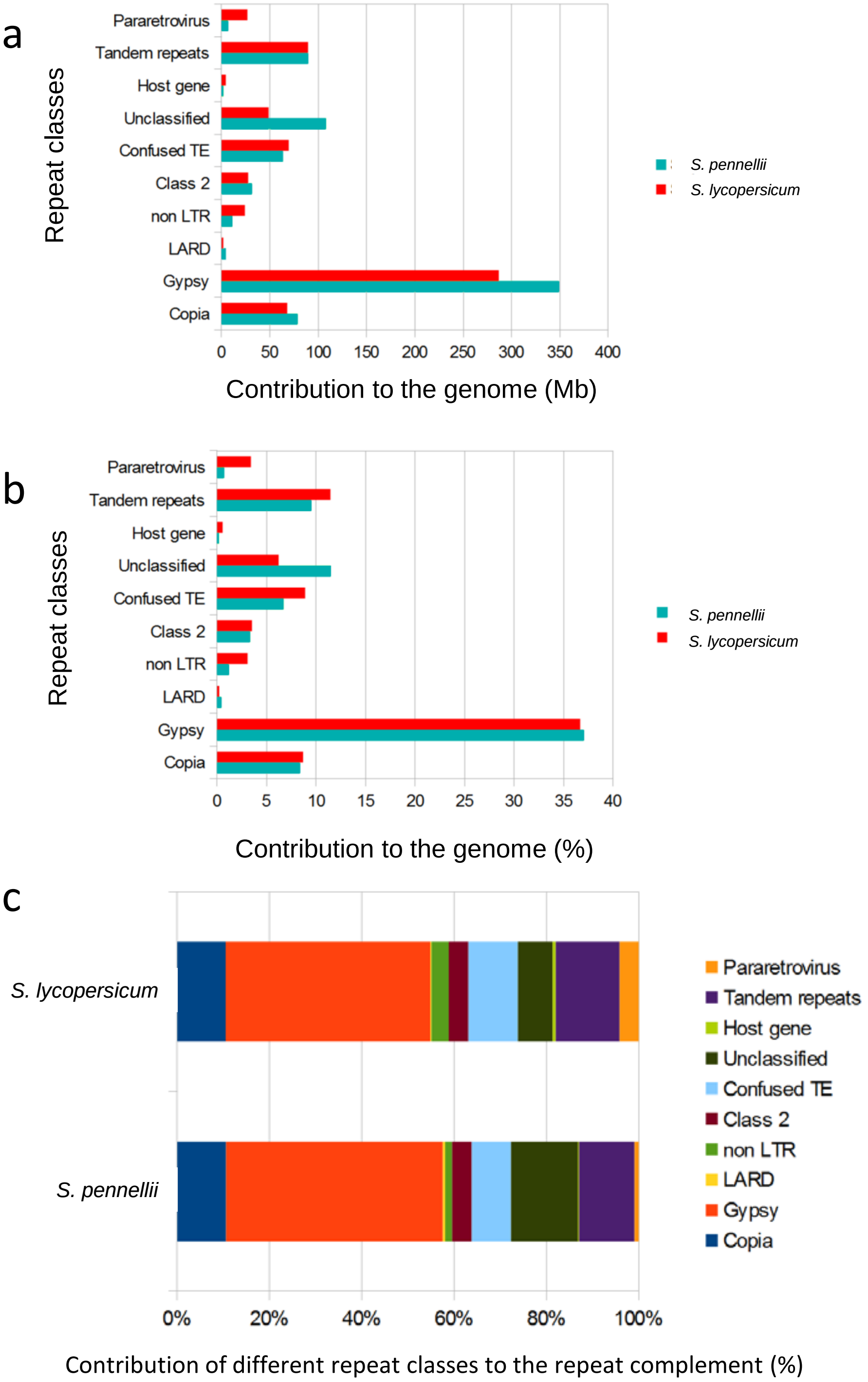
Supplementary Figure 8: SNP density distribution *between S. lycopersicum* cv. Heinz, *S. lycopersicum* cv. M82 and *S. pimpinellifolium* for chromosome 4 (a), 5 (b), 11 (c) and 1 (d). The blue line indicates the M82 versus Heinz cultivars. The red line indicates the *S. pimpinellifolium* (pimp) versus Heinz and the orange line indicates pimp versus M82. These plots show a greater similarity between M82 and Heinz in chromosome 1 (d), but a greater similarity between M82 and pimp in chromosomes 4 (a), 5 (b) and 11 (c).



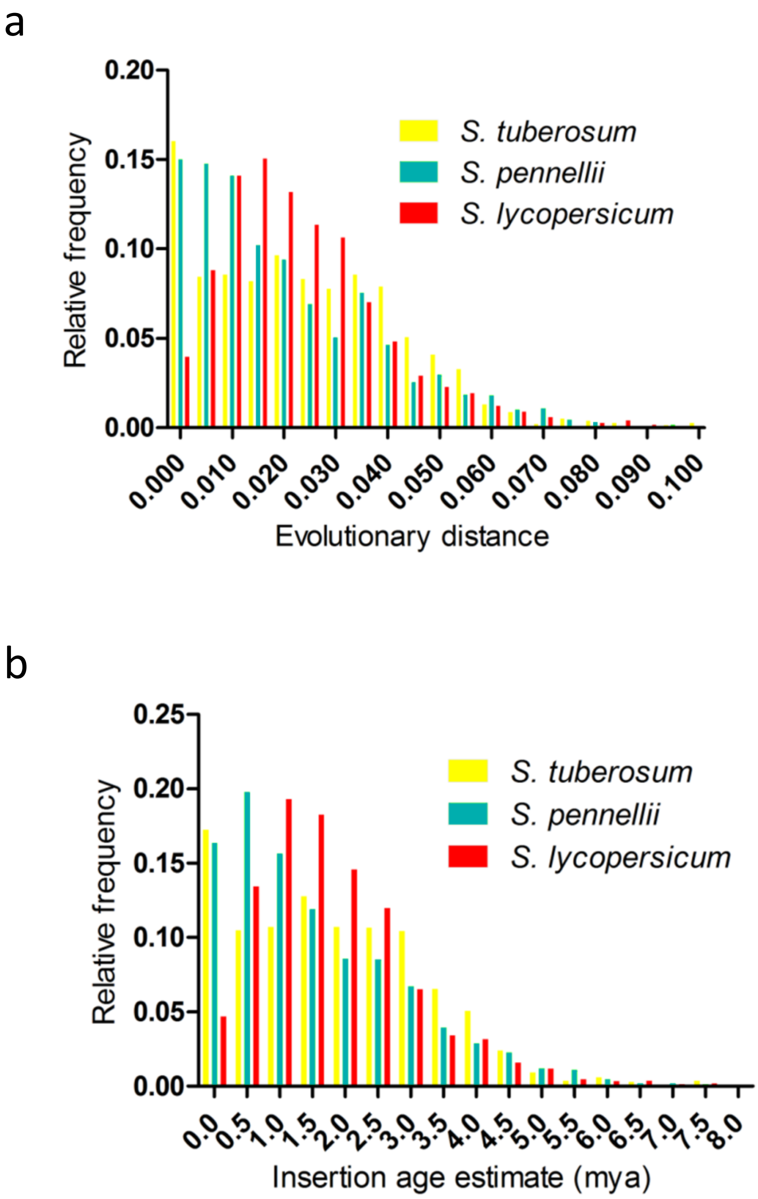
Supplementary Figure 9: SNP density distribution between *S. lycopersicum* cv. Heinz, *S. lycopersicum* cv. M82 and *S. pimpinellifolium* for chromosomes 2, 3, 6-10 and 12. The blue line indicates the M82 versus Heinz cultivars. The red line indicates the *S. pimpinellifolium* (pimp) versus Heinz and the orange line indicates the pimp versus M82.



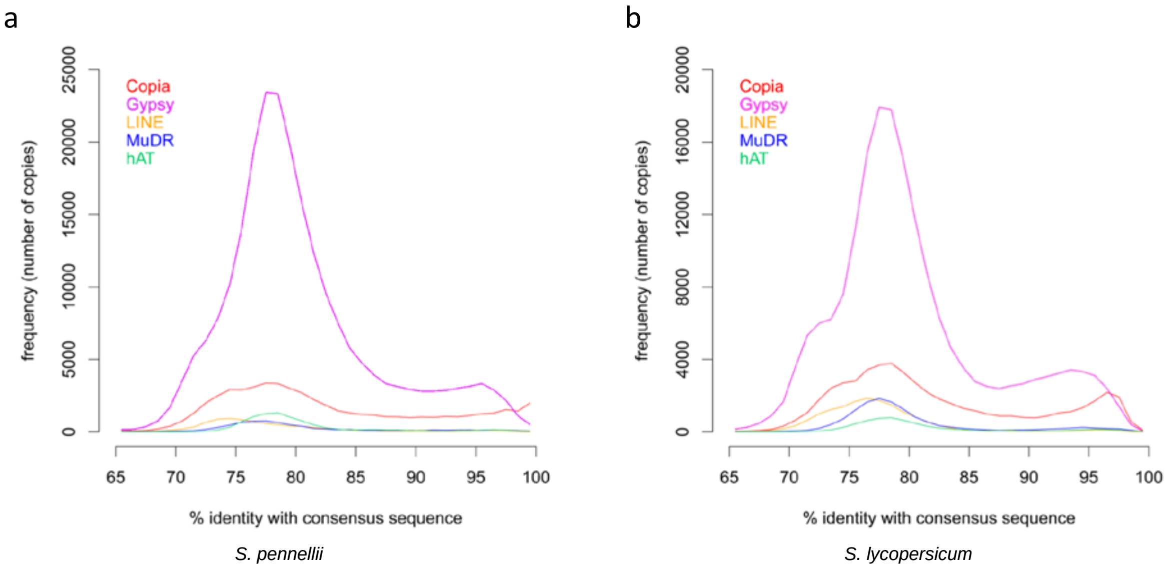
Supplementary Figure 10: Genomic unique and repeat sequence composition of *S. pennellii* and *S. lycopersicum* in mega base pairs (a) and percentage (b). Repeats were only considered when at least 5 copies were found. Repeats were categorized using the REPET dedicated utility followed by semi-manual curation.



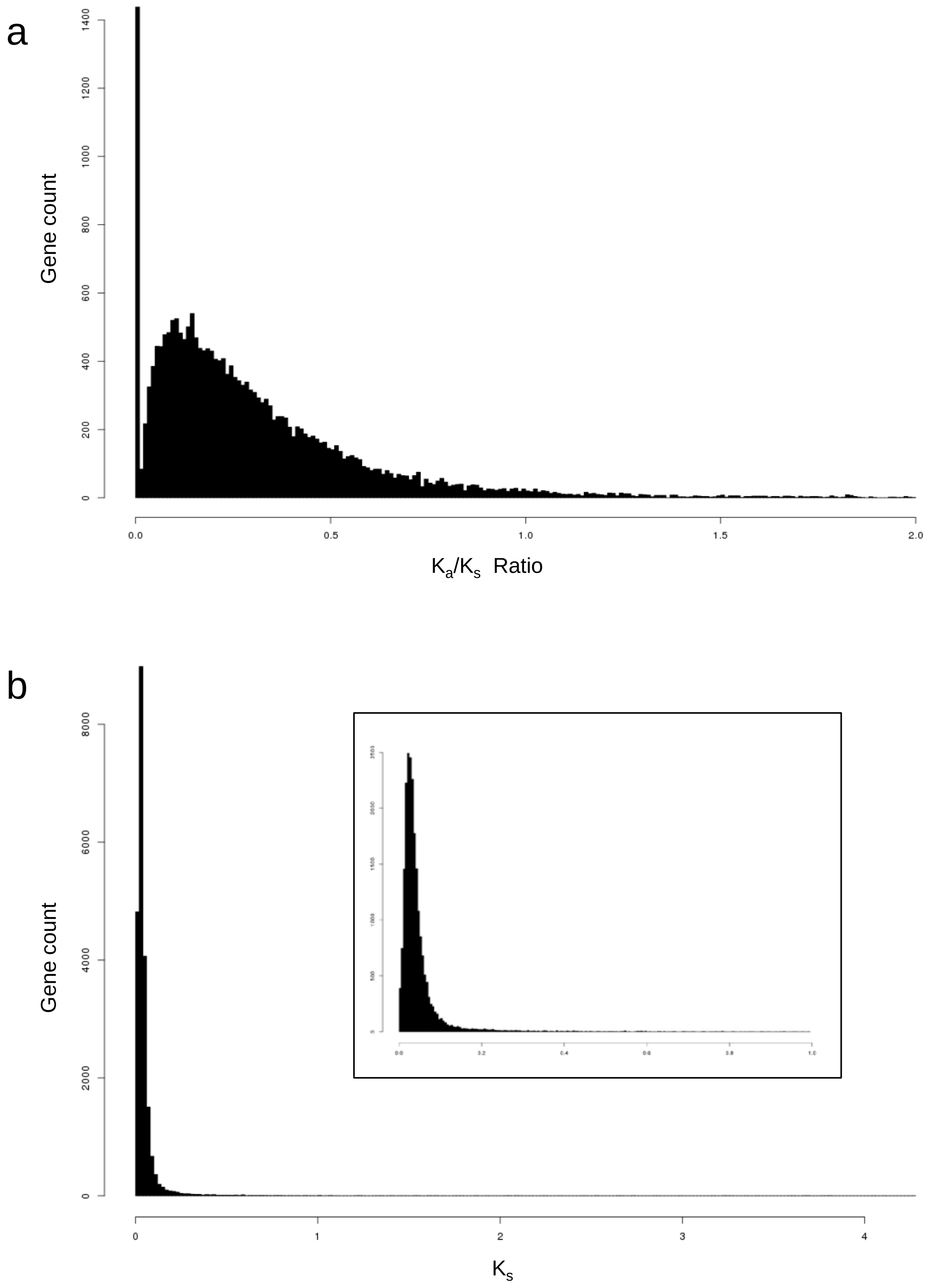
Supplementary Figure 11: Contribution of the different repeat classes to the *S. pennellii* and *S. lycopersicum* genome in mega base pairs (a) and percentage (b and c). Repeats were only considered when at least 5 copies were found. Repeats were categorized using the REPET dedicated utility followed by semi-manual curation.



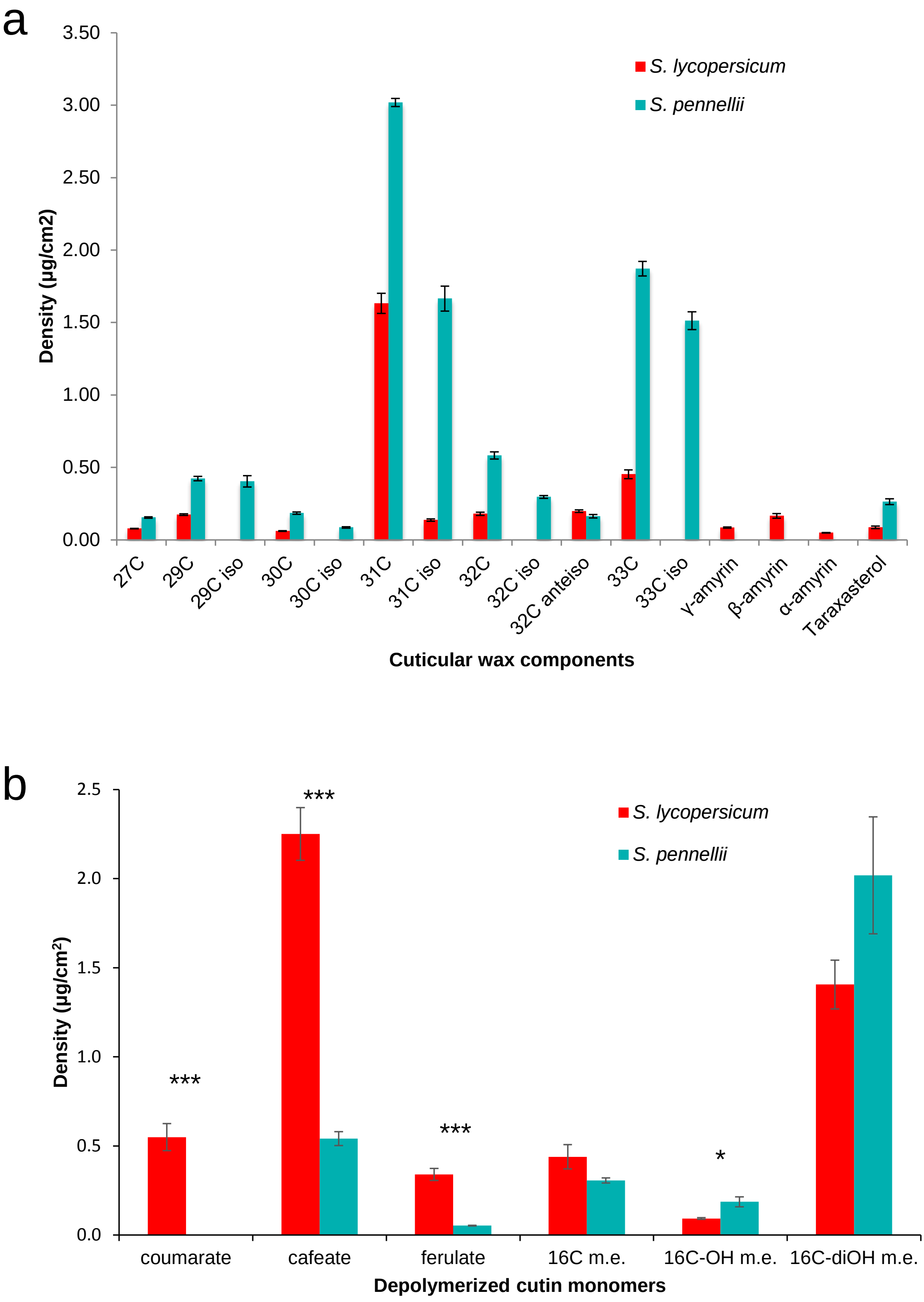
Supplementary Figure 12: Long terminal repeat (LTR) reterotransposon divergence analysis in *S. pennellii*, *S. lycopersicum* and *S. tuberosum* in evolutionary distance (a) and estimated insert age (b). Evolutionary distance was calculated using "Distmat" from the Emboss package and distance per site was then used to estimate insertion age.



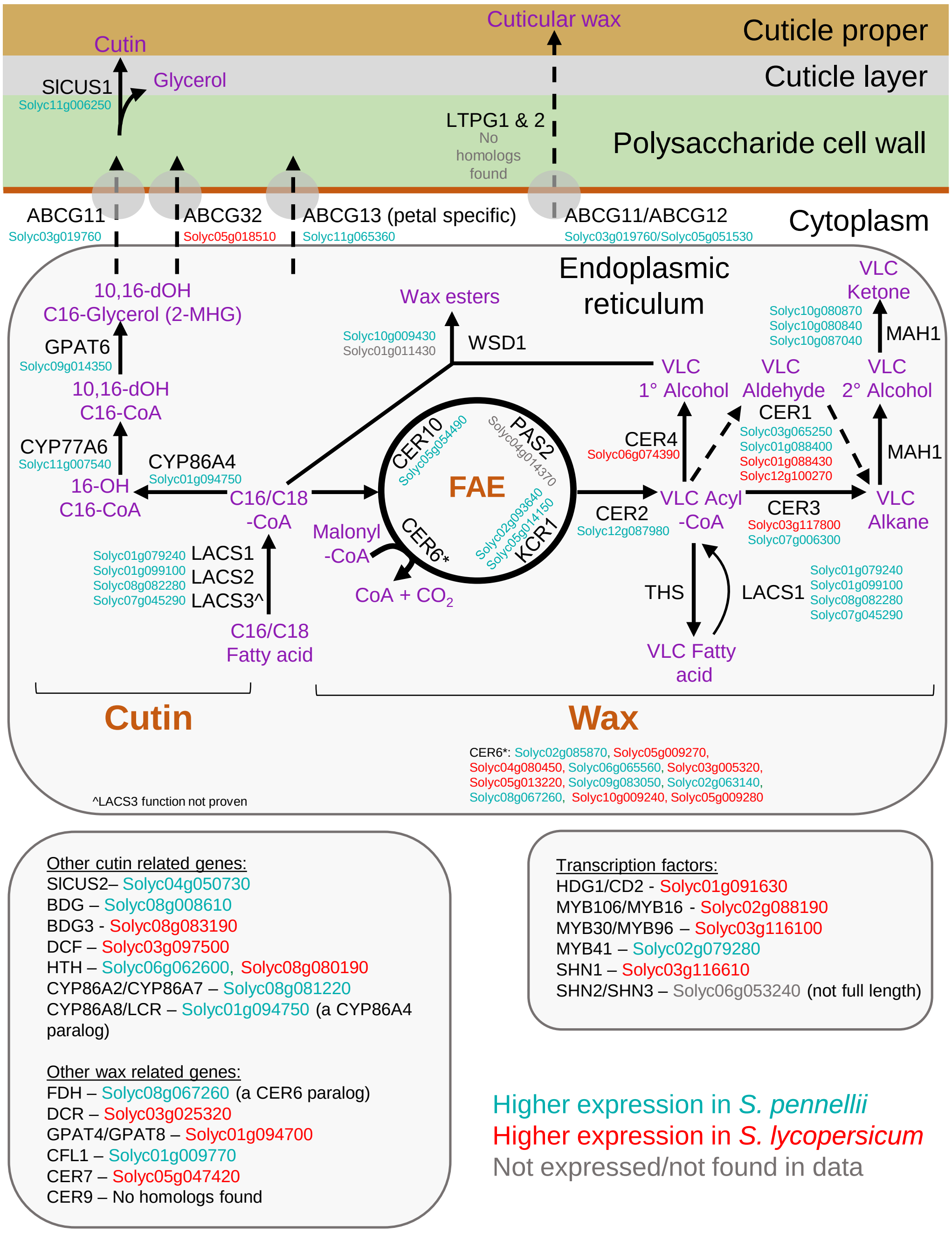
Supplementary Figure 13: Age development of the different transposable elements in *S. pennellii* (a) and *S. lycopersicum* (b). Recent activity is indicated by a greater % identity to the consensus sequence. Young Copia insertions can be seen in *S. pennellii* (a) but is not seen in *S. lycopersicum* (b).



Supplementary Figure 14: Histogram of K_a/K_s values for orthologous genes from *S. pennellii* and *S. lycopersicum*. The histogram of the K_a/K_s values for orthologous genes from *S. pennellii* and *S. lycopersicum* is shown in (b). The histogram of the K_s values are shown in (a) with the insert depicting the K_s values below 1.



Supplementary Figure 15: (a) Composition of extracted cuticular waxes from mature leaves of *S. lycopersicum* cv. M82 and *S. pennellii*. The alkanes are listed by chain length as well as by whether they are straight chain or a branched isomer (iso/ anteiso). The differences between the two species are statistically significant for each compound, with p-values less than 0.001 for all except the 32C anteiso alkane, which has a p-value of 0.047. (b) Composition of depolymerized cutin monomers from mature leaves of *S. lycopersicum* cv. M82 and *S. pennellii*. The aliphatic components were all 16-carbon methyl esters (m.e.) of fatty acids with 0, 1 or 2 hydroxyl groups. The p-values for the differences between the two species, based on t-tests, are represented by *** for p<0.001, ** for p=0.001-0.01, * for p=0.01-0.05. Error bars represent s.e.m.



CER6*: Solyc02g085870, Solyc05g009270, Solyc04g080450, Solyc06g065560, Solyc03g005320, Solyc05g013220, Solyc09g083050, Solyc02g063140, Solyc08g067260, Solyc10g009240, Solyc05g009280

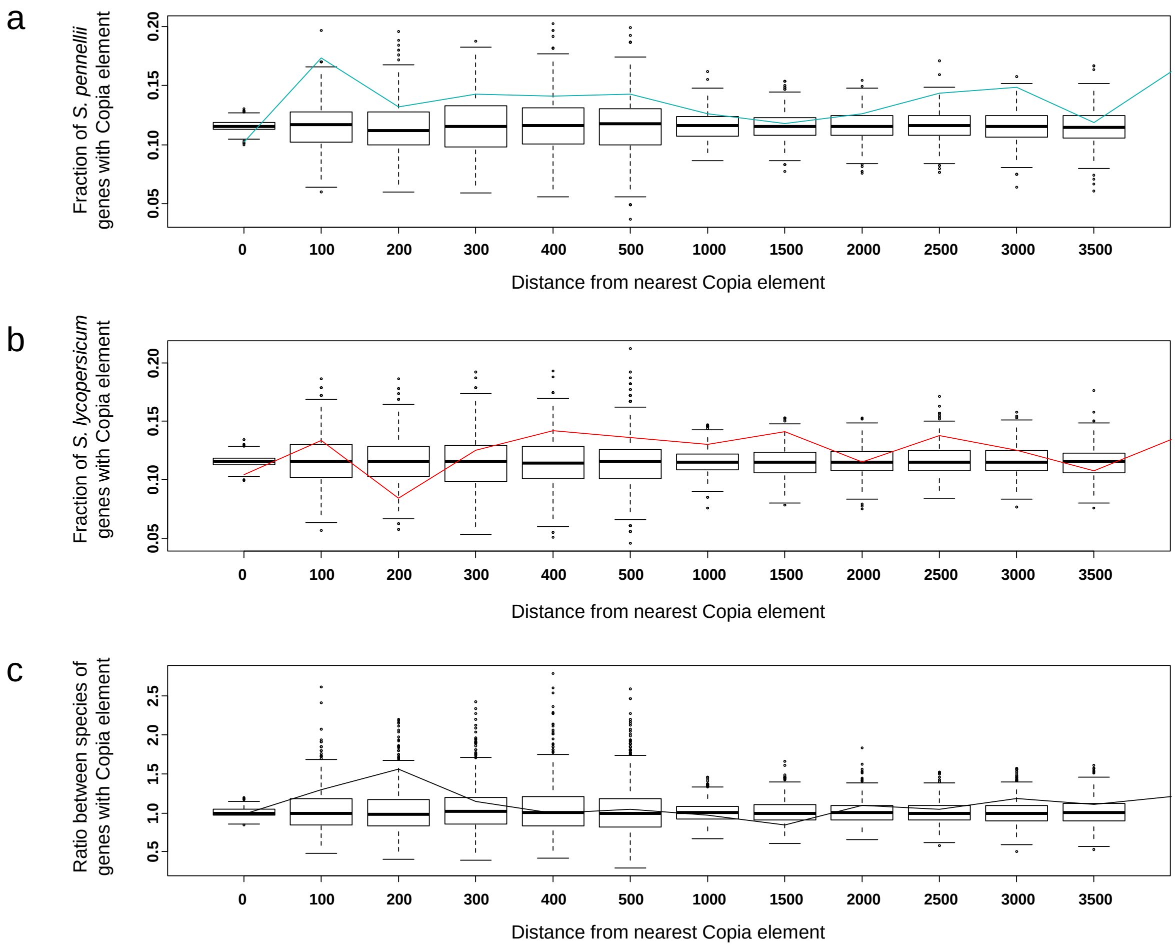
Other cutin related genes:
SlCUS2– Solyc04g050730
BDG – Solyc08g008610
BDG3 - Solyc08g083190
DCF – Solyc03g097500
HTH – Solyc06g062600, Solyc08g080190
CYP86A2/CYP86A7 – Solyc08g081220
CYP86A8/LCR – Solyc01g094750 (a CYP86A4 paralog)

Other wax related genes:
FDH – Solyc08g067260 (a CER6 paralog)
DCR – Solyc03g025320
GPAT4/GPAT8 – Solyc01g094700
CFL1 – Solyc01g009770
CER7 – Solyc05g047420
CER9 – No homologs found

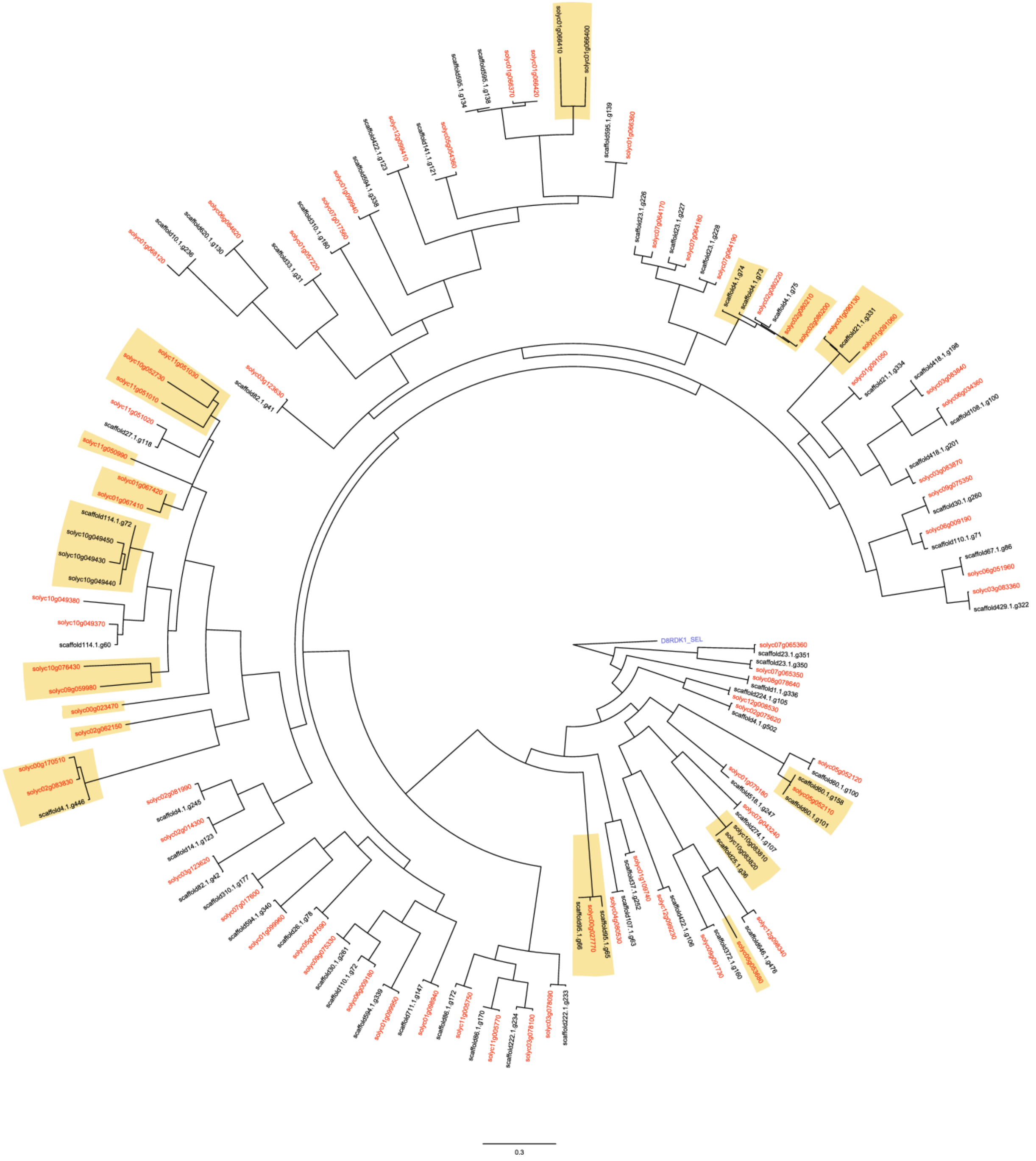
Transcription factors:
HDG1/CD2 - Solyc01g091630
MYB106/MYB16 - Solyc02g088190
MYB30/MYB96 – Solyc03g116100
MYB41 – Solyc02g079280
SHN1 – Solyc03g116610
SHN2/SHN3 – Solyc06g053240 (not full length)

Higher expression in *S. pennellii*
Higher expression in *S. lycopersicum*
Not expressed/not found in data

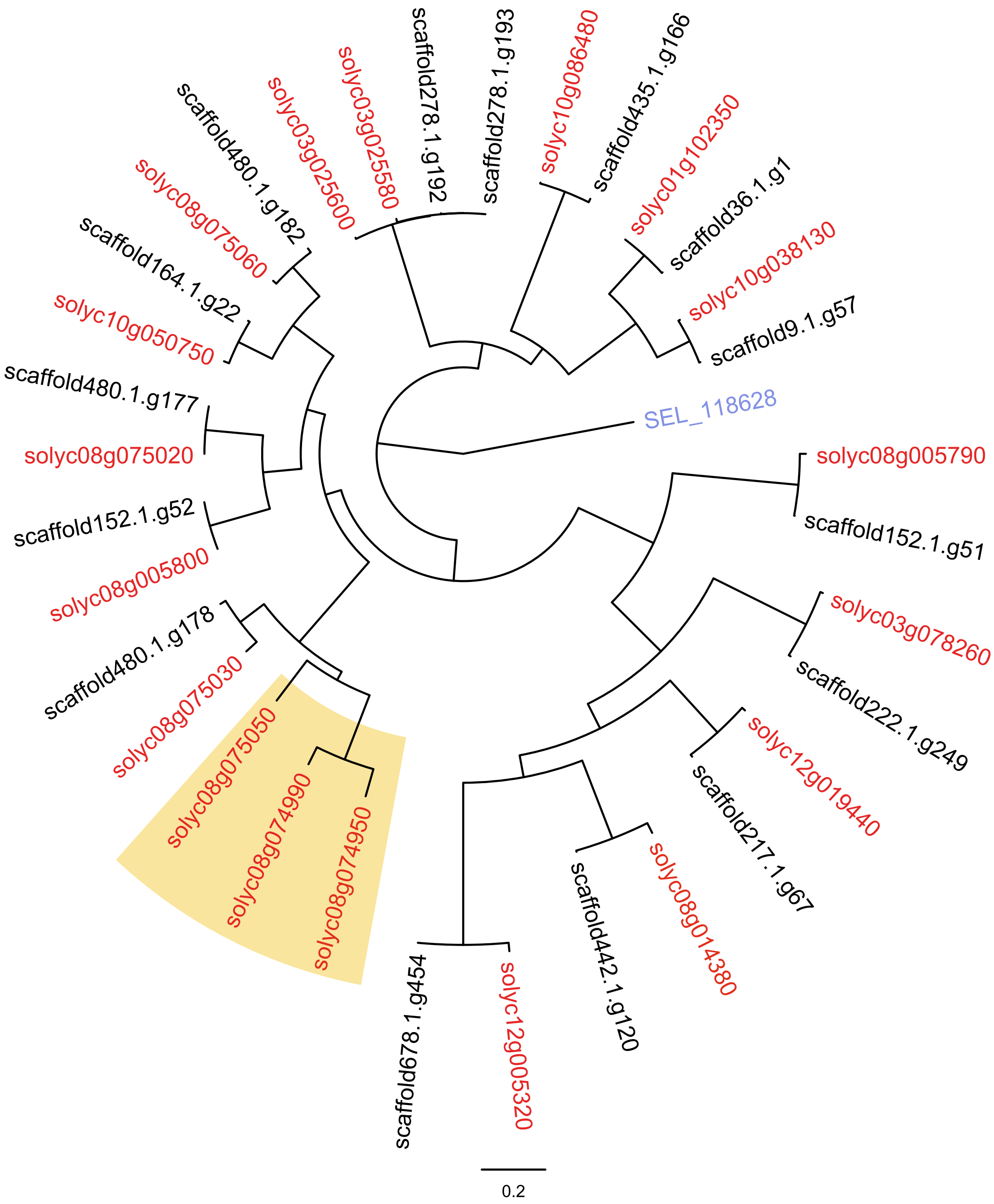
Supplementary Figure 16: Expression of cuticle related genes. This figure is drawn based on the biosynthetic model[147] and includes a color-coded summary of RNA-Seq expression data derived from expanding leaves, taken from Koenig et al.[27]. A. thaliana protein sequences of known cuticle related genes were used to identify the most closely related homologs from the S. lycopersicum and S. pennellii genome sequences. Gene IDs for closely related paralogs are also shown. The relative expression patterns are color coded as indicated.



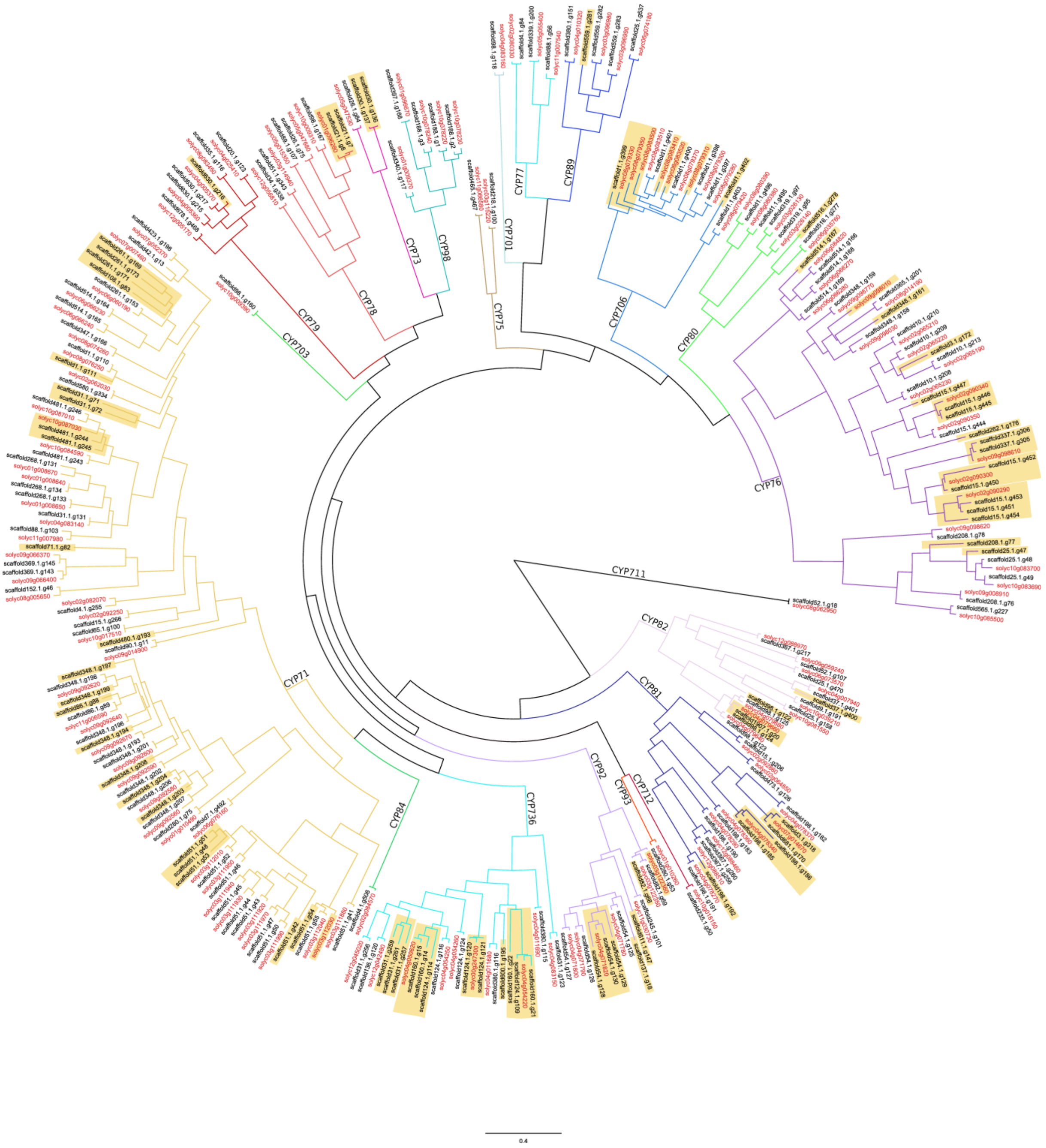
Supplementary Figure 17: Co-location analysis of stress genes. The plotted line depicts the fraction of genes which are stress responsive, at specified distances from Copia elements (x-axis), in *S. pennellii* (a), *S. lycopersicum* (b), and the ratio between *the S. pennellii* and *S. lycopersicum* values (c). The box-plots indicate the background distribution of genes at this distance, generated by 1000 trials drawing an equal number of randomly selected genes from the orthologous gene set. The whiskers represent the most extreme data points within 1.5 times the inter-quartile range from that quartile.



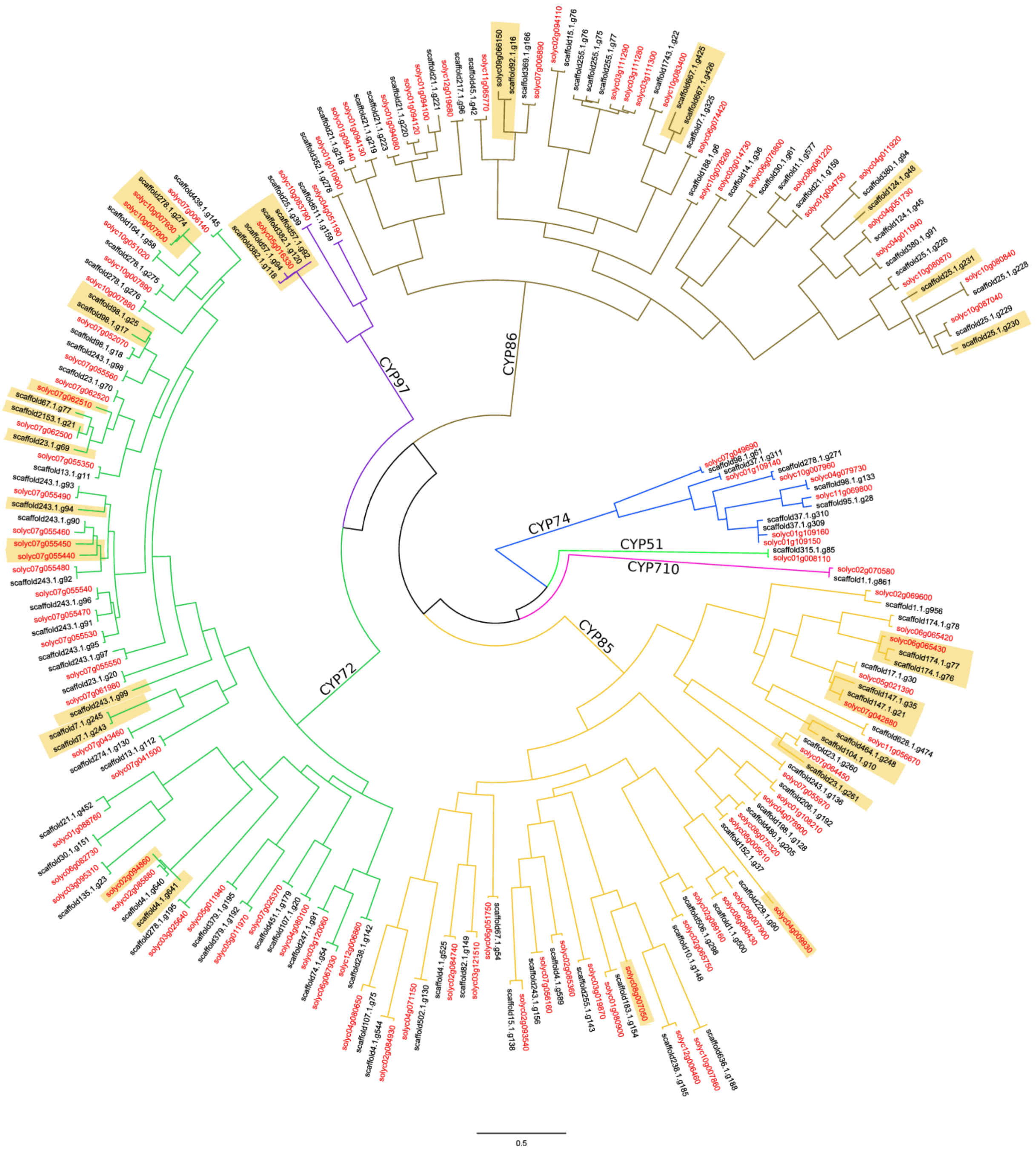
Supplementary Figure 18: Maximum likelihood phylogenetic tree of the pectin methyl esterase (PME) genes from *S. pennellii* and *S. lycopersicum*. A PME from *Selaginella moellendorffii* (labeled in purple) was used to root the tree. *S. lycopersicum* and *S. pennellii* genes are labeled in red and black respectively. Clades where there are a greater number of nodes from one species versus the other and where nodes only exist in one species have been highlighted.



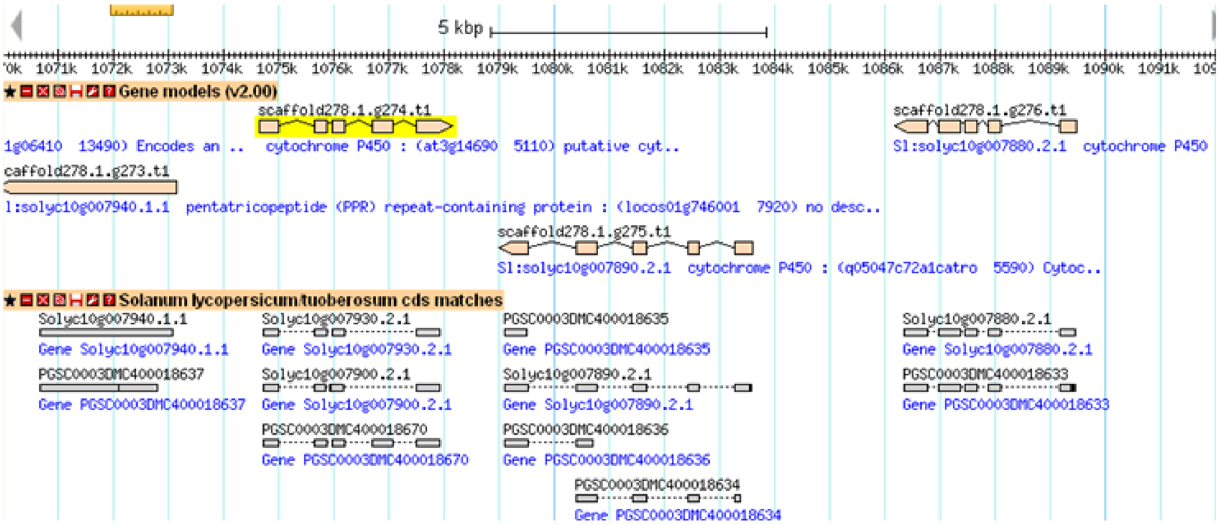
Supplementary Figure 19: Maximum likelihood phylogenetic tree of the pectin acetyl esterase (PAE) genes from *S. pennellii* and *S. lycopersicum*. A PAE from *S. moellendorffii* (labeled in purple) was used to root the tree. *S. lycopersicum* and *S. pennellii* genes are labeled in red and black respectively. Clades where there are a greater number of nodes from one species versus the other and where nodes only exist in one species have been highlighted.



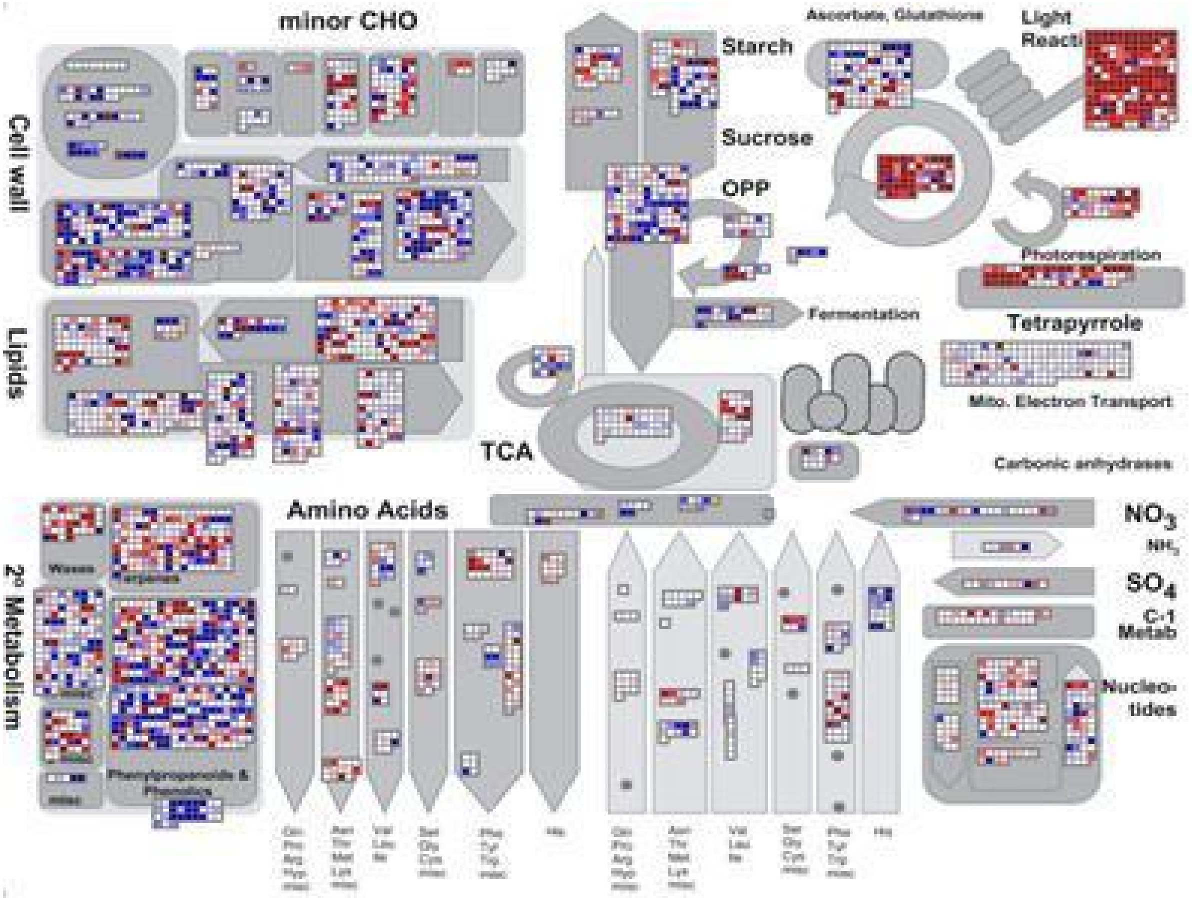
Supplementary Figure 20: Maximum likelihood phylogenetic tree of the cytochrome P450 CYP71 clan from *S. pennellii* and *S. lycopersicum*. CYP711, which is not a member of the CYP71 clan was used to root the tree. The branches of each family within the different CYP71 clan are uniquely colored and labeled to allow them to be easily distinguished. *S. lycopersicum* and *S. pennellii* genes are labeled in red and black respectively. Clades where there are a greater number of nodes from one species versus the other and where nodes only exist in one species have been highlighted.



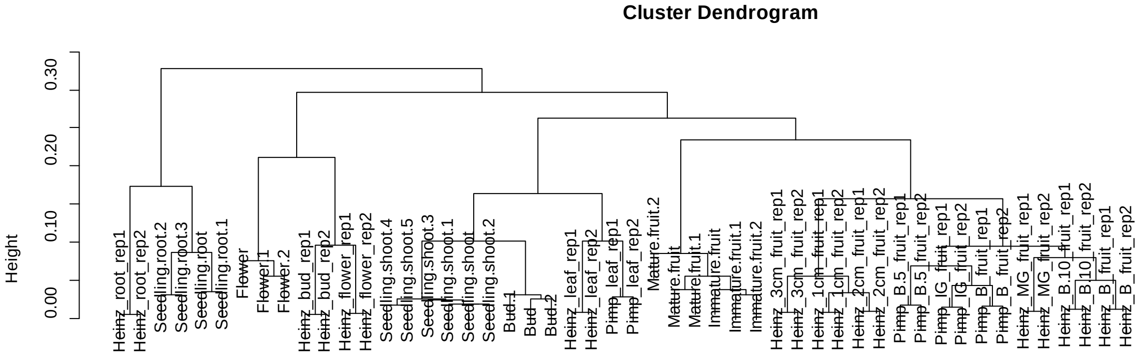
Supplementary Figure 21: Maximum likelihood phylogenetic tree depicting seven clans of the cytochrome P450 family from *S. lycopersicum* and *S. pennellii*. The other two clades found in these species, CYP71 and CYP711, are shown on **Supplementary Figure** 20. The CYP74 clan was used to root the tree. The branches of each clan are uniquely colored and labeled to allow then to be easily distinguished. *S. lycopersicum* and *S. pennellii* genes are labeled in red and black respectively. Clades where there are a greater number of nodes from one species versus the other and where nodes only exist in one species have been highlighted.


5 kbp
'0k 1071k 1072k 1073k 1074k 1075k 1076k 1077k 1078k 1079k 1080k 1081k 1082k 1083k 1084k 1085k 1086k 1087k 1088k 1089k 1090k 1091k 109
Gene models (v2.00)
scaffold278.1.g274.t1
1g06410 13490) Encodes an .. cytochrome P450 : (at3g14690 5110) putative cyt..
scaffold278.1.g276.t1
Sl:solyc10g007880.2.1 cytochrome P450
caffold278.1.g273.t1
l:solyc10g007940.1.1 pentatricopeptide (PPR) repeat-containing protein : (locos01g746001 7920) no desc..
scaffold278.1.g275.t1
Sl:solyc10g007890.2.1 cytochrome P450 : (q05047c72a1catro 5590) Cytoc..
Solanum lycopersicum/tuoberosum cds matches
Solyc10g007940.1.1
Gene Solyc10g007940.1.1
PGSC0003DMC400018637
Gene PGSC0003DMC400018637
Solyc10g007930.2.1
Gene Solyc10g007930.2.1
Solyc10g007900.2.1
Gene Solyc10g007900.2.1
PGSC0003DMC400018670
Gene PGSC0003DMC400018670
PGSC0003DMC400018635
Gene PGSC0003DMC400018635
Solyc10g007890.2.1
Gene Solyc10g007890.2.1
PGSC0003DMC400018636
Gene PGSC0003DMC400018636
PGSC0003DMC400018634
Gene PGSC0003DMC400018634
Solyc10g007880.2.1
Gene Solyc10g007880.2.1
PGSC0003DMC400018633
Gene PGSC0003DMC400018633

Supplementary Figure 22: GBrowse visualization of cytochrome P450 genes from *S. pennellii* together with homologs from *S. lycopersicum* and *S. tuberosum*. The bottom panel depicts 4 genes from both *S. lycopersicum* and *S. tuberosum* for three genes from *S. pennellii* indicating a likely loss of a cytochrome P450 gene from *S. pennellii*.



Supplementary Figure 23: MapMan visualization of root versus shoot data from *S. pennellii* RNA-Seq data. Genes up-regulated in the shoot are indicated in blue and down-regulated genes in red. Stark differences between shoot and root transcripts can be seen in the 'light reaction' category, as expected.



Supplementary Figure 24: Clustering of tissues from *S. pennellii*, *S. lycopersicum* and *S. pimpinellifolium*. RPKM expression estimates were clustered using 1-r_s as a distance measure and hierarchically clustered using average linkage. The y-axis shows the distance between individual items.

Supplementary Tables

Supplementary Table 1: 205 bp insert library.

Read Length	Raw Reads (M)	Raw Bases (Bn)	Trimmed Reads (M)	Trimmed Bases (Bn)	Chloroplast DNA (%)	Mitochondrial DNA (%)
2 x 80	38.3	3.06	18.7	1.00	5.80%	2.86%
2 x 80	30.5	2.44	14.1	0.79	5.49%	2.82%
2 x 42	24.2	1.02	22.4	0.93	5.61%	2.94%
2 x 79	43.3	3.45	37.2	2.35	5.81%	2.17%
2 x 79	40.2	3.19	24.8	1.47	5.20%	2.18%
2 x 79	39.9	3.17	21.9	1.27	5.51%	2.26%
2 x 79	32.6	2.59	17.8	1.04	5.60%	2.34%
2 x 79	28.2	2.25	15.4	0.92	4.89%	2.25%
2 x 80	41.9	3.35	37.6	2.64	6.26%	2.35%
2 x 80	42.4	3.40	37.2	2.56	6.22%	2.45%
2 x 80	41.7	3.34	36.7	2.52	6.15%	2.54%
2 x 80	37.5	3.00	33.4	2.30	6.05%	2.58%
2 x 80	37.3	2.98	34.4	2.41	5.74%	2.59%
2 x 80	41.2	3.30	37.8	2.59	5.74%	2.64%
2 x 80	41.1	3.29	37.7	2.57	5.72%	2.67%
2 x 80	37.1	2.97	34.3	2.32	5.69%	2.69%
2 x 101	41.8	4.22	26.6	1.98	5.73%	1.15%
2 x 101	42.9	4.33	23.3	1.82	5.85%	1.19%
2 x 101	43.9	4.43	24.9	1.91	5.84%	1.19%
2 x 101	47.0	4.75	37.5	2.81	5.76%	1.18%
Total	**773**	**64.53**	**573.7**	**38.2**		

Supplementary Table 2: 275 insert library.

Read Length	Raw Reads (M)	Raw Bases (Bn)	Trimmed Reads (M)	Trimmed Bases (Bn)	Chloroplast DNA (%)	Mitochondrial DNA(%)
2 x 151	88.4	13.3	80.3	10.47	4.98%	0.83%
2 x 151	90.1	13.6	81.7	10.77	4.93%	0.79%
2 x 151	89.1	13.5	80.1	10.11	4.91%	0.77%
2 x 151	91.1	13.8	81.8	10.80	4.93%	0.77%
2 x 151	89.6	13.5	80.6	10.66	4.90%	0.74%
2 x 151	73.5	11.1	65.6	7.57	5.15%	0.94%
Total	**521.8**	**78.8**	**470.1**	**60.38**		

Supplementary Table 3: 515 bp insert library.

Read Length	Raw Reads (M)	Raw Bases (Bn)	Trimmed Reads (M)	Trimmed Bases (Bn)	Chloroplast DNA (%)	Mitochondrial DNA (%)
2 x 80	28.7	2.29	12.1	0.63	4.48%	2.14%
2 x 80	37.1	2.97	31.7	2.15	4.65%	1.94%
2 x 80	36.1	2.88	31.2	2.13	4.58%	1.95%
2 x 80	35.7	2.85	31.1	2.14	4.51%	1.96%
2 x 80	36.2	2.89	31.2	2.13	4.49%	1.94%
2 x 80	34.4	2.75	29.9	1.93	4.34%	1.95%
2 x 80	33.6	2.69	29.3	1.91	4.31%	1.95%
2 x 80	36.6	2.92	32.3	2.15	4.18%	1.94%
2 x 80	34.9	2.79	30.6	2.03	4.15%	1.92%
2 x 101	45.0	4.54	33.8	2.59	4.55%	0.77%
2 x 101	45.4	4.58	34.5	2.67	4.55%	0.77%
2 x 101	44.9	4.54	34.6	2.69	4.50%	0.75%
2 x 101	44.6	4.50	34.2	2.67	4.53%	0.76%
2 x 101	55.8	5.64	54. 0	5.17	5.66%	0.95%
2 x 101	57.1	5.77	55.3	5.25	5.73%	0.96%
2 x 101	52.0	5.25	50.5	4.89	5.75%	0.95%
2 x 101	65.7	6.64	63.3	6.02	5.72%	0.91%
2 x 151	70.6	10.66	67.7	8.99	5.56%	0.84%
2 x 151	64.3	9.71	62.0	8.40	5.59%	0.86%
Total	**858.7**	**86.86**	**695.3**	**66.54**		

Supplementary Table 4: Paired End Libraries.

Library Nominal Size	Lib name (internal)	Raw Read Pairs (M)	Trimmed Read Pairs (M)	Size Estimating Pairs (M)	Estimated Size (Mean / SD)
205bp	A	386.5	212.4	36.1	204/21
275bp	D	261.0	212.2	125.7	275/37
515bp	B	429.3	331.7	107.1	515/70

Supplementary Table 5: Mate Pair Libraries. For the 5kb library marked with (1) the libraries were prepared by the Illumina mate pair preparation kit. The 40kb library marked by (2) was prepared with the Lucigen Fosmid kit; all other libraries were prepared using the hybrid approach explained in the Supplementary Note.

Library Nominal Size	Library name	Raw Read Pairs (M)	Trimmed Read Pairs (M)	Size Estimating Pairs (M)	Estimated Size (Mean / SD)	Mate Pair grouping name
3kbp	EA	34.2	27.3	4.60	1964 / 767	MP04
5kbp(1)	CA	26.8	22.7	1.20	4197 / 524	MP05
	CB	27.7	23.2	1.22	4197 / 524	
	CC	37.7	36.3	1.44	4179 / 530	
	CD	36.0	34.8	1.42	4179 / 529	
5kbp	CE	27.4	17.5	1.74	4627 / 541	
8kbp	FA	36.3	23.7	5.64	6086 / 1007	MP06
	FB	35.7	23.3	5.54	6086 / 1007	
10kbp	FC	37.1	23.6	1.48	9053 / 1177	MP07
10kbp	FD	39.9	18.8	1.12	8702 / 1427	
20kbp	GB	41.9	27.8	0.35	17796 / 4167	MP08
20kbp	GC	36.1	16.7	0.31	17912 / 4031	
30kbp	GA	42.8	21.4	0.12	24895 / 6794	MP09
30kbp	HA	37.8	12.2	0.18	26272 / 9012	
30kbp	HB	48.4	24.0	0.63	24549 / 7508	
40kbp	HC	45.2	13.9	0.25	28741 / 12707	MP10
40kbp(2)	HD	105.4	100.3	0.32	37224 / 4659	

Supplementary Table 6: Detailed genome assembly statistics. The statistics for the genome assembly are given after each stage of the SOAP de novo assembly. The statistics were calculated using the 'statsContigAll' tool from http://code.google.com/p/curtain/, against the contigs and/or scaffolds produced at each stage. After PE refers to after using paired end libraries only. After 3/5 8/10 20, 30 and 40kb refers to the results after scaffolding with the corresponding mate-pair library. Min and max are minimum and maximum sequence sizes and total refers to sum of the lengths of all sequences. Split after anchoring refers to the data set that was anchored using RAD-Seq and genetic markers.

Stage	n10	n20	n30	n40	n50	n60	n70	n80	n90	min	max	total	# Contigs/ Scaffold	n > n50	N
Contigs	17,180	10,976	7,148	4,370	2,176	779	246	108	68	64	80,781	1,117,562,721	4,315,954	81,824	0
After PE	114,924	78,931	59,405	45,150	33,782	24,683	16,565	9,015	2,281	100	391,544	907,124,276	464,970	6,982	10,971,966
After 3/5Kb	291,764	200,811	152,980	116,613	88,502	64,533	43,654	23,524	3,572	100	925,062	919,696,645	435,788	2,750	24,672,194
After 8/10Kb	726,647	506,420	390,275	300,355	227,029	170,160	119,359	71,492	19,034	100	1,973,027	982,759,790	413,693	1,162	87,631,776
After 20Kb	2,215,469	1,476,079	1,168,714	944,145	738,109	566,421	394,021	238,393	45,648	100	4,240,888	1,006,163,183	409,235	386	110,997,002
After 30Kb	4,951,557	3,283,930	2,463,944	1,937,804	1,541,925	1,184,815	918,483	562,572	107,406	100	9,085,884	1,016,203,713	407,707	182	120,600,869
After 40Kb	5,003,978	3,370,668	2,482,965	2,019,941	1,603,317	1,212,905	944,646	600,486	98,078	100	10,126,651	1,021,472,455	407,506	177	125,806,430
Scaff-unfilled	5,003,978	3,370,668	2,482,965	2,019,941	1,603,317	1,212,905	944,646	600,486	98,078	100	10,126,651	1,021,472,455	407,506	177	125,806,430
Scaff	4,970,060	3,341,192	2,463,860	1,997,983	1,590,935	1,196,704	935,772	594,885	95,443	100	10,011,355	1,012,612,203	407,506	177	67,624,937
Scaff_L2000	5,114,738	3,416,689	2,670,319	2,113,751	1,741,129	1,353,889	1,059,177	763,066	437,042	2,000	10,011,355	942,624,776	4,591	156	67,190,024
Scaff_L2000_ decon	**5,114,738**	**3,416,689**	**2,670,319**	**2,113,751**	**1,741,129**	**1,353,889**	**1,059,177**	**763,066**	**437,042**	**2,000**	**10,011,355**	**942,595,034**	**4,579**	**156**	**67,190,021**
Split after anchoring	3,823,417	2,758,321	2,199,110	1,833,122	1,452,825	1,145,865	928,427	658,741	370,741	2,000	8,769,512	942,398,177	4,726	192	66,993,164

Supplementary Table 7: Comparison of genomic assembly statistics between *S. pennellii*, *S. lycopersicum* and *S. tuberosum.*

Type	Species	n10	n20	n30	n40	n50	n60	n70	n80	n90	min	max	total	# Contigs/ Scaffold	n > n50	N
Scaffolds	*S. lycopersicum*	32987597	28223487	22422656	18607109	16467796	11730995	8223970	6294186	3041128	2000	42121211	781345411	3223	17	43,709,063
	S. pennellii	5114738	3416689	2670319	2113751	1741129	1353889	1059177	763066	437042	2000	10011355	942595034	4579	156	67,190,021
	S. tuberosum	3297431	2506659	2030316	1655652	1354002	1092554	827814	538315	291899	2000	7100477	714593138	2310	162	44,527,276
Contigs	*S. lycopersicum*	374812	234946	168244	120160	86909	62879	44201	28258	14131	100	2487452	737634768	26905	2011	0
	S. pennellii	132269	94561	72667	57821	45715	35384	26348	18054	10035	100	381625	875372209	47968	5410	0
	S. tuberosum	82181	60476	48064	38702	31429	24812	18866	13084	6858	100	253599	682647040	113093	6446	0

Supplementary Table 9: Sequence assessment using BACs. The table lists BAC ID, its Genbank ID, sequencing technology used, the length of the individual BACs, SNPs and GAPs found after aligning it to the *S. pennellii* genome, the coverage and finally SNPs and Gaps after resequencing most of the BACS using high coverage long read Illumina technology. The BAC marked with # had a different sequence than the original and likely represents new BAC sequence.

Scaffold	Sequencing	BAC	Genbank	BAC length	SNPs	Gaps	% BAC cover	SNPs post	Gaps post
scaffold403.1	Sanger	C02SpCP013J021.P3B05.Contig12	FJ812349	27,267	33	15	95.49	10	8
scaffold403.1	Sanger	C02SpCP013J021.P3B05.Contig13	FJ812349	66,803	139	49	99.68	64	18
scaffold268.1	Sanger	C01SpBP028F006.P5D08	FJ809742	100,386	137	60	99.99	26	4
scaffold70.1	Sanger	C02SpCP020G005.P6A06	FJ809746	50,564	0	5	100	split	split
scaffold107.1	Roche454	C04SpBP093E005.P4C04	FJ809740	83,193	0	20	97.74	3	6
scaffold34.1	Roche454	C11SpBP029K005.P4E08	FJ809741	94,556	1	25	98.34	-	-
scaffold23.1	Roche454	C07SpCP018I014.P4G10	FJ809743	97,026	40	35	99.82	38	11
scaffold23.1	Roche454	C07SpCP034K014.P4H06	FJ809744	148,978	20	51	99.99	-	-
scaffold23.1	Roche454	C07SpCP066B007.P5C04	FJ809745	146,349	41	25	97.47	18	5
scaffold23.1	Roche454	C07SpCP101P015.P5F03	FJ809747	111,488	5	20	99.96	6	7
scaffold70.1	*Sanger*	*C02SpCP020G005.P6A06*	-	196,440				*71*	*20*
Scaffold23.1	*Sanger*	*C02SpCP020G005.P6A06*	-	146,941				*18*	*4*
scaffold34.1	*Roche454*	*C11SpBP029K005.P4E08#*	-	*103,329*				*1*	*9*

Supplementary Table 10: Completeness assessment using publicly available EST data.

Dataset	Number	Mapped	Identity threshold	Completeness Threshold	% mapped
***S. lycopersicum* EST**	307,350	239,445	95	95	77.91
		257,259	95	90	83.70
		271,414	95	80	88.31
		252,538	90	95	82.17
		271,677	90	90	88.39
		286,556	90	80	93.23
***S. pennellii* EST**	7,812	6,940	95	95	88.84
		7,286	95	90	93.77
		7,503	95	80	96.04
Tomato Unigenes	42,257	29,225	90	95	69.16
		32,277	90	90	76.38
		35,154	90	80	83.19
		30,182	80	95	71.42
		33,418	80	90	79.08
		36,508	80	80	86.40

Supplementary Table 11: Codon usage in the *S. pennellii* and *S. lycopersicum* species. Data is given in total codons used as well as in %.

Amino Acid	Codon	*S.lycopersicum*	*S.pennellii*	*(%) S.lycopersicum*	*(%) S.pennellii*
ala	GCA	233257	250327	34.350	33.857
	GCC	93840	106805	13.819	14.446
	GCG	52955	59472	7.798	8.044
	GCT	299017	322758	44.033	43.654
arg	AGA	190136	199624	35.241	33.164
	AGG	132512	146636	24.561	24.361
	CGA	63684	74669	11.804	12.405
	CGC	36369	43102	6.741	7.161
	CGG	39700	45521	7.358	7.562
	CGT	77130	92384	14.296	15.348
asn	AAC	168946	175342	33.490	33.844
	AAT	335524	342740	66.510	66.156
asp	GAC	151158	167800	26.458	27.055
	GAT	420153	452409	73.542	72.945
cys	TGC	72821	82406	36.959	38.202
	TGT	124213	133306	63.041	61.798
gln	CAA	237624	249226	60.392	59.973
	CAG	155845	166338	39.608	40.027
glu	GAA	399780	421925	58.063	57.261
	GAG	288749	314920	41.937	42.739
gly	GGA	244283	258730	36.016	35.489
	GGC	94335	106100	13.908	14.553
	GGG	113160	123097	16.684	16.885
	GGT	226479	241124	33.391	33.074
his	CAC	78885	85870	31.036	31.794
	CAT	175286	184216	68.964	68.206
ile	ATA	165267	165529	27.238	26.883
	ATC	139048	144046	22.917	23.394
	ATT	302442	306159	49.846	49.723
leu	CTA	111236	114832	11.082	11.092
	CTC	110689	114564	11.027	11.066
	CTG	111432	118739	11.101	11.469
	CTT	255386	263747	25.443	25.476
	TTA	164203	162212	16.359	15.669
	TTG	250831	261166	24.989	25.227
lys	AAA	356284	365969	52.434	50.981
	AAG	323210	351883	47.566	49.019
met	ATG	262768	283046	100.000	100.000
phe	TTC	174581	180041	38.177	38.782
	TTT	282717	284198	61.823	61.218

pro	CCA	197771	218952	38.671	38.495
	CCC	60838	69151	11.896	12.158
	CCG	50146	61356	9.805	10.787
	CCT	202662	219324	39.628	38.560
ser	AGC	103392	108645	10.926	10.806
	AGT	180358	189828	19.059	18.880
	TCA	240993	251697	25.466	25.033
	TCC	105214	116676	11.118	11.604
	TCG	64657	73727	6.832	7.333
	TCT	251709	264882	26.599	26.344
stop	TAA	10373	9752	36.664	35.216
	TAG	6829	6830	24.138	24.664
	TGA	11090	11110	39.198	40.120
thr	ACA	193338	199617	36.308	35.159
	ACC	85040	93775	15.970	16.517
	ACG	49693	58182	9.332	10.248
	ACT	204427	216173	38.390	38.076
trp	TGG	135256	140435	100.000	100.000
tyr	TAC	113911	118902	36.109	36.718
	TAT	201556	204920	63.891	63.282
val	GTA	130726	135545	18.605	17.768
	GTC	101213	114571	14.405	15.019
	GTG	167815	185636	23.884	24.335
	GTT	302867	327098	43.105	42.878

Supplementary Table 12: Distribution of MapMan annotations, assigned by Mercator, for *S. pennellii*, *S. lycopersicum*, and *S. tuberosum*. An in-depth comparison is included in Supplementary dataset 2.

BINcounts 1st level				
S.pennellii	***S.lycopersicum***	***S.tuberosum***	***BinName***	***Bin***
252	386	241	PS	1
110	116	104	major CHO metabolism	2
141	135	127	minor CHO metabolism	3
85	79	77	glycolysis	4
22	26	22	fermentation	5
15	13	16	gluconeogenesis / glyoxylate cycle	6
33	33	32	OPP	7
72	77	71	TCA / org transformation	8
133	189	119	mitochondrial electron transport / ATP synthesis	9
519	553	521	cell wall	10
523	561	570	lipid metabolism	11
36	36	35	N-metabolism	12
308	313	282	amino acid metabolism	13
11	11	10	S-assimilation	14
72	71	69	metal handling	15
669	695	894	secondary metabolism	16
811	929	1002	hormone metabolism	17
84	90	84	Co-factor and vitamine metabolism	18
51	50	43	tetrapyrrole synthesis	19
1057	1113	1477	stress	20
245	260	263	redox	21
21	21	18	polyamine metabolism	22
171	189	162	nucleotide metabolism	23
50	51	52	Biodegradation of Xenobiotics	24
25	26	29	C1-metabolism	25
1696	1728	1954	misc	26
3241	3352	2956	RNA	27
3241	668	536	DNA	28
3538	3947	3592	protein	29
1469	1576	1613	signalling	30
849	879	784	cell	31
1	1	1	micro RNA, natural antisense etc	32
820	865	758	development	33
1161	1232	1141	transport	34
24890	15851	21150	not assigned	35

Supplementary Table 14: Categories over- and under represented for KaKs<=0.01 and p<=0.01; Overrepresented Categories were determined using the online tool http://mapman.mpimp-golm.mpg.de/general/ora/ora.shtml. Overrepresented and Underrepresented Processes ordered separately by p-value for KaKs value <=0.01 and p<=0.01.

BINCode	BINName	Found	Background	p-value	Ratio
29.2	protein.synthesis	113	480	4.85E-46	4.802101
29.2.1	protein.synthesis.ribosomal protein	86	274	8.18E-46	6.402388
29.2.1.2	protein.synthesis.ribosomal protein.eukaryotic	68	164	1.69E-45	8.457834
29	protein	271	2987	1.78E-25	1.850666
29.2.1.2.2	protein.synthesis.ribosomal protein.eukaryotic.60S subunit	41	112	2.34E-25	7.467237
29.2.1.2.1	protein.synthesis.ribosomal protein.eukaryotic.40S subunit	27	52	4.67E-22	10.59143
28.1.3.2	DNA.synthesis/chromatin structure.histone.core	16	30	7.58E-14	10.8791
28.1.3	DNA.synthesis/chromatin structure.histone	16	31	1.49E-13	10.52816
30.5	signalling.G-proteins	37	221	5.27E-11	3.415101
28.1.3.2.1	DNA.synthesis/chromatin structure.histone.core.H2A	11	22	1.60E-09	10.19915
29.5.11.3	protein.degradation.ubiquitin.E2	15	47	3.53E-09	6.510097
29.5.11.20	protein.degradation.ubiquitin.proteasom	16	54	3.61E-09	6.043942
31.1.1.1	cell.organisation.cytoskeleton.actin	11	28	3.71E-08	8.01362
31.1.1.1.1	cell.organisation.cytoskeleton.actin.Actin	7	10	7.03E-08	14.27881
29.2.1.1	protein.synthesis.ribosomal protein.prokaryotic	20	101	8.18E-08	4.039268
29.2.1.1.1	protein.synthesis.ribosomal protein.prokaryotic.chloroplast	12	53	7.48E-06	4.618484
31.1.1	cell.organisation.cytoskeleton	16	96	1.63E-05	3.399718
27.3.71	RNA.regulation of transcription.SNF7	6	13	1.74E-05	9.414602
28.1.3.2.3	DNA.synthesis/chromatin structure.histone.core.H3	4	5	2.76E-05	16.31864
27.2	RNA.transcription	14	80	3.10E-05	3.569703
29.3	protein.targeting	31	294	5.89E-05	2.150842
10.1	cell wall.precursor synthesis	10	49	0.00011	4.162919
29.2.1.2.2.9	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L9	3	3	0.000118	20.39831
29.2.1.2.1.27	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S27	3	3	0.000118	20.39831
29.2.1.1.1.2	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit	9	41	0.000134	4.477677
29.3.4	protein.targeting.secretory pathway	20	165	0.000174	2.472522
1.3.6	PS.calvin cyle.aldolase	4	7	0.000178	11.65617
29.3.4.99	protein.targeting.secretory pathway.unspecified	13	83	0.000191	3.194915
31	cell	60	754	0.000209	1.623207
8	TCA / org. transformation	11	63	0.00022	3.561609
29.2.1.1.3.2	protein.synthesis.ribosomal protein.prokaryotic.unknown organellar.50S subunit	7	27	0.00025	5.288449
29.6	protein.folding	12	75	0.000272	3.263729
9.1	mitochondrial electron transport / ATP synthesis.NADH-DH	8	36	0.000288	4.532957
8.2	TCA / org. transformation.other organic acid transformaitons	6	20	0.000293	6.119492
34.1	transport.p- and v-ATPases	9	46	0.000338	3.990973
34.1.1	transport.p- and v-ATPases.H+-transporting two-sector ATPase	5	14	0.000387	7.285109
29.2.1.2.2.518	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L18A	3	4	0.000453	15.29873
27.1	RNA.processing	29	301	0.000615	1.965285
28.1	DNA.synthesis/chromatin structure	26	257	0.000631	2.063642
29.2.3	protein.synthesis.initiation	12	83	0.000703	2.949153
28.1.3.2.2	DNA.synthesis/chromatin structure.histone.core.H2B	4	10	0.000951	8.159322

1	PS	20	188	0.00103	2.170032
27.3.75	RNA.regulation of transcription.GRP	3	5	0.001091	12.23898
29.2.1.2.1.8	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S8	3	5	0.001091	12.23898
9	mitochondrial electron transport / ATP synthesis	14	113	0.001263	2.527224
27.1.1	RNA.processing.splicing	9	58	0.001931	3.165254
29.2.1.2.1.19	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S19	2	2	0.002401	20.39831
29.2.1.2.2.36	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L36	2	2	0.002401	20.39831
29.2.1.2.1.6	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S6	2	2	0.002401	20.39831
29.2.1.2.2.22	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L22	2	2	0.002401	20.39831
29.2.1.2.2.21	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L21	2	2	0.002401	20.39831
29.2.1.2.1.12	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S12	2	2	0.002401	20.39831
29.2.1.2.1.14	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S14	2	2	0.002401	20.39831
29.2.1.1.3	protein.synthesis.ribosomal protein.prokaryotic.unknown organellar	7	39	0.002589	3.661234
8.2.10	TCA / org. transformation.other organic acid transformaitons.malic	3	7	0.003544	8.742131
29.5.11	protein.degradation.ubiquitin	60	842	0.004197	1.453561
29.3.4.1	protein.targeting.secretory pathway.ER	4	15	0.005077	5.439548
34.19	transport.Major Intrinsic Proteins	7	44	0.005219	3.245185
8.1	TCA / org. transformation.TCA	6	35	0.006554	3.496852
1.2.3	PS.photorespiration.aminotransferases peroxisomal	2	3	0.006969	13.59887
7.2.4	OPP.non-reductive PP.ribose 5-phosphate isomerase	2	3	0.006969	13.59887
8.1.6	TCA / org. transformation.TCA.succinyl-CoA ligase	2	3	0.006969	13.59887
10.1.3	cell wall.precursor synthesis.AXS	2	3	0.006969	13.59887
29.2.1.2.2.5	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L5	2	3	0.006969	13.59887
29.2.1.2.2.26	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L26	2	3	0.006969	13.59887
29.2.1.2.1.16	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S16	2	3	0.006969	13.59887
29.2.1.1.3.2.510	protein.synthesis.ribosomal protein.prokaryotic.unknown organellar.50S subunit.L10A	2	3	0.006969	13.59887
29.2.1.2.2.30	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L30	2	3	0.006969	13.59887
29.2.1.2.2.510	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L10A	2	3	0.006969	13.59887
35	not assigned	227	7358	3.76E-20	0.629304
35.2	not assigned.unknown	174	5621	2.75E-14	0.631437
26	misc	23	1303	1.71E-09	0.360062
30.2	signalling.receptor kinases	4	474	1.38E-06	0.172138
35.1	not assigned.no ontology	53	1737	9.85E-05	0.622401
20.1	stress.biotic	5	374	0.000381	0.272705
26.1	misc.cytochrome P450	1	172	0.003701	0.118595

Supplementary Table 18: Expression data for genes involved in wax and cutin synthesis.

At1g68530	Solyc02g085870	CER6	1.573389
	Solyc05g009270	CER6 paralog	-0.56941
	Solyc04g080450	CER6 paralog	-0.77084
	Solyc06g065560	CER6 paralog	0.376149
	Solyc03g005320	CER6 paralog	-3.0641
	Solyc05g013220	CER6 paralog	-0.13871
	Solyc09g083050	CER6 paralog	0.93926
	Solyc02g063140	CER6 paralog	2.625329
	Solyc08g067260	CER6 paralog	0.04407
	Solyc10g009240	CER6 paralog	-1.2921
	Solyc05g009280	CER6 paralog	-0.03468
At1g67730	Solyc02g093640	KCR1	0.365012
	Solyc05g014150	KCR1 paralog	-0.4269
At5g10480	Solyc04g014370	PAS2	ND
At3g55360	Solyc05g054490	CER10	0.97172
At4g33790	Solyc06g074390	CER4	-3.42908
At5g37300	Solyc10g009430	WSD1	0.30011
	Solyc01g011430	WSD1 paralog	0
At5g57800	Solyc03g117800	CER3	-3.28207
	Solyc07g006300	CER3 paralog	0.744964
At1g02205	Solyc03g065250	CER1	0.929766
	Solyc01g088400	CER1 paralog	0.250621
	Solyc01g088430	CER1 paralog	-0.73598
	Solyc12g100270	CER1 paralog	-1.63196
At1g57750	Solyc10g080870	MAH1/CYP96A15	0.162028
	Solyc10g080840	MAH1/CYP96A15 paralog	-0.01971
	Solyc10g087040	MAH1/CYP96A15 paralog	6.3736
At2g47240	Solyc01g079240	CER8/LACS1	0.26063
	Solyc01g099100	CER8/LACS1 paralog	0.89263
	Solyc08g082280	CER8/LACS1 paralog	0.229408
	Solyc07g045290	CER8/LACS1 paralog	0.292331
At1g51500	Solyc03g019760	CER5/ABCG12	0.323063
	Solyc05g051530	CER5/ABCG12 paralog	1.528516
At3g60500	Solyc05g047420	CER7	-0.99639
At1g15360	Solyc03g116610	SHN1/WIN1	-8.5485
At3g28910	Solyc03g116100	MYB30	-0.25104
At4g28110	Solyc02g079280	MYB41	3.507606
At4g24510	Solyc12g087980	CER2	3.121495
At1g63710	Solyc08g081220	CYP86A7	0.225761
At2g45970	Solyc01g094750	CYP86A8/LCR	0.191935
At3g10570	Solyc11g007540	CYP77A6	0.51304
At1g01610	Solyc01g094700	GPAT4	-0.08013
At2g38110	Solyc09g014350	GPAT6	0.797
At3g48720	Solyc03g097500	DCF	-3.83595
At3g04290	Solyc04g050730	LTL1	1.335836
Pp1s34 98V6.1	Solyc11g006250	PpCUS1	0.072828
At1g64670	Solyc08g008610	BDG	0.404505
At5g23940	Solyc03g025320	DCR/PEL3	-0.37336
At1g72970	Solyc06g062600	HOTHEAD	1.799705
	Solyc08g080190	Hothead paralog	-0.34475
At1g51460	Solyc11g065360	ABCG13	0.587499
At2g26910	Solyc05g018510	ABCG32	-0.62994
At2g33510	Solyc01g009770	CFL1	0.214187
At3g61150	Solyc01g091630	CD2/HDG1	-0.36241
At4g24140	Solyc08g083190	BDG3	-0.30148
	Solyc10g081450	BDG3 paralog	2.174514
At3g01140	Solyc02g088190	MYB106	-0.09389

Supplementary Table 19: Primers used for cutin validation.

Gene	Primer Sequence
Solyc05g054480_actin_F	AGATCCTCACCGAGCGTGGTTA
Solyc05g054480_actin_R	GAGCTGGTCTTTGAAGTCTCGA
Solyc02g079280_MYB41_F	GGATATGGTAATTGGAGGACTC
Solyc02g079280_MYB41_R	CTGGCCTTAGATAATTAGTCCA
Solyc02g085870_cer6_F	CCGTTACGTGCAGAGTACCC
Solyc02g085870_cer6_R	CACCAAGACCTGACCTTTCAAG
Solyc03g065250_cer1_F	GTGGGACGTAGCATTGAGTC
Solyc03g065250_cer1_R	TTCGATTATCACCCTTTGCAGT
Solyc03g117800_cer3_F	GAGCATGGAGGATATTTGGTG
Solyc03g117800_cer3_R	CTTCATAAGACACCCTTCGC
Solyc04g050730_SlCUS2_F	CGAGCCTTCTTCGTGTTTG
Solyc04g050730_SlCUS2_R	ATGAGTAGGATAGTCAATGCC
Solyc09g014350_GPAT6_F	GCTCATCCCCATATTGAACCA
Solyc09g014350_GPAT6_R	ATGAGACTACCACCTTCTAAG
Solyc09g007920_PAL1_F	CAGGTTGGTGAGACAAGAACT
Solyc09g007920_PAL1_R	GATCTGTCCATTGCACATTGC
Solyc03g097500_ASFT_F	CTTTCCACACGACGGATTTCG
Solyc03g097500_ASFT_R	AGCACATTGACACTTCTCCTCT
Solyc07g005760_ASFTparal_F	CCAATTCATATGGGACCAGCTT
Solyc07g005760_ASFTparal_R	AACAGCCAAACGCAAGTTCCTA

Supplementary Table 20: Summary of IL-QTL related to drought and salinity tolerance. Red boxes denote significant QTL detected in the IL population for several phenotypic traits positively related to drought/salt tolerance.

Introgression line	Drought stress								Salt stress					
	Plant vegetative weight[1]	Total yield[1]	Mean fruit weight[1]	Brix X yield[1]	Fruit number[1]	Seedling survival rates[2]	Leaf damage rates[2]	Seed germination rate[3]	Plant height[4,5]	Less damage[6]	leaf dry mass[4]	root dry mass[4]	lower NaCl content[4]	seedling salt tolerance[7]
IL1-1	🟥													
IL1-1-2														
IL1-1-3														
IL1-2	🟥													
IL1-3									🟥					
IL1-4									🟥					
IL1-4-18			🟥							🟥				
IL2-1														
IL2-1-1														
IL2-2			🟥											
IL2-3	🟥			🟥						🟥				
IL2-4	🟥			🟥										
IL2-5	🟥				🟥	🟥	🟥			🟥			🟥	
IL2-6														
IL2-6-5			🟥	🟥										
IL3-1														
IL3-2														
IL3-3														
IL3-4										🟥				
IL3-5														
IL4-1														
IL4-1-1														
IL4-2														
IL4-3										🟥				
IL4-3-2														
IL4-4	🟥													
IL5-1										🟥				
IL5-2									🟥					
IL5-3			🟥											
IL5-4	🟥													
IL5-5														
IL6-1														
IL6-2														🟥
IL6-2-2														
IL6-3	🟥													
IL6-4														
IL7-1														🟥
IL7-2	🟥													
IL7-3	🟥													
IL7-4		🟥	🟥											
IL7-4-1		🟥	🟥	🟥										
IL7-5			🟥	🟥										
IL7-5-5														
IL8-1										🟥				
IL8-1-1														
IL8-1-3														
IL8-2														
IL8-2-1														
IL8-3		🟥		🟥	🟥			🟥						
IL8-3-1														
IL9-1						🟥	🟥			🟥				
IL9-1-2														
IL9-1-3														
IL9-2														
IL9-2-5														
IL9-2-6														
IL9-3														
IL9-3-1														
IL9-3-2														
IL10-1														
IL10-1-1										🟥				
IL10-2										🟥				
IL10-2-2														
IL10-3														
IL11-1											🟥			
IL11-2	🟥											🟥		
IL11-3														
IL11-4														
IL11-4-1														
IL12-1														
IL12-1-1			🟥	🟥										
IL12-2														
IL12-3										🟥				
IL12-3-1														
IL12-4														
IL12-4-1			🟥											

Supplementary Table 21: List of candidate stress genes (from Supplementary Table 22) contained in 4 major IL-QTL related to salt and drought tolerance (ILs 2-5, 7-4-1, 8-3 and 9-1).

Gene ID	IL	Tomato gene annotation
Solyc02g081390	2-5	Amine oxidase family protein (AHRD V1 **** Q1EPI3_MUSAC); contains Interpro domain(s) IPR002937 Amine oxidase
Solyc02g081700	2-5	Proteasome subunit alpha type (AHRD V1 ***- Q38HT0_SOLTU); contains Interpro domain(s) IPR001353 Proteasome, subunit alpha/beta
Solyc02g082590	2-5	Superoxide dismutase (AHRD V1 ***- B8B5M4_ORYSI); contains Interpro domain(s) IPR009003 Peptidase, trypsin-like serine and cysteine
Solyc02g084240	2-5	H1 histone-like protein (AHRD V1 ***- Q43511_SOLLC); contains Interpro domain(s) IPR005818 Histone H1/H5
Solyc02g084440	2-5	Fructose-bisphosphate aldolase (AHRD V1 ***- Q9SXX4_NICPA); contains Interpro domain(s) IPR000741 Fructose-bisphosphate aldolase, class-I
Solyc02g084840	2-5	Dehydrin DHN1 (AHRD V1 *-*- DHN1_PEA); contains Interpro domain(s) IPR000167 Dehydrin
Solyc02g084850	2-5	Unknown Protein (AHRD V1); contains Interpro domain(s) IPR000167 Dehydrin
Solyc02g086670	2-5	Glycogen synthase kinase (AHRD V1 **** C7AE95_SOYBN); contains Interpro domain(s) IPR002290 Serine/threonine protein kinase
Solyc02g088710	2-5	4-coumarate CoA ligase-like (AHRD V1 **** Q84K86_NICSY); contains Interpro domain(s) IPR000873 AMP-dependent synthetase and ligase
Solyc02g089610	2-5	S-adenosylmethionine decarboxylase proenzyme (AHRD V1 **** Q7XZQ9_VITVI); contains Interpro domain(s) IPR001985 S-adenosylmethionine decarboxylase
Solyc02g089620	2-5	Proline dehydrogenase (AHRD V1 **** A1E289_ACTDE); contains Interpro domain(s) IPR015659 Proline oxidase
Solyc02g089630	2-5	Proline dehydrogenase (AHRD V1 **** A1E289_ACTDE); contains Interpro domain(s) IPR015659 Proline oxidase
Solyc02g090680	2-5	Cyclin-dependent kinase inhibitor 7 (AHRD V1 *-** KRP7_ARATH); contains Interpro domain(s) IPR016701 Cyclin-dependent kinase inhibitor, plant
Solyc02g093050	2-5	WRKY transcription factor 26 (AHRD V1 ***- C9DI15_9ROSI); contains Interpro domain(s) IPR003657 DNA-binding WRKY
Solyc07g005650	7-4-1	WRKY transcription factor (AHRD V1 ***- D4P3Y2_9ROSI); contains Interpro domain(s) IPR003657 DNA-binding WRKY
Solyc07g005760	7-4-1	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase (AHRD V1 **-* B2Z6Q6_POPTR); contains Interpro domain(s) IPR003480 Transferase
Solyc07g006500	7-4-1	Alpha alpha-trehalose-phosphate synthase (UDP-forming) (AHRD V1 **** D2REU5_ARCPA); contains Interpro domain(s) IPR001830 Glycosyl transferase, family 20
Solyc07g007670	7-4-1	Purple acid phosphatase 3 (AHRD V1 **** Q6J5M8_SOLTU); contains Interpro domain(s) IPR015914 Purple acid phosphatase, N-terminal
Solyc07g007870	7-4-1	NADH flavin oxidoreductase/12-oxophytodienoate reductase (AHRD V1 **-* Q2U7C4_ASPOR); contains Interpro domain(s) IPR001155 NADH:flavin oxidoreductase/NADH oxidase, N-terminal
Solyc07g008310	7-4-1	Rieske (2Fe-2S) domain protein (AHRD V1 **-- Q024N8_SOLUE); contains Interpro domain(s) IPR001663 Aromatic-ring-hydroxylating dioxygenase, alpha subunit
Solyc07g008320	7-4-1	Calcium-transporting ATPase 1 (AHRD V1 **** Q7XBH9_CERRI); contains Interpro domain(s) IPR006408 ATPase, P-type, calcium-transporting,

		PMCA-type
Solyc07g014680	7-4-1	Potassium transporter (AHRD V1 *-** A0MNZ1_THEHA); contains Interpro domain(s) IPR003445 Cation transporter
Solyc07g014690	7-4-1	Potassium transporter (AHRD V1 **** A0MNZ1_THEHA); contains Interpro domain(s) IPR003445 Cation transporter
Solyc07g015960	7-4-1	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase-like protein (AHRD V1 **-* B9N329_POPTR); contains Interpro domain(s) IPR003480 Transferase
Solyc07g026720	7-4-1	Calcium ATPase (AHRD V1 ***- D5JXY5_NICBE); contains Interpro domain(s) IPR006068 ATPase, P-type cation-transporter, C-terminal
Solyc07g039310	7-4-1	Polyamine oxidase (AHRD V1 **** Q8LL67_AMAHP); contains Interpro domain(s) IPR002937 Amine oxidase
Solyc07g043590	7-4-1	Amine oxidase family protein (AHRD V1 **** Q1EPI3_MUSAC); contains Interpro domain(s) IPR002937 Amine oxidase
Solyc07g043640	7-4-1	Acyl-CoA synthetase/AMP-acid ligase II (AHRD V1 **** D0C359_9GAMM); contains Interpro domain(s) IPR000873 AMP-dependent synthetase and ligase
Solyc08g078880	8-3	Cation/H(+) antiporter 15 (AHRD V1 **** CHX15_ARATH); contains Interpro domain(s) IPR006153 Cation/H+ exchanger
Solyc08g079180	8-3	Elongation factor G (AHRD V1 ***- Q9SI75_ARATH); contains Interpro domain(s) IPR004540 Translation elongation factor EFG/EF2
Solyc08g079430	8-3	Primary amine oxidase (AHRD V1 ***- B9RBR2_RICCO); contains Interpro domain(s) IPR000269 Copper amine oxidase
Solyc08g079830	8-3	Cu/Zn-superoxide dismutase copper chaperone (AHRD V1 **** Q9BBU5_SOYBN); contains Interpro domain(s) IPR001424 Superoxide dismutase, copper/zinc binding
Solyc08g080190	8-3	Choline dehydrogenase (AHRD V1 ***- A8NUQ9_COPC7); contains Interpro domain(s) IPR012132 Glucose-methanol-choline oxidoreductase
Solyc08g080370	8-3	Acetylornithine aminotransferase (AHRD V1 **** D8THT2_VOLCA); contains Interpro domain(s) IPR004636 Acetylornithine and succinylornithine aminotransferase
Solyc08g080590	8-3	Osmotin 81 (Fragment) (AHRD V1 **-- Q84ML3_SOLTU); contains Interpro domain(s) IPR001938 Thaumatin, pathogenesis-related
Solyc08g080600	8-3	Osmotin 81 (Fragment) (AHRD V1 **-- Q84MK2_SOLTU); contains Interpro domain(s) IPR001938 Thaumatin, pathogenesis-related
Solyc08g080610	8-3	Osmotin-like protein (Fragment) (AHRD V1 **-- Q8S4L1_SOLNI); contains Interpro domain(s) IPR001938 Thaumatin, pathogenesis-related
Solyc08g080660	8-3	Osmotin-like protein (Fragment) (AHRD V1 **-- Q8S4L1_SOLNI); contains Interpro domain(s) IPR001938 Thaumatin, pathogenesis-related
Solyc08g080940	8-3	Glutathione peroxidase (AHRD V1 ***- Q4VY91_CAPCH); contains Interpro domain(s) IPR000889 Glutathione peroxidase
Solyc08g081220	8-3	Cytochrome P450
Solyc08g081530	8-3	Reductase (AHRD V1 ***- B5HVX9_9ACTO); contains Interpro domain(s) IPR013027 FAD-dependent pyridine nucleotide-disulphide oxidoreductase
Solyc08g081810	8-3	Cation/H(+) antiporter 18 (AHRD V1 **** CHX18_ARATH); contains Interpro domain(s) IPR006153 Cation/H+ exchanger
Solyc08g081820	8-3	Cation/H+ antiporter (AHRD V1 **** Q8A6U3_BACTN); contains Interpro domain(s) IPR006153 Cation/H+ exchanger
Solyc09g005620	9-1	Glutaredoxin (AHRD V1 *-*- D7G070_ECTSI); contains Interpro domain(s)

		IPR004480 Glutaredoxin-related protein
Solyc09g007180	9-1	Adenylate kinase (AHRD V1 **** B6SLP1_MAIZE); contains Interpro domain(s) IPR006259 Adenylate kinase, subfamily
Solyc09g007290	9-1	Nuclear transcription factor Y subunit B-3 (AHRD V1 **** B6UBN3_MAIZE); contains Interpro domain(s) IPR003957 Transcription factor, CBFA/NFYB, DNA topoisomerase
Solyc09g007910	9-1	Phenylalanine ammonia-lyase (AHRD V1 **** B5LAW0_CAPAN); contains Interpro domain(s) IPR005922 Phenylalanine ammonia-lyase
Solyc09g007920	9-1	Phenylalanine ammonia-lyase (AHRD V1 **** B5LAW0_CAPAN); contains Interpro domain(s) IPR005922 Phenylalanine ammonia-lyase
Solyc09g008770	9-1	Group 3 late embryogenesis abundant protein (AHRD V1 **-- Q2N1E0_PHAVU); contains Interpro domain(s) IPR004238 Late embryogenesis abundant protein
Solyc09g009100	9-1	Heat stress transcription factor A3 (AHRD V1 ***- D1M7W9_SOLLC); contains Interpro domain(s) IPR000232 Heat shock factor (HSF)-type, DNA-binding
Solyc09g009390	9-1	Monodehydroascorbate reductase (NADH)-like protein (AHRD V1 **** Q0WUJ1_ARATH); contains Interpro domain(s) IPR013027 FAD-dependent pyridine nucleotide-disulphide oxidoreductase
Solyc09g010530	9-1	Cation/H+ antiporter (AHRD V1 **** D1JSI8_9BACE); contains Interpro domain(s) IPR006153 Cation/H+ exchanger
Solyc09g010630	9-1	heat shock protein (AHRD V1 ***- B2D2G5_CAPSN); contains Interpro domain(s) IPR013126 Heat shock protein 70

Supplementary Table 23: Differential expression and prediction of *cis*-regulatory elements detected in a subset of candidate genes related to salt and drought tolerance.

Gene ID	Annotation	Log2 fold-change (seedling, *S. pennellii*/*S. lycopersicum*)[27]	IL	References	cis-acting elements located in promoter regions (sequence)	insertion/deletion with respect to *S. pen* sequence	Function
Solyc02g084240	H1 histone-like protein	2.14	2-5	Scippa et al 2004[138]	Interspersed SNPs		
Solyc02g084850	Dehydrin (tas14 gene)	2.59	2-5	Godoy et al 1994;[121] Munoz-mayor et al 2012[122]	MYCATRD22 (CATGTG)	insertion	binding site for MYC factors; involved in response to drought
					POLYASIG3 (AATAAT)	insertion	polyadenylation signal of plant mRNAs
					SURECOREATSULTR11 (GAGAC)	insertion	core of sulfur-responsive element (SURE), involved in responses to sulphur deficiency
					TCA1MOTIF (TCATCTTCTT)	insertion	element for salicylic acid response and multiple abiotic stresses
Solyc06g005170	Mitogen-activated protein kinase 3	-0.68	6-1	Sinha et al 2011[134]	TATABOX5 (TTATTT)	insertion	common promoter element
					CAATBOX1 (CAAT)	insertion	common promoter element
					MYBCORE (CNGTTR)	insertion	consensus binding site for MYB factors
					POLYASIG3 (AATAAT)	insertion	polyadenylation signal of plant mRNAs
					WBOXHVISO1 (TGACT)	insertion	binding element of WRKY transcription factors, sugar signalling
Solyc06g060630	Cation/H(+) antiporter 15	-6.24	6-1	Pardo et al 2006[158]	MARTBOX (TTWTWTTWTT)	deletion	motif found in matrix attachment regions (MARs)
					TATABOX5 (TTATTT)	deletion	common promoter element
					POLYASIG1 (AATAAA)	deletion	polyadenylation signal of plant mRNAs
					NTBBF1ARROLB (ACTTTA)	deletion	Dof binding site required for auxin-responsive expression
					CACTFTPPCA1 (YACT)	deletion	motif found in the distal promoter of phosphoenolpyruvate carboxylase
Solyc06g019170	Gamma-glutamyl phosphate reductase	0.98	6-1	Garcia-Rios et al 1997[159]	DOFCOREZM (AAAG)	insertion	core motif for binding of Dof proteins

					POLYASIG1 (AATAAA)	insertion	polyadenylation signal of plant mRNAs
Solyc06g035580	Choline dehydrogenase	-2.33	6-1	Sakamoto & Murata 2000[160]	DOFCOREZM (AAAG)	deletion	core motif for binding of Dof proteins
					POLYASIG1 (AATAAA)	deletion	polyadenylation signal of plant mRNAs
					TATABOX5 (TTATTT)	deletion	common promoter element
					MYBCORE (CNGTTR)	deletion	consensus binding site for MYB factors
					GTGANTG10 (GTGA)	deletion	motif found in the promoter of the tobacco pollen gene g10
Solyc06g049080	Manganese/iron superoxide dismutase	-0.59	6-1	Mittova et al 2002a,b[136,137] Tsang et al 1991	DOFCOREZM (AAAG)	insertion	core motif for binding of Dof proteins
					ARR1AT (NGATT)	insertion	sequence involved in cytokinin-induced reponses
					SEF4MOTIFGM7S (RTTTTTR)	insertion	SEF4 binding site
Solyc07g008320	Ca2+ transporting ATPase	0.59	7-4-1	Huda et al 2013[161]	SEF3MOTIFGM (AACCCA)	insertion	SEF3 binding site, sequence found in upstream regions of soya globulin genes
					ANAERO1CONSENSUS (AAACAAA)	insertion	motif commonly found in genes induced by anaerobiosis
					CAATBOX1 (CAAT)	insertion	common promoter element
					TATABOX5 (TTATTT)	insertion	common promoter element
Solyc07g014680	Potassium transporter	0.74	7-4-1	Asins et al 2013;[116]	RAV1AAT (CAACA)	deletion	binding consensus sequence for Arabidopsis RAV1, a TF involved in responses to pathogens and osmotic stress
					POLYASIG3 (AATAAT)	deletion	polyadenylation signal of plant mRNAs
					GT1CONSENSUS (GRWAAW)	deletion	common element found in light- and SA-induced genes
Solyc07g006500	alpha-trehalose-phosphate synthase	-1.31	7-4-1	Chary et al 2008[162]	PYRIMIDINEBOXOSRAMY1A (CCTTTT)	deletion	pyrimidine box found in rice involved in sugar repression
					GTGANTG10 (GTGA)	deletion	motif found in the promoter of the tobacco pollen gene g10
					DOFCOREZM (AAAG)	deletion	core motif for binding of Dof proteins
					SREATMSD (TTATCC)	deletion	sugar-repressive element (SRE): involved in regulation of gene expression during axillary bud outgrowth in Arabidopsis
Solyc07g065500	Nuclear transcription factor Y subunit B-3	0.59	7-1	Gong et al 2010[140]	GATABOX (GATA)	insertion	element commonly found upstream of light-regulated genes
					-10PEHVPSBD (TATTCT)	insertion	element involved in light-regulated transcription of chlorplasto gene psbD

					POLYASIG3 (AATAAT)	deletion	polyadenylation signal of plant mRNAs
					CAATBOX1 (CAAT)	deletion	common promoter element
					CARGCW8GAT (CWWWWWWWWG)	deletion	a variant of CArG motif; binding site for MADS domain protein AGL15
					RAV1AAT (CAACA)	deletion	binding consensus sequence for Arabidopsis RAV1, a TF involved in responses to pathogens and osmotic stress
Solyc08g006720	Glutathione peroxidase	0.73	8-1-1	Mittova et al 2002a,b[136,137]	CACTFTPPCA1 (YACT)	insertion	motif found in the distal promoter of phosphoenolpyruvate carboxylase
Solyc08g014420	Mitogen-activated PK 2	0.31	8-1-1	Sinha et al 2011[134]	Interspersed SNPs		
Solyc08g016160	Cation/H(+) antiporter 15	not expressed in S. pennellii	8-1-1	Pardo et al 2006[159]	TATABOX5 (TTATTT)	insertion	common promoter element
					POLYASIG1 (AATAAA)	insertion	polyadenylation signal of plant mRNAs
					DOFCOREZM (AAAG)	insertion	core motif for binding of Dof proteins
					ABRELATERD1 (ACGTG)	insertion	ABRE-like sequences involved in dehydration-induced expression
					CAATBOX1 (CAAT)	insertion	common promoter element
Solyc08g075750	ATP-dependent Clp protease proteolytic subunit	0.35	8-3	Sjogren et al 2006[163]	CAREOSREP1 (CAACTC)	insertion	cis-acting element found upstream of gibberellin-induced cysteine proteinase gene
					CCA1ATLHCB1	insertion	binding site for CCA1 (MYB-related), involved in phytochrome signalling
					ARR1AT (NGATT)	insertion	sequence involved in cytokinin-induced reponses
					CAATBOX1 (CAAT)	insertion	common promoter element
					MYBATRD22 (CTAACCA)	insertion	binding site for AtMYB2, involved in dehydration and ABA-induced responses
					DOFCOREZM (AAAG)	insertion	core motif for binding of Dof proteins
					WBOXHVISO1 (TGACT)	insertion	binding element of WRKY transcription factors, sugar signalling
Solyc08g079830	Cu/Zn-superoxide dismutase	1.32	8-3	Mittova et al 2002a,b[136,137]	MYBPLANT (MACCWAMC)	insertion	MYB binding site, commonly found in promoters of phenylpropanoid synthesis genes
					SURECOREATSULTR11 (GAGAC)	insertion	core of sulfur-responsive element (SURE), involved in responses to sulphur deficiency
					DPBFCOREDCDC3 (ACACNNG)	insertion	binding sequence for bZIP transcription factors DPBF1 and 2, involved in ABA signalling
					DOFCOREZM (AAAG)	insertion	core motif for binding of Dof proteins
					GT1CONSENSUS (GRWAAW)	insertion	common element found in light- and SA-induced genes
					GATABOX (GATA)	insertion	element commonly found upstream of light-regulated genes

					IBOX (GATAAG)	insertion	conserved sequence upstream of light-regulated genes
					CAATBOX1 (CAAT)	insertion	common promoter element
Solyc08g080370	Acetylornithine aminotransferase	0.22	8-3	Sharma & Verslues 2010[164]	Interspersed SNPs		
Solyc09g008770	LEA protein (group 3)	4.24	9-1	Battaglia et al 2008[119]	GT1CONSENSUS	deletion	common element found in light- and SA-induced genes
					CAATBOX1 (CAAT)	deletion	common promoter element
					IBOXCORE (GATAA)	deletion	element for light-regulated transcritpion
Solyc09g010530	Cation/H+ antiporter	0.94	9-1	Pardo et al 2006[159]	Interspersed SNPs		
Solyc09g009100	Heat stress transcription factor A3	-0.41	9-1	Bharti et al 2000[165]	MYCCONSENSUSAT (CANNTG)	insertion	MYC recognition sequence; involved in response to drought and other abiotic stresses
					ARR1AT (NGATT)	insertion	sequence involved in cytokinin-induced reponses
					POLYASIG3 (AATAAT)	insertion	polyadenylation signal of plant mRNAs
					MYBGAHV (TAACAAA)	insertion	MYB recognition sequence found in promoters of gibberellin-responsive genes
					SEF4MOTIFGM7S (RTTTTTR)	insertion	SEF4 binding site
Solyc11g010500	Mitochondrial carrier family	-0.98	11-1	Palmieri et al 2011[166]	BOXIINTPATPB (ATAGAA)	deletion	conserved sequence found in plastid atpB gene promoter
					CACTFTPPCA1 (YACT)	deletion	motif found in the distal promoter of phosphoenolpyruvate carboxylase
					GT1GMSCAM4 (GAAAAA)	insertion	GT-1 motif involved in pathogen and salt stress-induced expression
					MARTBOX (TTWTWTTWTT)	insertion	motif found in matrix attachment regions (MARs)
Solyc11g011340	Alcohol dehydrogenase	2.7	11-1	Bird & Wilson 1994[167]	WBOXATNPR1 (TTGAC)	insertion	W-box; binding sequence for WRKY transcription factros, involeved in SA response
					POLYASIG2 (AATTAAA)	insertion	polyadenylation signal of plant mRNAs
					CACTFTPPCA1 (YACT)	insertion	motif found in the distal promoter of phosphoenolpyruvate carboxylase
					MYCCONSENSUSAT (CANNTG)	insertion	MYC recognition sequence; involved in response to drought and other abiotic stresses
					TATABOX5 (TTATTT)	insertion	common promoter element
Solyc11g017470	NAC domain protein	2.1	11-1	Puranik et al 2012[130]	GATABOX (GATA)	insertion	element commonly found upstream of light-regulated genes
					TATABOX5 (TTATTT)	insertion	common promoter element
					POLYASIG3 (AATAAT)	insertion	polyadenylation signal of plant mRNAs

Supplementary Table 28: Samples used for RNA sequencing. This table shows the different tissues and/or conditions used to represent a maximal sampling space of the *S. pennellii* transcriptome. The column experimental group indicates experiments where some replication existed and the resulting tests are shown in Supplementary Dataset 15.

Internal Label	ng/ul	A260	260/280	Sample Label	Experimental Group
a1	561.65	14.04	2.14	Diurnal, D + 4, Leaf	
a2	383.46	9.59	2.08	Diurnal, D + 8, Leaf	
a3	965.41	24.13	2.17	Diurnal, D + 12, Leaf	
a4	755.15	22.91	2.18	Diurnal, ED (D + 14), Leaf	
a5	696.48	18.88	2.18	Diurnal, N + 4, Leaf	
a6	501.86	17.41	2.14	Diurnal, N + 8, Leaf	
a7	530.56	12.55	2.17	Diurnal, EN, Leaf	
a8	530.56	13.26	2.16	Diurnal, XN + 12, Leaf	
a9	455.46	11.39	2.11	Diurnal, XN + 28, Leaf	
a10	753.78	18.84	2.17	Diurnal, XN + 36, Leaf	
a11	1165.93	29.15	2.15	6wk, Small leaves	
a12	597.35	14.93	2.14	6wk, Mature leaves	
a13	269.05	6.73	2.15	6wk, Meristem	
a14	269.88	6.75	2.15	6wk, Stem	
a15	112.14	2.80	1.52	Pseudomonas(24h), Infected leaves	
a16	21.98	0.55	1.82	Pseudomonas(24h), Uninfected leaves	
a17	1040.03	26.00	2.15	Pseudomonas(24h), Small leaves	
a18	1237.10	30.93	2.16	Cold(4C, 24h), Small leaves	
a19	698.08	17.45	2.15	Cold(4C, 24h), Mature leaves	
a20	1632.55	40.81	2.14	Strong UV(72h), Small leaves	
a21	312.02	7.80	1.54	Strong UV(72h), Mature leaves	
a22	885.13	22.13	2.14	Drought(1wk), Small leaves	
a23	604.86	15.12	2.16	Drought(1wk), Mature leaves	

a24	412.58	10.32	2.13	Drought(1wk), Root	
a25	1450.83	36.27	2.15	Weak UV(120h), Small leaves	
a26	591.94	14.80	2.16	Weak UV(120h), Mature leaves	
a27	821.13	20.53	2.17	Insect(72h), Damaged leaves	
a28	466.04	11.65	2.18	Insert(72h), Undamaged leaves	
a29	417.03	10.43	2.12	Anthocynin, Mature leaves	
a30	1058.96	26.47	2.21	Pollen	
a31	1132.94	28.32	2.13	Exterior(10d), Small leaves	
a32	395.35	9.88	2.00	Exterior(10d), Mature leaves	
a33	1266.60	31.67	2.14	HP lo salt, Shoot	b
a34	1113.84	27.85	2.15	HP hi salt, Shoot	b
a35	77.97	1.95	2.07	HP -N, Shoot	b
a36	2174.79	54.37	2.13	HP -Fe, Shoot	b
a37	1482.29	37.06	2.13	HP -Mg, Shoot	b
a38	734.37	18.36	2.13	HP -Ca, Shoot	b
a39	493.69	12.34	2.13	HP H20, Shoot	b
a40	536.55	13.41	2.13	HP H20, Shoot	b
a41	409.28	10.23	2.11	HP lo salt, Root	c
a42	340.63	8.52	2.13	HP hi salt, Root	c
a43	233.38	5.84	2.14	HP -N, Root	c
a44	479.12	11.98	2.14	HP -Fe, Root	c
a45	371.67	9.29	2.12	HP -Mg, Root	c
a46	612.76	15.32	2.16	HP -Ca, Root	c
a47	209.97	5.25	2.13	HP H20, Root	c
a48	401.40	10.04	2.12	HP H20, Root	c
d1	724.99	18.13	2.13	Seedling shoot	a
d2	784.31	19.61	2.13	Seedling shoot	a
d3	915.36	22.88	2.13	Seedling shoot	a
d4	738.71	18.47	2.13	Seedling shoot	a

d5	759.28	18.98	2.14	Seedling shoot	a
d6	844.00	21.10	2.12	Seedling shoot	a
d7	433.57	10.84	2.07	Seedling root	a
d8	399.55	9.99	2.03	Seedling root	a
d9	432.10	10.80	2.07	Seedling root	a
d10	342.07	8.55	2.09	Seedling root	a
d11	93.16	2.33	2.27	Seedling root	a
d12	47.67	1.19	2.24	Seedling root	a
d13	1212.97	30.32	13.89	Bud	a
d14	703.54	17.59	7.92	Bud	a
d15	1436.48	35.91	16.39	Bud	a
d16	206.92	5.17	2.37	Flower	a
d17	111.72	2.79	1.33	Flower	a
d18	302.63	7.57	3.50	Flower	a
d19	553.39	13.84	2.12	Immature fruit	a
d20	352.83	8.82	2.17	Immature fruit	a
d21	194.11	4.85	2.15	Immature fruit	a
d22	431.63	10.79	2.19	Mature fruit	a
d23	401.59	10.04	2.17	Mature fruit	a
d24	283.42	7.09	2.16	Mature fruit	a

Supplementary Table 29: Correlation between expression in different tissues across species. The expression estimates (rpkm values) for the *S. lycopersicum* cv. Heinz and *S. pimpinellifolium* were extracted from the tomato genome publication and compared to the values obtained here. *S. pennellii* tissues are in the rows (green) and *S. lycopersicum* cv. Heinz and *S. pimpinellifolium* are in columns (red and violet, respectively). The highest correlation value for a row and per species is marked in bold.

	Heinz_bud	Heinz_flower	Heinz_leaf	Heinz_root	Heinz_1cm_fruit	Heinz_2cm_fruit	Heinz_3cm_fruit	Heinz_MG_fruit	Heinz_B_fruit	Heinz_B.10_fruit	Pimp_leaf	Pimp_IG_fruit	Pimp_B_fruit	Pimp_B.5_fruit
Bud	0.542915	**0.557337**	0.525325	0.260548	0.537159	0.435556	0.355012	0.259259	0.23249	0.17986	**0.652359**	0.18879	0.18109	0.10208
Flower	0.212942	**0.337884**	0.068381	0.039934	0.052605	0.054957	0.055831	0.050656	0.046892	0.039901	0.088358	0.043445	0.041648	0.023977
Seedling.shoot	0.356604	0.48018	**0.620167**	0.089461	0.258284	0.259111	0.213637	0.146482	0.118686	0.072338	**0.870568**	0.122814	0.095886	0.044678
Seedling.root	0.291455	0.448914	0.284995	**0.69159**	0.440649	0.501226	0.531237	0.436322	0.360833	0.296096	0.338161	0.308715	0.275758	0.153328
Immature.fruit	0.116957	0.18151	0.036364	0.21586	0.230324	0.261448	0.269892	**0.277571**	0.202928	0.157734	0.077559	**0.248437**	0.176186	0.083662
Mature.fruit	0.118139	0.197049	0.065472	0.178238	0.179174	0.209355	0.220549	**0.22446**	0.178808	0.158937	0.11734	**0.230908**	0.175755	0.091145

12 References

1 Peterson, D. G., Boehm, K. S. & Stack, S. M. Isolation of milligram quantities of nuclear DNA from tomato (Lycopersicon esculentum), a plant containing high levels of polyphenolic compounds. *Plant Molecular Biology Reporter* **15**, 148-153, doi:Doi 10.1007/Bf02812265 (1997).

2 Kamenetzky, L. *et al.* Genomic analysis of wild tomato introgressions determining metabolism- and yield-associated traits. *Plant Physiol* **152**, 1772-1786, doi:10.1104/pp.109.150532 (2010).

3 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, doi:10.1093/bioinformatics/btu170 (2014).

4 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

5 Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317, doi:10.1038/nature08696 (2010).

6 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).

7 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).

8 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

9 Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol* **13**, R56, doi:10.1186/gb-2012-13-6-r56 (2012).

10 Fulton, T. M., Van der Hoeven, R., Eannetta, N. T. & Tanksley, S. D. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* **14**, 1457-1467 (2002).

11 Tanksley, S. D. *et al.* High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**, 1141-1160 (1992).

12 Van Schalkwyk, A. *et al.* Bin mapping of tomato diversity array (DArT) markers to genomic regions of Solanum lycopersicum x Solanum pennellii introgression lines. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **124**, 947-956, doi:10.1007/s00122-011-1759-5 (2012).

13 Sim, S. C. *et al.* High-density SNP genotyping of tomato (Solanum lycopersicum L.) reveals patterns of genetic variation due to breeding. *PLoS One* **7**, e45520, doi:10.1371/journal.pone.0045520 (2012).

14 Chitwood, D. H. *et al.* A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *Plant Cell* **25**, 2465-2481, doi:10.1105/tpc.113.112391 (2013).

15 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

16 Homer, N. & Nelson, S. F. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol* **11**, R99, doi:10.1186/gb-2010-11-10-r99 (2010).

17 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

18 Engstrom, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**, 1185-1191, doi:10.1038/nmeth.2722 (2013).

19 Kahlau, S., Aspinall, S., Gray, J. C. & Bock, R. Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. *J Mol Evol* **63**, 194-207, doi:10.1007/s00239-005-0254-5 (2006).

20 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

21 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644, doi:10.1093/bioinformatics/btn013 (2008).

22 Sato, S. *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641, doi:Doi 10.1038/Nature11119 (2012).

23 Bugos, R. C. *et al.* RNA isolation from plant tissues recalcitrant to extraction in guanidine. *Biotechniques* **19**, 734-737 (1995).

24 Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664, doi:10.1101/gr.229202 (2002).

25 Lohse, M. *et al.* Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ*, doi:10.1111/pce.12231 (2013).

26 Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**, D1202-1210, doi:10.1093/nar/gkr1090 (2012).

27 Koenig, D. *et al.* Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc Natl Acad Sci U S A* **110**, E2655-2662, doi:10.1073/pnas.1309606110 (2013).

28 Harriman, R. W., Tieman, D. M. & Handa, A. K. Molecular cloning of tomato pectin methylesterase gene and its expression in rutgers, ripening inhibitor, nonripening, and never ripe tomato fruits. *Plant Physiol* **97**, 80-87 (1991).

29 Tieman, D. M., Harriman, R. W., Ramamohan, G. & Handa, A. K. An Antisense Pectin Methylesterase Gene Alters Pectin Chemistry and Soluble Solids in Tomato Fruit. *Plant Cell* **4**, 667-679, doi:10.1105/tpc.4.6.667 (1992).

30 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

31 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 (2003).

32 Dievart, A. & Clark, S. E. LRR-containing receptors regulating plant development and defense. *Development* **131**, 251-261, doi:10.1242/dev.00998 (2004).

33 Afzal, A. J., Wood, A. J. & Lightfoot, D. A. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Molecular plant-microbe interactions : MPMI* **21**, 507-517, doi:10.1094/MPMI-21-5-0507 (2008).

34 Caplan, J., Padmanabhan, M. & Dinesh-Kumar, S. P. Plant NB-LRR immune receptors: from recognition to transcriptional reprogramming. *Cell host & microbe* **3**, 126-135, doi:10.1016/j.chom.2008.02.010 (2008).

35 Andolfo, G. *et al.* Overview of tomato (Solanum lycopersicum) candidate pathogen recognition genes reveals important Solanum R locus dynamics. *New Phytol* **197**, 223-237, doi:10.1111/j.1469-8137.2012.04380.x (2013).

36 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS computational biology* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

37 Felsenstein, J. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

38 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113, doi:10.1186/1471-2105-5-113 (2004).

39 Usadel, B. *et al.* PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics* **7**, 535 (2006).

40 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**, 276-277 (2000).

41 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

42 Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Research* **11**, 863-874, doi:Doi 10.1101/Gr.176601 (2001).

43 Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database (vol 40, pg D306, 2011). *Nucleic Acids Res* **40**, 4725-4725 (2012).

44 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**, D225-D229, doi:Doi 10.1093/Nar/Gkq1189 (2011).

45 Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M. & Solovyev, V. V. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res* **31**, 114-117, doi:Doi 10.1093/Nar/Gkg041 (2003).

46 Higo, K., Ugawa, Y., Iwamoto, M. & Korenaga, T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res* **27**, 297-300, doi:Doi 10.1093/Nar/27.1.297 (1999).

47 Schauer, N. *et al.* Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology* **24**, 447-454, doi:10.1038/nbt1192 (2006).

48 Schauer, N. *et al.* Mode of inheritance of primary metabolic traits in tomato. *Plant Cell* **20**, 509-523, doi:10.1105/tpc.107.056523 (2008).

49 Schilmiller, A. L. *et al.* Monoterpenes in the glandular trichomes of tomato are synthesized from a neryl diphosphate precursor rather than geranyl diphosphate. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 10865-10870, doi:10.1073/pnas.0904113106 (2009).

50 Tieman, D. *et al.* Functional analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor volatile methyl salicylate. *Plant J* **62**, 113-123, doi:10.1111/j.1365-313X.2010.04128.x (2010).

51 Kochevenko, A. & Fernie, A. R. The genetic architecture of branched-chain amino acid accumulation in tomato fruits. *J Exp Bot* **62**, 3895-3906, doi:10.1093/jxb/err091 (2011).

52 Mageroy, M. H., Tieman, D. M., Floystad, A., Taylor, M. G. & Klee, H. J. A Solanum lycopersicum catechol-O-methyltransferase involved in synthesis of the flavor molecule guaiacol. *Plant J* **69**, 1043-1051, doi:10.1111/j.1365-313X.2011.04854.x (2012).

53 Schilmiller, A. L., Charbonneau, A. L. & Last, R. L. Identification of a BAHD acetyltransferase that produces protective acyl sugars in tomato trichomes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16377-16382, doi:10.1073/pnas.1207906109 (2012).

54 Klee, H. J. & Giovannoni, J. J. Genetics and control of tomato fruit ripening and quality attributes. *Annual review of genetics* **45**, 41-59, doi:10.1146/annurev-genet-110410-132507 (2011).

55 Olson, D. C., White, J. A., Edelman, L., Harkins, R. N. & Kende, H. DIFFERENTIAL EXPRESSION OF 2 GENES FOR 1-AMINOCYCLOPROPANE-1-CARBOXYLATE SYNTHASE IN TOMATO FRUITS. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 5340-5344, doi:10.1073/pnas.88.12.5340 (1991).

56 Lincoln, J. E. *et al.* LE-ACS4, A FRUIT RIPENING AND WOUND-INDUCED 1-AMINOCYCLOPROPANE-1-CARBOXYLATE SYNTHASE GENE OF TOMATO (LYCOPERSICON-ESCULENTUM) - EXPRESSION IN ESCHERICHIA-COLI, STRUCTURAL CHARACTERIZATION, EXPRESSION CHARACTERISTICS, AND PHYLOGENETIC ANALYSIS. *J Biol Chem* **268**, 19422-19430 (1993).

57 Sitrit, Y. & Bennett, A. B. Regulation of tomato fruit polygalacturonase mRNA accumulation by ethylene: A re-examination. *Plant Physiol* **116**, 1145-1150, doi:10.1104/pp.116.3.1145 (1998).

58 Lashbrook, C. C., Tieman, D. M. & Klee, H. J. Differential regulation of the tomato ETR gene family throughout plant development. *Plant J* **15**, 243-252, doi:10.1046/j.1365-313X.1998.00202.x (1998).

59 Kevany, B. M., Tieman, D. M., Taylor, M. G., Dal Cin, V. & Klee, H. J. Ethylene receptor degradation controls the timing of ripening in tomato fruit. *Plant J* **51**, 458-467, doi:10.1111/j.1365-313X.2007.03170.x (2007).

60 Krasnyanski, S. F., Sandhu, J., Domier, L. L., Buetow, D. E. & Korban, S. S. Effect of an enhanced CaMV 35S promoter and a fruit-specific promoter on UIDA gene expression in transgenic tomato plants. *In Vitro Cellular & Developmental Biology-Plant* **37**, 427-433 (2001).

61 Kesanakurti, D., Kolattukudy, P. E. & Kirti, P. B. Fruit-specific overexpression of wound-induced tap1 under E8 promoter in tomato confers resistance to fungal pathogens at ripening stage. *Physiologia Plantarum* **146**, 136-148, doi:10.1111/j.1399-3054.2012.01626.x (2012).

62 Vrebalov, J. *et al.* Fleshy Fruit Expansion and Ripening Are Regulated by the Tomato SHATTERPROOF Gene TAGL1. *Plant Cell* **21**, 3041-3062, doi:10.1105/tpc.109.066936 (2009).

63 Bemer, M. *et al.* The Tomato FRUITFULL Homologs TDR4/FUL1 and MBP7/FUL2 Regulate Ethylene-Independent Aspects of Fruit Ripening. *Plant Cell* **24**, 4437-4451, doi:10.1105/tpc.112.103283 (2012).

64 Lin, Z. *et al.* SlTPR1, a tomato tetratricopeptide repeat protein, interacts with the ethylene receptors NR and LeETR1, modulating ethylene and auxin responses and development. *J Exp Bot* **59**, 4271-4287, doi:10.1093/jxb/ern276 (2008).

65 Chung, M. Y. *et al.* A tomato (Solanum lycopersicum) APETALA2/ERF gene, SlAP2a, is a negative regulator of fruit ripening. *Plant J* **64**, 936-947, doi:10.1111/j.1365-313X.2010.04384.x (2010).

66 Karlova, R. *et al.* Transcriptome and Metabolite Profiling Show That APETALA2a Is a Major Regulator of Tomato Fruit Ripening. *Plant Cell* **23**, 923-941, doi:10.1105/tpc.110.081273 (2011).

67 Lanahan, M. B., Yen, H. C., Giovannoni, J. J. & Klee, H. J. The never ripe mutation blocks ethylene perception in tomato. *Plant Cell* **6**, 521-530 (1994).

68 Giovannoni, J. J. Genetic regulation of fruit development and ripening. *Plant Cell* **16 Suppl**, S170-180, doi:10.1105/tpc.019158 (2004).

69 Grumet, R., Fobes, J. F. & Herner, R. C. Ripening behavior of wild tomato species. *Plant Physiol* **68**, 1428-1432 (1981).

70 Liu, Y. S. *et al.* There is more to tomato fruit colour than candidate carotenoid genes. *Plant Biotechnology Journal* **1**, 195-207, doi:10.1046/j.1467-7652.2003.00018.x (2003).

71 Rousseaux, M. C. *et al.* QTL analysis of fruit antioxidants in tomato using Lycopersicon pennellii introgression lines. *Theoretical and Applied Genetics* **111**, 1396-1408, doi:10.1007/s00122-005-0071-7 (2005).

72 Alba, R. *et al.* Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* **17**, 2954-2965, doi:10.1105/tpc.105.036053 (2005).

73 Nguyen, C. *et al.* Tomato GOLDEN2-LIKE Transcription Factors Reveal Molecular Gradients That Function during Fruit Development and Ripening. *Plant Cell* **26**, 585-601 (2014).

74 Vrebalov, J. *et al.* A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. *Science* **296**, 343-346, doi:10.1126/science.1068181 (2002).

75 Frary, A. *et al.* fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85-88, doi:10.1126/science.289.5476.85 (2000).

76 Chakrabarti, M. *et al.* A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17125-17130, doi:10.1073/pnas.1307313110 (2013).

77 Schauer, N. & Fernie, A. R. Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci* **11**, 508-516, doi:10.1016/j.tplants.2006.08.007 (2006).

78 Centeno, D. C. *et al.* Malate Plays a Crucial Role in Starch Metabolism, Ripening, and Soluble Solid Content of Tomato Fruit and Affects Postharvest Softening. *Plant Cell* **23**, 162-184, doi:10.1105/tpc.109.072231 (2011).

79 Fridman, E., Carrari, F., Liu, Y. S., Fernie, A. R. & Zamir, D. Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* **305**, 1786-1789, doi:10.1126/science.1101666 (2004).

80 Morgan, M. J. *et al.* Metabolic Engineering of Tomato Fruit Organic Acid Content Guided by Biochemical Analysis of an Introgression Line. *Plant Physiol* **161**, 397-407, doi:10.1104/pp.112.209619 (2013).

81 Schauer, N., Zamir, D. & Fernie, A. R. Metabolic profiling of leaves and fruit of wild species tomato: a survey of the Solanum lycopersicum complex. *J Exp Bot* **56**, 297-307, doi:10.1093/jxb/eri057 (2005).

82 Ho, L. C. The mechanism of assimilate partitioning and carbohydrate compartmentation in fruit in relation Ito the quality and yield of tomato. *J Exp Bot* **47**, 1239-1243, doi:10.1093/jxb/47.Special_Issue.1239 (1996).

83 Balibrea, M. E., Estan, M. T., Bolarin, M. C., Perez-Alfocea, F. & Dell'amico, J. M. Carbon partitioning in tomato plants growing under salinity. *J Exp Bot* **47**, 1320-1321 (1996).

84 Balibrea, M. E., Parra, M., Bolarin, M. C. & Perez-Alfocea, F. Cytoplasmic sucrolytic activity controls tomato fruit growth under salinity. *Australian Journal of Plant Physiology* **26**, 561-568 (1999).

85 Sun, J. D., Loboda, T., Sung, S. J. S. & Black, C. C. SUCROSE SYNTHASE IN WILD TOMATO, LYCOPERSICON-CHMIELEWSKII, AND TOMATO FRUIT SINK STRENGTH. *Plant Physiol* **98**, 1163-1169, doi:10.1104/pp.98.3.1163 (1992).

86 D'Aoust, M. A., Yelle, S. & Nguyen-Quoc, B. Antisense inhibition of tomato fruit sucrose synthase decreases fruit setting and the sucrose unloading capacity of young fruit. *Plant Cell* **11**, 2407-2418, doi:10.1105/tpc.11.12.2407 (1999).

87 Tohge, T., Alseekh, S. & Fernie, A. On the regulation and function of secondary metabolism during fruit development and ripening. *J Exp Bot.* **[Epub ahead of print]** (2014).

88 Alberstein, M., Eisenstein, M. & Abeliovich, H. Removing allosteric feedback inhibition of tomato 4-coumarate:CoA ligase by directed evolution. *Plant J* **69**, 57-69, doi:10.1111/j.1365-313X.2011.04770.x (2011).

89 Niggeweg, R., Michael, A. J. & Martin, C. Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nat Biotechnol* **22**, 746-754 (2004).

90 Teutschbein, J. *et al.* Identification and Localization of a Lipase-like Acyltransferase in Phenylpropanoid Metabolism of Tomato (Solanum lycopersicum). *J Biol Chem* **285**, 38374-38381, doi:10.1074/jbc.M110.171637 (2010).

91 Oneill, S. D., Tong, Y., Sporlein, B., Forkmann, G. & Yoder, J. I. MOLECULAR GENETIC-ANALYSIS OF CHALCONE SYNTHASE IN LYCOPERSICON-ESCULENTUM AND AN ANTHOCYANIN-DEFICIENT MUTANT. *Molecular & General Genetics* **224**, 279-288, doi:10.1007/bf00271562 (1990).

92 Groenenboom, M. *et al.* The Flavonoid Pathway in Tomato Seedlings: Transcript Abundance and the Modeling of Metabolite Dynamics. *Plos One* **8**, doi:10.1371/journal.pone.0068960 (2013).

93 Schmidt, A., Li, C., Shi, F., Jones, A. D. & Pichersky, E. Polymethylated Myricetin in Trichomes of the Wild Tomato Species Solanum habrochaites and Characterization of Trichome-Specific 3 '/5 '- and 7/4 '-Myricetin O-Methyltransferases. *Plant Physiol* **155**, 1999-2009, doi:10.1104/pp.110.169961 (2011).

94 Schmidt, A., Li, C., Jones, A. D. & Pichersky, E. Characterization of a flavonol 3-O-methyltransferase in the trichomes of the wild tomato species Solanum habrochaites. *Planta* **236**, 839-849, doi:10.1007/s00425-012-1676-0 (2012).

95 Bonguebartelsman, M., Oneill, S. D., Tong, Y. S. & Yoder, J. I. CHARACTERIZATION OF THE GENE ENCODING DIHYDROFLAVONOL 4-REDUCTASE IN TOMATO. *Gene* **138**, 153-157, doi:10.1016/0378-1119(94)90799-4 (1994).

96 Olsen, K. M. *et al.* Identification and characterisation of CYP75A31, a new flavonoid 3 ' 5 '-hydroxylase, isolated from Solanum lycopersicum. *BMC plant biology* **10**, doi:10.1186/1471-2229-10-21 (2010).

97 Schreiber, G. *et al.* ANTHOCYANIN1 from Solanum chilense is more efficient in accumulating anthocyanin metabolites than its Solanum lycopersicum counterpart in association with the ANTHOCYANIN FRUIT phenotype of tomato. *Theoretical and Applied Genetics* **124**, 295-307, doi:10.1007/s00122-011-1705-6 (2012).

98 Itkin, M. *et al.* GLYCOALKALOID METABOLISM1 Is Required for Steroidal Alkaloid Glycosylation and Prevention of Phytotoxicity in Tomato. *Plant Cell* **23**, 4507-4525, doi:10.1105/tpc.111.088732 (2011).

99 Itkin, M. *et al.* Biosynthesis of Antinutritional Alkaloids in Solanaceous Crops Is Mediated by Clustered Genes. *Science* **341**, 175-179, doi:10.1126/science.1240230 (2013).

100 Tieman, D. *et al.* Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 8287-8292, doi:10.1073/pnas.0602469103 (2006).

101 Simkin, A. J., Schwartz, S. H., Auldridge, M., Taylor, M. G. & Klee, H. J. The tomato carotenoid cleavage dioxygenase 1 genes contribute to the formation of the flavor volatiles beta-ionone, pseudoionone, and geranylacetone. *Plant J* **40**, 882-892, doi:10.1111/j.1365-313X.2004.02263.x (2004).

102 Longhurst, T., Lee, E., Hinde, R., Brady, C. & Speirs, J. STRUCTURE OF THE TOMATO ADH2 GENE AND ADH2 PSEUDOGENES, AND A STUDY OF ADH2 GENE-EXPRESSION IN FRUIT. *Plant Mol Biol* **26**, 1073-1084, doi:10.1007/bf00040690 (1994).

103 Chen, G. P. *et al.* Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol* **136**, 2641-2651, doi:10.1104/pp.104.041608 (2004).

104 Falara, V. *et al.* The Tomato Terpene Synthase Gene Family. *Plant Physiol* **157**, 770-789, doi:10.1104/pp.111.179648 (2011).

105 Tieman, D. M. *et al.* Identification of loci affecting flavour volatile emissions in tomato fruits. *J Exp Bot* **57**, 887-896, doi:10.1093/jxb/erj074 (2006).

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

107 Klee, H. J. Purple Tomatoes: Longer Lasting, Less Disease, and Better for You. *Curr Biol* **23**, R520-R521, doi:10.1016/j.cub.2013.05.010 (2013).

108 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

109 Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-U194, doi:Doi 10.1038/Nature10158 (2011).

110 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526, doi:10.1371/journal.pone.0016526 (2011).

111 Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS computational biology* **1**, 166-175, doi:10.1371/journal.pcbi.0010022 (2005).

112 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-268, doi:10.1093/nar/gkm286 (2007).

113 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980).

114 Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14**, 860-869, doi:10.1101/gr.1466204 (2004).

115 Jakowitsch, J., Mette, M. F., van Der Winden, J., Matzke, M. A. & Matzke, A. J. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc Natl Acad Sci U S A* **96**, 13241-13246 (1999).

116 Asins, M. J. *et al.* Two closely linked tomato HKT coding genes are positional candidates for the major tomato QTL involved in Na+ /K+ homeostasis. *Plant Cell Environ* **36**, 1171-1191, doi:10.1111/pce.12051 (2013).

117 Galvez, F. J. *et al.* Expression of LeNHX isoforms in response to salt stress in salt sensitive and salt tolerant tomato species. *Plant Physiol Biochem* **51**, 109-115, doi:10.1016/j.plaphy.2011.10.012 (2012).

118 Rodriguez-Rosales, M. P. *et al.* Overexpression of the tomato K+/H+ antiporter LeNHX2 confers salt tolerance by improving potassium compartmentalization. *New Phytol* **179**, 366-377 (2008).

119 Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F. & Covarrubias, A. A. The enigmatic LEA proteins and other hydrophilins. *Plant Physiol* **148**, 6-24, doi:10.1104/pp.108.120725 (2008).

120 Fischer, I., Camus-Kulandaivelu, L., Allal, F. & Stephan, W. Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. *New Phytol* **190**, 1032-1044, doi:10.1111/j.1469-8137.2011.03648.x (2011).

121 Godoy, J. A. *et al.* Expression, tissue distribution and subcellular localization of dehydrin TAS14 in salt-stressed tomato plants. *Plant Mol Biol* **26**, 1921-1934 (1994).

122 Munoz-Mayor, A. *et al.* Overexpression of dehydrin tas14 gene improves the osmotic stress imposed by drought and salinity in tomato. *J Plant Physiol* **169**, 459-468, doi:10.1016/j.jplph.2011.11.018 (2012).

123 Goel, D., Singh, A. K., Yadav, V., Babbar, S. B. & Bansal, K. C. Overexpression of osmotin gene confers tolerance to salt and drought stresses in transgenic tomato (Solanum lycopersicum L.). *Protoplasma* **245**, 133-141, doi:10.1007/s00709-010-0158-0 (2010).

124 Guo, Y., Huang, C., Xie, Y., Song, F. & Zhou, X. A tomato glutaredoxin gene SlGRX1 regulates plant responses to oxidative, drought and salt stresses. *Planta* **232**, 1499-1509, doi:10.1007/s00425-010-1271-1 (2010).

125 Huang, S. *et al.* Genome-wide analysis of WRKY transcription factors in Solanum lycopersicum. *Mol Genet Genomics* **287**, 495-513, doi:10.1007/s00438-012-0696-6 (2012).

126 Huang, W. *et al.* SlNAC1, a stress-related transcription factor, is fine-tuned on both the transcriptional and the post-translational level. *New Phytol* **197**, 1214-1224, doi:10.1111/nph.12096 (2013).

127 Lata, C. & Prasad, M. Role of DREBs in regulation of abiotic stress responses in plants. *J Exp Bot* **62**, 4731-4748, doi:10.1093/jxb/err210 (2011).

128 Li, J. *et al.* Tomato SlDREB gene restricts leaf expansion and internode elongation by downregulating key genes for gibberellin biosynthesis. *J Exp Bot* **63**, 6407-6420, doi:10.1093/jxb/ers295 (2012).

129 Orellana, S. *et al.* The transcription factor SlAREB1 confers drought, salt stress tolerance and regulates biotic and abiotic stress-related genes in tomato. *Plant Cell Environ* **33**, 2191-2208, doi:10.1111/j.1365-3040.2010.02220.x (2010).

130 Puranik, S., Sahu, P. P., Srivastava, P. S. & Prasad, M. NAC proteins: regulation and role in stress tolerance. *Trends Plant Sci* **17**, 369-381, doi:10.1016/j.tplants.2012.02.004 (2012).

131 Loukehaich, R. *et al.* SpUSP, an annexin-interacting universal stress protein, enhances drought tolerance in tomato. *J Exp Bot* **63**, 5593-5606, doi:10.1093/jxb/ers220 (2012).

132 Lu, Y. *et al.* Genomic organization, phylogenetic comparison and expression profiles of annexin gene family in tomato (Solanum lycopersicum). *Gene* **499**, 14-24, doi:10.1016/j.gene.2012.03.026 (2012).

133 Mboup, M., Fischer, I., Lainer, H. & Stephan, W. Trans-species polymorphism and allele-specific expression in the CBF gene family of wild tomatoes. *Mol Biol Evol* **29**, 3641-3652, doi:10.1093/molbev/mss176 (2012).

134 Sinha, A. K., Jaggi, M., Raghuram, B. & Tuteja, N. Mitogen-activated protein kinase signaling in plants under abiotic stress. *Plant Signal Behav* **6**, 196-203 (2011).

135 Stulemeijer, I. J., Stratmann, J. W. & Joosten, M. H. Tomato mitogen-activated protein kinases LeMPK1, LeMPK2, and LeMPK3 are activated during the Cf-4/Avr4-induced hypersensitive response and have distinct phosphorylation specificities. *Plant Physiol* **144**, 1481-1494, doi:10.1104/pp.107.101063 (2007).

136 Mittova, V., Guy, M., Tal, M. & Volokita, M. Response of the cultivated tomato and its wild salt-tolerant relative Lycopersicon pennellii to salt-dependent oxidative stress: increased activities of antioxidant enzymes in root plastids. *Free Radic Res* **36**, 195-202 (2002).

137 Mittova, V., Tal, M., Volokita, M. & Guy, M. Salt stress induces up-regulation of an efficient chloroplast antioxidant system in the salt-tolerant wild tomato species Lycopersicon pennellii but not in the cultivated species. *Physiol Plant* **115**, 393-400 (2002).

138 Scippa, G. S. *et al.* The histone-like protein H1-S and the response of tomato leaves to water deficit. *J Exp Bot* **55**, 99-109, doi:10.1093/jxb/erh022 (2004).

139 Langenkamper, G. *et al.* Accumulation of plastid lipid-associated proteins (fibrillin/CDSP34) upon oxidative stress, ageing and biotic stress in Solanaceae and in response to drought in other species. *J Exp Bot* **52**, 1545-1554 (2001).

140 Gong, P. *et al.* Transcriptional profiles of drought-responsive genes in modulating transcription signal transduction, and biochemical pathways in tomato. *J Exp Bot* **61**, 3563-3575, doi:10.1093/jxb/erq167 (2010).

141 Ouyang, B. *et al.* Identification of early salt stress response genes in tomato root by suppression subtractive hybridization and microarray analysis. *J Exp Bot* **58**, 507-520, doi:10.1093/jxb/erl258 (2007).

142 Sun, W. *et al.* Comparative transcriptomic profiling of a salt-tolerant wild tomato species and a salt-sensitive tomato cultivar. *Plant Cell Physiol* **51**, 997-1006, doi:10.1093/pcp/pcq056 (2010).

143 Claussen, W. Proline as a measure of stress in tomato plants. *Plant Science* **168**, 241-248, doi:http://dx.doi.org/10.1016/j.plantsci.2004.07.039 (2005).

144 Tal, M., Katz, A., Heikin, H. & Dehan, K. SALT TOLERANCE IN THE WILD RELATIVES OF THE CULTIVATED TOMATO: PROLINE ACCUMULATION IN LYCOPERSICON ESCULENTUM MILL., L. PERUVIANUM MILL. AND SOLANUM PENNELLI COR. TREATED WITH NaCI AND POLYETHYLENE GLYCOLE. *New Phytologist* **82**, 349-355, doi:10.1111/j.1469-8137.1979.tb02660.x (1979).

145 Santa-Cruz, A., Estañ, M. T., Rus, A., Bolarin, M. C. & Acosta, M. Effects of NaCl and mannitol iso-osmotic stresses on the free polyamine levels in leaf discs of tomato species differing in salt tolerance. *Journal of Plant Physiology* **151**, 754-758, doi:http://dx.doi.org/10.1016/S0176-1617(97)80074-0 (1997).

146 Santa-Cruz, A., Perez-Alfocea, F., Caro, M. & Acosta, M. Polyamines as short-term salt tolerance traits in tomato. *Plant Science* **138**, 9-16, doi:http://dx.doi.org/10.1016/S0168-9452(98)00143-5 (1998).

147 Yeats, T. H. & Rose, J. K. C. The Formation and Function of Plant Cuticles. *Plant Physiol* **163**, 5-20, doi:DOI 10.1104/pp.113.222737 (2013).

148 Eshed, Y., Abu-Abied, M., Saranga, Y. & Zamir, D. Lycopersicon esculentum lines containing small overlapping introgressions from L. pennellii. *Theoretical and Applied Genetics* **83**, 1027-1034, doi:10.1007/BF00232968 (1992).

149 Frary, A. *et al.* Salt tolerance in Solanum pennellii: antioxidant response and related QTL. *BMC plant biology* **10**, 58, doi:10.1186/1471-2229-10-58 (2010).

150 Frary, A., Keles, D., Pinar, H., Gol, D. & Doganlar, S. NaCl tolerance in Lycopersicon pennellii introgression lines: QTL related to physiological responses. *Biol Plantarum* **55**, 461-468, doi:DOI 10.1007/s10535-011-0111-x (2011).

151 Gur, A. *et al.* Yield quantitative trait loci from wild tomato are predominately expressed by the shoot. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **122**, 405-420, doi:10.1007/s00122-010-1456-9 (2011).

152 Lei，L., Y. S., and Jun-ming, L. Mapping of QTLs for Drought Tolerance During Seedling Stage Using Introgression Line Populations in Tomato. *ACTA HORTICULTURAE SINICA* **38**, 1921-1928 (2011).

153 Li, J. M. *et al.* Seedling salt tolerance in tomato. *Euphytica* **178**, 403-414, doi:DOI 10.1007/s10681-010-0321-x (2011).

154 Uozumi, A. *et al.* Tolerance to salt stress and blossom-end rot in an introgression line, IL8-3, of tomato. *Scientia Horticulturae* **138**, 1-6, doi:10.1016/j.scienta.2012.01.036 (2012).

155 Filippis, I., Lopez-Cobollo, R., Abbott, J., Butcher, S. & Bishop, G. J. Using a periclinal chimera to unravel layer-specific gene expression in plants. *The Plant Journal* **75**, 1039-1049 (2013).

156 Labate, J. A. & Robertson, L. D. Evidence of cryptic introgression in tomato (Solanum lycopersicum L.) based on wild tomato species alleles. *BMC plant biology* **12**, 133, doi:10.1186/1471-2229-12-133 (2012).

157 Li-Beisson, Y. *et al.* Acyl-lipid metabolism. *The Arabidopsis book / American Society of Plant Biologists* **11**, e0161, doi:10.1199/tab.0161 (2013).

158 Pardo, J. M., Cubero, B., Leidi, E. O. & Quintero, F. J. Alkali cation exchangers: roles in cellular homeostasis and stress tolerance. *J Exp Bot* **57**, 1181-1199, doi:10.1093/jxb/erj114 (2006).

159 Garcia-Rios, M. *et al.* Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase. *Proc Natl Acad Sci U S A* **94**, 8249-8254 (1997).

160 Sakamoto, A. & Murata, N. Genetic engineering of glycinebetaine synthesis in plants: current status and implications for enhancement of stress tolerance. *J Exp Bot* **51**, 81-88 (2000).

161 Huda, K. M., Banu, M. S., Tuteja, R. & Tuteja, N. Global calcium transducer P-type Ca(2)(+)-ATPases open new avenues for agriculture by regulating stress signalling. *J Exp Bot* **64**, 3099-3109, doi:10.1093/jxb/ert182 (2013).

162 Chary, S. N., Hicks, G. R., Choi, Y. G., Carter, D. & Raikhel, N. V. Trehalose-6-phosphate synthase/phosphatase regulates cell shape and plant architecture in Arabidopsis. *Plant Physiol* **146**, 97-107, doi:10.1104/pp.107.107441 (2008).

163 Sjogren, L. L., Stanne, T. M., Zheng, B., Sutinen, S. & Clarke, A. K. Structural and functional insights into the chloroplast ATP-dependent Clp protease in Arabidopsis. *Plant Cell* **18**, 2635-2649, doi:10.1105/tpc.106.044594 (2006).

164 Sharma, S. & Verslues, P. E. Mechanisms independent of abscisic acid (ABA) or proline feedback have a predominant role in transcriptional regulation of proline metabolism during low water potential and stress recovery. *Plant Cell Environ* **33**, 1838-1851, doi:10.1111/j.1365-3040.2010.02188.x (2010).

165 Bharti, K. *et al.* Isolation and characterization of HsfA3, a new heat stress transcription factor of Lycopersicon peruvianum. *Plant J* **22**, 355-365 (2000).

166 Palmieri, F., Pierri, C. L., De Grassi, A., Nunes-Nesi, A. & Fernie, A. R. Evolution, structure and function of mitochondrial carriers: a review with new insights. *Plant J* **66**, 161-181, doi:10.1111/j.1365-313X.2011.04516.x (2011).

167 Bird, D. M. & Wilson, M. A. DNA sequence and expression analysis of root-knot nematode-elicited giant cell transcripts. *Molecular plant-microbe interactions : MPMI* **7**, 419-424 (1994).