SUPPLEMENTARY INFORMATION
doi:10.1038/nature11119

The tomato genome sequence provides insights into fleshy fruit evolution

The Tomato Genome Consortium (TGC)

Supplementary Information

CONTENTS

SEQUENCING, ASSEMBLY AND MAPPING OF *S. LYCOPERSICUM*

1.1 Plant materials

'Heinz 1706' seeds were provided by Heinz Corporation (Pittsburgh, PA). The pedigree of 'Heinz 1706' has been described[29]. Grandparents include 'Fireball', 'Roma' and 'VR Moscow'; pedigree records indicate that 'VR Moscow' is the likely maternal (cytoplasm) contributor. The fourth grandparent is a selection from a recombination of 'Burgess Crackproof' and an Eastern States line ('ES 25'), selected from a variable population developed by O. Pearson in the 1950's. The ES population is reported to have a background of 'Andrus' 2153', 'Firesteel', Yeager's high vitamin line and Hanna's *S. pimpinellifolium* hybrid 17-5. Seeds for analyses were greenhouse grown for three weeks in flats that were then transferred to growth chambers with no light for 72 hours to promote starch degradation. Fresh meristematic expanding leaves were harvested and used either directly for isolation of nuclei (BACs) or frozen in liquid nitrogen and stored at -80°C prior to DNA extraction for fosmid library construction and 454 sequencing library preparation.

1.2 High Molecular Weight DNA preparation and BAC library construction

Extraction of high molecular weight (HMW) genomic DNA from leaf nuclei was performed as described previously[30]. Genomic DNA was partially digested with either EcoRI or MboI restriction enzyme followed by two rounds of size selection from agarose gels following separation by pulsed field gel electrophoresis (PFGE). In the first size selections, the pulsed field gel was run with a 70 second pulse at 170 V and 11°C for 18 hours. DNA ranging from 200 kb to 400 kb was excised from the gel and used in the next size selection. The second size selection gel was run using a 4 second pulse at 150 V and 11°C for 16 hours and the compressed band representing DNA fragments greater than 100 kb was excised. Prior to ligation the final size selected DNA was released from agarose by electroelution. The agarose gel slice containing the > 100 kb DNA was fragmented with a razor blade and the resulting pieces placed in dialysis bags (Gibco-BRL). Electroelution was carried out for 2 hours at 200 V with a 90 second pulse at 11°C. Eluted DNA was quantified by gel electrophoresis against lambda phage standards (Sigma) and a molar ratio of approximately 3:1 vector: insert was used for ligation. EcoRI-digested and dephosphorylated Copy Control pCC1BAC vector (derived from pBelobac 11, Epicentre, Madison, WI) was used for EcoRI library construction while pECBACI (pBELO BACII variant with BamHI cloning site) was used for the MboI library. Transformations were performed by electroporation using Gene Hogs electrocompetent cells

(Invitrogen). The EcoRI and MboI libraries consist of 75,264 and 52,992 ordered clones, respectively. In addition, a sheared genomic DNA library (80,256 clones) was constructed by Lucigen Corp. (Middletown, WI) in the BstXI site of vector pSMART-BAC. A HindIII library (129,024 clones) was described previously[31]. The fosmid library was made in vector pCC1FOS Copy Control (Epicentre, WI) cloned into the Eco72 site under contract by Warwick Plant Genomic Libraries Ltd. with a titre of 2.9 million unamplified clones. A total of 307,200 unamplified phages were converted to bacterial colonies and individually picked into 384-well microtiter plates (800 total) using a Q-bot robotic workstation. The EcoRI, MboI, HindIII random sheared and fosmid libraries are named SL_EcoRI, SL_MboI, SL_HindIII (formerly LeHBa), SLs and SL_Fos, respectively. All clones from all libraries described are available without restriction individually as bacterial stabs or as complete 384-well arrayed libraries from the Boyce Thompson Institute for Plant Research, Cornell University and can be requested via an ordering form available at the Tomato Functional Genomics Database (http://ted.bti.cornell.edu).

1.3 Sanger sequencing of Selected BAC Mixture

To enrich for gene-rich regions of the genome, two sets of BAC clone pools were generated using the BAC end sequences available at the SOL Genomics Network (SGN) website, and denoted as SBM (Selected BAC clone Mixture). The first set (SBM-I) comprised 20,000 BAC clones of which neither end sequence showed sequence similarity with any of the tomato repetitive sequences registered in the tomato genome repeat dataset provided by SGN (ftp://ftp.sgn.cornell.edu/tomato_genome/repeats/curr/). The set contained 10,000 clones from the HindIII library and 5,000 clones from each of the MboI and EcoRI libraries. The second set (SBM-II) was selected to include clones with a repetitive sequence at one end and a unique sequence at the other. A total of 10,800 clones (5,400 clones from the HindIII library, and 2,700 clones from each of the MboI and EcoRI libraries) were pooled. Shotgun libraries with average insert sizes of 2.5 kb were generated using pBluescript SK- as the cloning vector, and these were used to transform *E. coli* ElectroTen-Blue (Agilent Technologies, Santa Clara, CA). The shotgun clones were propagated in microtiter plates and the plasmid DNA was amplified for the sequencing reaction using TempliPhi (GE healthcare, UK). Sequencing was performed using a cycle sequencing kit (Big Dye-terminator Cycle Sequencing kit, Applied Biosystems, USA) with DNA sequencers ABI 3730XL Genome Analyzer (Applied Biosystems, USA) or Denova (Shimadzu, Japan) according to the protocol recommended by the manufacturer. The total number of sequence reads produced was

4,248,000 (2,916,000 from SBM-I and 1,332,000 from SBM-II). Reads with low quality were masked using TRIM3 and vector sequences were masked using Cross_match. In total, 4,039,383 reads passed the quality trimming (`the 'raw' column in **Supplementary Table 15**).

1.4 Sanger sequencing of clone ends

The BAC clones were provided to a sequencing company (SeqWright Inc, Houston, TX) as glycerol stocks arrayed in 384-well microtiter plates. SeqWright extracted the BAC DNA in 96-well format using a commercially available Qiagen kit and by following the manufacturer's instructions. DNA quality was verified by agarose gel Quality Control (QC) analysis of a subset of the extracted BACs from each plate of clones. Plates containing extracted BAC DNA that passed the QC analysis were BigDye Terminator v3.1 (Applied Biosystems) cycle sequenced with forward and reverse primers in 384-well format following the manufacturer's instructions. Sequencing reactions were assayed on Applied Biosystems 3730XL DNA Analyzers. Chromatogram, sequence and Phred quality data were supplied for all sequencing reactions.

The fosmid clones were cultured overnight in 384-well microtiter plates after induction with 0.01% L-arabinose (Sigma-Aldrich). DNA extraction was performed in 384-well format with a protocol based on alkaline lysis and subsequent recovery with filter plates (Millipore). The sequencing reaction was performed in nanoliter format (final volume 0.8 μl) by means of the Microlab Star nanopipettor (Hamilton), using the Big Dye-terminator Cycle Sequencing kit (Applied Biosystems) and a 3730XL sequencer (Applied Biosystems) according to the manufacturer's protocol. The sequencing yielded 211,359 reads that were quality checked with Phred, allowing the identification and removal of bad quality sequences. The remaining reads (151,301, corresponding to the column labelled 'raw' in **Supplementary Table 16**) were screened for low-quality regions (quality cut-off <20) and vector sequence, which were trimmed before further analysis.

The insert size distribution and the corresponding average size of the BAC and fosmid libraries were determined through BLASTN alignment of the end sequences to the *de novo* genome assembly (**Supplementary Fig. 18** and **Supplementary Table 16**).

1.5 454 Sequencing

DNA sample preparation for sequencing on the 454/Roche GS FLX were performed as described by the manufacturer[32] with modifications to improve the overall yields that included replacing the

Qiagen MinElute centrifuge columns with an Agencourt AMPure SPRI bead-based purification, removing the manipulations to enrich DNA molecules that contain a single A and B adapter ligated on each end of the fragment, and to eliminate the steps that resulted in generating a single stranded DNA library[33]. Briefly, this procedure entailed shearing the DNA via nebulisation and subsequent end repair, as described[34], followed by ligation of adapter sequences and a second round of end repair to yield a blunt ended DNA library that then was quantified and diluted prior to amplification via emPCR[32]. For BAC-based sequencing, the BAC clones were tagged individually using 454-recommended MID tags and pooled in groups of twelve for emPCR amplification and subsequent sequencing on one-quarter plate of the 454/Roche GS FLX using the Titanium chemistry. For whole genome shotgun sequencing, a total of five distinct shotgun libraries and five 3-kb, ten 8-kb and six 20-kb mate-pair libraries were constructed. Amplification and sequencing of these libraries was performed using GS FLX Titanium Sequencing Kits and 454 Genome Sequencer FLX Instruments following the manufacturer's protocols (Roche Applied Science, Mannheim, Germany). PicoTiter Plates were loaded with beads from any combination of one, two, four or eight different libraries, separated physically into distinct regions by use of appropriate gaskets. In total, 28.4 Gb of sequence data was generated, consisting of 14.4 Gb of shotgun reads, 7.1 Gb of 3-kb mate-pair reads, 3.9 Gb of 8-kb mate-pair reads, and 3.0 Gb of 20-kb mate-pair reads (**Supplementary Table 17**).

1.6 SOLiD Sequencing

The *S. lycopersicum* 'Heinz 1706' genome was also analysed on a SOLiD™ sequencer (Applied Biosystems, Foster City, CA) following the manufacturer's protocol. Four different mate-pair libraries with estimated fragment length of 1, 4, 7 and 8 kb were constructed. The starting amount of genomic DNA was about 50 µg for the 8-kb library and about 30 µg for the others. Every sample was then sequenced on a whole slide using the mate-pair procedure. The 7-kb library was treated as a fragment library because a technical problem did not allow sequencing from the R3 tag. In total, 3.6 billion reads and 133 Gb of sequence data were generated **(Supplementary Table 18)**.

1.7 Illumina sequencing

DNA of *S. lycopersicum* 'Heinz 1706' was prepared from 4-week old leaf tissue by enriching for nuclei[30] followed by a standard CTAB extraction[35]. Two paired-end libraries of ~450 and 500 bp

and four mate-pair libraries of 2, 3, 4 and 5 kb were prepared using Illumina's paired-end and mate-pair kits respectively (Illumina, San Diego, CA). 5 µg of DNA was sheared with a Covaris S2 ultrasonicator (Covaris Inc. Woburn, MA) for the paired-end libraries and 10 µg of DNA was sheared with a Hydroshear (Genomic Solutions, Ann Arbor, MI) for the mate-pair libraries. The two paired-end libraries were run as 2X90 bp runs on 16 flow cell lanes of an Illumina Genome Analyzer IIx (GAIIx) sequencer, yielding approximately 64 Gb of sequence data. The four mate-pair libraries were each run on two lanes of a flow cell as 2X54 bp runs, yielding approximately 3 Gb per run (**Supplementary Table 19**).

1.8 Construction of a Whole Genome Profiling (WGP) physical map

Subsets of the four 'Heinz 1706' BAC libraries, comprising a total of 92,160 BAC clones and an approximate 11X genome coverage, were used for construction of a WGP map:

- a HindIII library consisting of 15,360 clones, with an estimated average insert size of 105 kb, representing approximately 2X genome coverage
- an MboI library, consisting of 15,360 clones, with an estimated average insert size of 121 kb, representing approximately 2X genome coverage
- an EcoRI library, consisting of 15,360 clones, with an estimated average insert size of 103 kb, representing approximately 2X genome coverage, and
- a random sheared library, consisting of 46,080 clones with an estimated average insert size of 100 kb, representing approximately 5X genome coverage

For WGP, BACs were pooled in a two-dimensional (2D) format prior to DNA isolation. Specifically, the 92,160 BACs, stored in 240 384-well plates (termed Superpools; SP) were pooled 2D by pooling each row (24 BACs) and each column (16 BACs), yielding 40 pools per SP. A total of 9,600 (240 x 40) BAC pools were prepared and subjected to isolation of high-concentration DNA (Amplicon Express Inc., Pullman, WA). WGP sample preparation was performed as described[36]. AFLP templates[37] were prepared from 20 nanograms of pooled BAC DNA by digestion using 5 units EcoRI and 2 units MseI for at least 1 hour at 37°C. Next, adapter ligation using a universal P7 MseI adapter and a sample-specific tagged EcoRI-P5 adapter was carried out for 3 hours at 37°C. PCR was performed in 20 µl and contained 5 µl 10-fold diluted restriction ligation mixture, 30 ng Illumina P5 primer (5'-AATGATACGGCGACCACCG-3'), 30 ng Illumina P7 primer (5'-CAAGCAGAAGACGGCATACGA-3'), 0.2 mM dNTPs, 0.4 U AmpliTaq® (Applied Biosystems)

and 1x AmpliTaq buffer. PCR was performed with the following profile: 2 minutes 72°C followed by 22 cycles of 30 sec 94°C, 60 sec 56°C, 60 sec 72°C, and finally cooling down to 4°C. Next, equal amounts of SP samples were purified using the QIAquick PCR Purification Kit (Qiagen). All 320 sample-specific EcoRI-P5 adapters included a unique five- or six-nucleotide sample identification tag adjacent to the EcoRI restriction site overhang for identification of individual BACs by deconvolution. A total of 35 lanes of 36-cycle Illumina Genome Analyzer II (GAII) sequencing were performed, divided over six runs. Each run was done using a flow cell with eight lanes of physically separated samples such that the same set of sample tags could be used for each lane. GAII runs were performed comprising seven lanes with tomato WGP samples at 5.5 pM concentration, each covering six or eight SPs represented in 240 or 320 row- or column pools, respectively. A total of 240 SPs were sequenced, equalling 240 x 384 = 92,160 BACs. The Illumina pipeline software (GA_pipeline_v0.3) was used to analyse images into sequence reads of 36 nt length. An additional quality filter was applied to select only those reads with all base calls being at least 0 on the Illumina GA scale. Sequence reads were split into 3 parts to enable assignment of unique tags to pools and to allow for consecutive deconvolution into individual BACs: the first 5 (or 6) nt represented the sample (i.e. BAC pool) identification tag; the next 6 nt matched the EcoRI restriction site of the adapter and the remaining 20 nt defined the WGP tag. The assignment of unique WGP tags to individual BACs was based on the following criteria: 1) a specific WGP tag must occur in two pools to indicate its unique position on the plate: one column and one row pool with both being represented by at least two reads, and; 2) if WGP tags are inadvertently observed in a third or fourth pool, the number of reads in these other pools must be less than a tenth of those in the smallest true pool. WGP tags not matching these criteria were discarded. A script was used to recognize and trim the sample identification tags, the restriction site part of the sequence reads and to perform the deconvolution. Unique WGP tags were defined by grouping them in 100% identical read sets. The output of this procedure consisted of a list of all WGP tags, the corresponding number of reads, and the identification number of the BAC to which they were assigned. WGP tags matching *E. coli* or chloroplast sequences and WGP tags occurring on just a single BAC were then removed. BACs were grouped into contigs using the FPC program[38] that originally was developed for analysing BAC fingerprint data: restriction fragments determined by their length. The WGP tags were adapted for use in FPC by converting them into numbers, yielding pseudo restriction-fragment sizes for which the software was designed. As the WGP tags are uniquely defined by their sequence composition, FPC could be used at the highest stringency setting of tolerance (value = 0). Different cut-off values were tested, specifying the threshold on the probability of BAC coincidence, i.e. the

likelihood that different BACs share overlapping WGP tags, while not originating from the same genomic region. The output of FPC consisted of a list with contigs and the corresponding order of BACs within each contig. The genome coverage, average contig size and N50 contig size in million basepairs were estimated by multiplying FPC band units by the average distance between two WGP tags of 3,433 basepairs. The latter was estimated by dividing the average BAC insert size of 114 kb by the average number of WGP tags of 33.2 (see below). A total of 326.9 million 36 nt GA II reads were produced to construct the WGP map. Of these, a subset of 136.7 million (42%) were deconvolutable to individual BACs, representing 261,913 unique WGP tags each consisting of the 6 nt EcoRI recognition sequence and 20 nt flanking sequence). A total of 66,084 BACs (72% of 92,160) contained at least one WGP tag, of which 37,912 were BACs from the libraries generated by restricted DNA and 28,172 were BACs from the random sheared library (**Supplementary Table 20**). These BACs served as input for contig building using FPC. The average number of WGP tags per BAC was 33.2 (35.0 for the enzyme libraries and 30.7 for the random sheared library) and the average number of reads per WGP tag was 50.0 (46 and 54 for the enzyme libraries and random sheared library, respectively). A sequence-based physical BAC map was assembled using an improved version of the FPC software (Keygene N.V.), capable of processing sequence-based BAC fingerprint (WGP) data instead of fragment mobility information as used in the original FPC[38]. WGP data were used as input in the FPC map assembly. The assembly was performed using an e^{-30} stringency level, resulting in a high stringency map. The results of the assembly in terms of the numbers of BAC clones incorporated in contigs, broken down by BAC library, are provided in **Supplementary Table 20**. Summary FPC results for the WGP map are provided in **Supplementary Table 21**.

1.9 Construction of a SNaPshot physical map

Clones from the four BAC libraries (**Supplementary Table 22**) were fingerprinted using the SNaPshot method[39-41]. Briefly, clones were digested with BamHI, EcoRI, XbaI, XhoI and HaeIII. Protruding termini from the first four enzymes were filled with fluorescent tagged ddNTP's. Restriction fragments were separated by capillary electrophoresis using an ABI 3730XL Genome Analyzer. Signals from labelled fragments were processed to filter data for clone contamination, small inserts or chimeras and generate size files that comprise the input data for the mapping software.

Size files were assembled using FPC v.9.3[42] at fix tolerance and cutoff values of 4 and $1*e^{-50}$, respectively. The "CpM" function was activated for the assembly, taking advantage of the inclusion of 4,123 markers (overgo and electronic markers, see below; **Supplementary Table 23**) to the clones included in the project. The initial assembly was subjected to removal of questionable clones in three incremental steps each at a higher stringency (cutoff $1*e^{-53}$, $1*e^{-56}$ and $1*e^{-59}$). Resulting contigs were subjected to end-end analysis and automatically merged at a cutoff $1*e^{-18}$. The same stringency was used for evaluation of single clones, which were also incorporated into the resulting contig population.

BAC end sequences (BES) from clones included in the map from the MboI, EcoRI and HindIIII BAC libraries were downloaded from SGN (http://solgenomics.net/) (**Supplementary Table 24**). These sequences were masked with RepeatMasker v3.2.7[43], using a custom redundant database which was mainly populated with Solanaceae repeat sequences obtained from SGN (http://solgenomics.net), Plant Repeats Database (http://plantrepeats.plantbiology.msu.edu/composition.html) and GIRI Repbase (Viridiplantae, v.13.01)[44] (http://www.girinst.org/repbase/index.html). BES were used to link physical mapped clones to tomato unigenes in the SNG database (ftp://ftp.solgenomics.net/unigene_builds), individually sequenced BACs (ftp://ftp.solgenomics.net/tomato_genome/bacs/), sequenced markers from the integrated EXPEN-Kazusa genetic map[45,46] (ftp://ftp.solgenomics.net/maps_and_markers/Tomato/ and http://www.kazusa.or.jp/tomato/) and to the sequence scaffolds and pseudomolecules (all available at the SGN webpage). The BES alignments of the mapped clones to the tomato pseudomolecules were graphically depicted using the SyMAP software[47] facilitating the editing of the physical map.

The SNaPshot physical map included 82,777 clones of which 86% were assembled into 3,534 contigs, with 10-fold genome coverage (**Supplementary Table 22** and **Supplementary Fig. 19**). The map and all underlying data are available via SGN (http://solgenomics.net/cview/index.pl). The physical contig alignments to tomato pseudomolecules are shown in **Supplementary Fig. 20**, indicating that anchored contigs cover 95% of the tomato genome. Detailed statistics of anchored contigs and clones to each chromosome are provided in **Supplementary Fig. 21** and **Supplementary Tables 25-26**, which are also available at the SGN website.

1.10 Construction of a high-density genetic map

An integrated genetic linkage map of tomato was previously constructed using the Tomato-EXPEN 2000 mapping population[45]. In total, 60 out of the 2,035 markers yielded problematic and/or inconsistent sequence data, and these were removed from further analyses. The location of an additional 108 Conserved Ortholog Set II (COSII) markers were obtained from Dr. Silvana Grandillo. The marker sequences are available from the Tomato marker database at the Kazusa DNA research institute (http://marker.kazusa.or.jp/Tomato/) and from SGN (http://solgenomics.net).

1.11 Anchoring BACs to the genetic map by OVERGO screening

Overgo design and screening was performed using ^{32}P-labeled probes derived from 20 nt primer pairs with 8 nt complementary sequences and methods and procedures as described on the NCBI Technologies website (http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechOvergo.shtml). A total of 1,536 probes derived from marker sequences spanning the 12 tomato chromosomes were used to screen a total of 128,560 BACs resulting in 7,972 high quality probe-BAC associations. A summary of these results including markers, primers and hybridizing BACs can be found on the SGN website (http://solgenomics.net/maps/physical/overgo_stats.pl).

1.12 Fluorescence in situ hybridisation (FISH)

FISH localisation of BACs on tomato chromosomes was performed in two different laboratories (Stack and de Jong labs). Although the techniques used are similar, there are enough differences that they are presented separately here. Even so, the two techniques yielded similar results when the same BACs were localized, and the localisations are shown together on diagrams of tomato pachytene chromosomes (**Fig. 1**, **Supplementary Figures 1, 15, 22**).

1.13 BAC-FISH on hypotonically spread pachytene chromosomes

Both the hypotonic chromosome spreading and FISH procedures have been described in detail elsewhere[48-51]. Briefly, living cherry tomato (accession LA444) primary microsporocytes at the pachytene stage of meiosis were extruded from anthers into a hypertonic aqueous sugar-salt medium, and the cells were then treated with cytohelicase to produce protoplasts. Protoplasts were burst on glass slides using a hypotonic detergent solution. This spreads pachytene chromosomes without significantly distorting their lengths or arm ratios. We refer to this as synaptonemal

complex (SC) spreading because it disperses chromatin enough to reveal SCs and kinetochores (centromeres). The spreads were immediately fixed lightly with a fine spray of 4% paraformaldehyde, air dried, washed briefly in water, air dried and stored at -80°C.

BAC DNA from the HindIII or MboI BAC libraries was isolated by standard protocols (AquaPlasmid, MultiTarget Pharmaceuticals, Salt Lake City, UT) and labelled with biotin, digoxigenin or dinitrophenol using nick translation according to the manufacturer's instructions (Roche Applied Science). Chromosome spreads on a glass slide were digested sequentially with RNase A and pepsin before hybridisation overnight to 50–1,000 ng of one or more labelled BAC probes plus 1–5 µg of Cot-100 tomato DNA[52,53] to suppress hybridisation of the probe to repeated sequences. Spreads were hybridized to DNA labelled with biotin, digoxigenin or dinitrophenol (DNP) and then incubated with: 1) mouse anti-biotin (Roche), biotinylated donkey anti-mouse (Jackson), and streptavidin-FITC (Roche); 2) sheep anti-digoxigenin (Roche) and donkey anti-sheep-TRITC (Jackson), or; 3) rat anti-DNP (Invitrogen) and donkey anti-rat-Dylight 649 (Jackson), respectively. After washing, slides were dehydrated through an ethanol series and air dried. Cover glasses were mounted using Vectashield (Vector Laboratories) containing 5 µg/ml 4,6-diamidino-2-phenylindole (DAPI). Microscopy was performed with one of three microscopes (Olympus Provis, Leica DM 5000B, or Leica DM5500B), all equipped for phase contrast illumination and fluorescence microscopy using DAPI, FITC, TRITC, and Cy5 filter cubes with zero pixel shift. Cooled Optronics and Hamamatsu monochrome cameras were used for photography. Images were captured and artificially coloured using either PictureFrame or IP Lab software (**Supplementary Fig. 23a**). Measurements of SCs and FISH signals were made using MicroMeasure[54] (www.biology.colostate.edu/MicroMeasure). The position of each localized BAC was measured as a percentage of the chromosome length from the end of the short arm. The positions of all of the BACs localized on SC spreads are available online (http://solgenomics.net/cview/map.pl?map_id=13). Each BAC localisation is based on measurements from at least ten SC spreads (with a few exceptions). The percentage positions of BACs were then converted to absolute (µm) positions based on the average length of each SC[55]. The estimated locations of distal euchromatin and pericentric heterochromatin are based on the data from Sherman and Stack[55] and on the observed distribution of recombination nodules[56]. Each tomato SC has a consistent relative length and arm ratio that can be used to identify the SC and that serve as the basis for the accurate location of BACs by FISH. However, two pairs of SCs (7 and 9; 5 and 12) are morphologically indistinguishable and require localisations of marker BACs to tell them

apart reliably. Also, SCs 5 and 12 have virtually equal "long" and "short" arms, so for these SCs, BAC markers were used to distinguish long arms from short arms.

1.14 BAC-FISH on stirred spreads of pachytene chromosomes

Young flower buds of *S. lycopersicum* 'Heinz 1706' plants were fixed in freshly prepared acetic ethanol (1 part glacial acetic acid to 3 parts ethanol). The next day the buds were transferred to 70% ethanol for storage at 4°C. Anthers were selected for cell wall digestion, and pachytene chromosomes were spread on glass slides by stirring according to Szinay *et al.*[57]. Isolation of Cot-100 DNA was carried out as described by Zwick *et al.*[53], without the phenol step during DNA extraction. BACs were isolated using High Pure Plasmid Isolation Kits (Roche 11754785001). BACs and repeat sequences were labelled with biotin and digoxigenin by nick translation following the manufacturer's instructions (http://www.roche.com). Repeat sequences were also directly labelled with Cy3-dUTP (Amersham, http://www5.amershambiosciences.com), Cy3.5-dCTP (Amersham) and Diethylaminocoumarin-5-dUTP (DEAC) (Perkin Elmer, http://www.perkinelmer.com). During hybridisation, Cot-100 DNA[58] was used for blocking repeats in the BACs as described[57,59]. After hybridisation, digoxigenin-labelled probes were amplified with anti-digoxigenin-FITC and anti-sheep-FITC. Biotin-labelled probes were amplified three times with Streptavidin-Cy5 and biotinylated-anti-streptavidin. Microscopy and image processing were performed as described[57]. Chromosome straightening was performed using the 'straighten-curved-objects' plug-in of Image J[60] (**Fig. 1**, **Supplementary Figures 1, 23b**).

1.15 Reproducibility of FISH and integration of results

Reproducibility of FISH was assessed by comparing results from independent hybridisations for each of the two different types of chromosome preparations. For both methods the order of BACs on chromosomes from different sets was consistent and reproducible **(Supplementary Fig. 24)**. To integrate BAC localisation data from stirred spreads (from the de Jong lab) with BAC localisations on SC spreads (from the Stack lab), common markers such as chromomeres (if present), telomeres, centromeres, and any BACs localized in both labs were used to divide each chromosome and each SC into distinct segments. At minimum, each chromosome and SC had two segments (telomere-centromere for short arm; centromere-telomere for long arm). In practice, most chromosomes and SCs had several BAC markers in common that were also used for alignment. BACs localized on

SCs were marked with horizontal red lines in **Fig. 1 and Supplementary Fig. 1**. Stirred chromosome spreads are accurate for determining the order of BACs along a chromosome, but the chromosomes are susceptible to variable stretching along their lengths during the spreading procedure (**Supplementary Fig. 24**). Therefore, the location of BACs on stirred spreads were interpolated and extrapolated to positions on corresponding SCs. The closer these BACs are to common markers on stirred chromosomes and SCs, the more accurate their placement. The locations of BACs on stirred spreads were integrated with the map of BACs on SCs in the SC idiogram (http://solgenomics.net/cview/).

1.16 BAC-FISH to determine order and orientation of scaffolds and gap sizes between scaffolds on DNA pseudomolecules/chromosomes

The order of scaffolds on pseudomolecules/chromosomes refers to their order from the end of the short arm to the end of the long arm. Orientation of a scaffold refers to its head-tail (H-T) orientation with its head directed toward the end of the short arm and its tail directed toward the end of the long arm. Order and orientation of scaffolds was first determined based on the Kazusa tomato molecular linkage map (**Fig. 1** and **Supplementary Fig. 1**). Scaffolds that contain one or more linkage markers can be ordered relative to one another (with unsequenced gaps between them); those with only one linkage marker cannot be oriented. Scaffolds with no linkage markers cannot be ordered or oriented and were assigned to chromosome *0*. BAC-FISH is an independent means of ordering and orienting scaffolds and determining the sizes of gaps between them. If the position of one BAC in a scaffold can be determined by FISH, the position of that scaffold on the chromosome and its order relative to other scaffolds that were also located by BAC-FISH can be determined. If BACs at both ends of a scaffold are localized, the position (order) and orientation of that scaffold is determined. Generally, the two methods for determining the order and orientation of scaffolds give the same results in euchromatin (where most crossing over occurs). However, in heterochromatin where there is little crossing over, the linkage method is not as reliable as the BAC-FISH method, as long as the distance between the two ends of the scaffold is large enough to permit resolution of the two BAC locations (≥100 kb in euchromatin and ≥500 kb in heterochromatin).

In addition, BAC-FISH was used to estimate the size of unsequenced gaps on a pseudomolecule. For this, the two BACs on either side of a gap (*i.e.* the terminal BACs of adjacent scaffolds) were localized (**Supplementary Fig. 22**, **Supplementary Table 13**). The distance between the two BACs was measured on ten or more SC spreads, and an average separation distance and standard

deviation were calculated. An adjustment was made for the average length of the SC, and depending on whether the gap is in euchromatin, heterochromatin, or centromere/kinetochore, the average distance between the BACs was multiplied by the DNA per unit length of SC (pachytene chromosome) in euchromatin (~1.54 Mb/µm), heterochromatin (~ 9.22 Mb/µm), or centromere/kinetochore (3.6 Mb/µm). These values for DNA per unit length of SC in euchromatin and heterochromatin were originally based on densitometry and area measurements of Feulgen stained tomato pachytene chromosomes[61], and more recently we generally confirmed them and determined DNA/µm in the centromere/kinetochore on the basis of the sequenced length of DNA in contigs that were localized by BAC FISH on spreads of SCs.

1.17 Data pre-processing of the reference genome

1.17.1 Construction of a non-redundant set of BAC contigs

Previously determined tomato BAC sequences[62], consisting of 984 HTGS phase 3 BACs (106.1 Mb), 266 phase 2 BACs (28.1 Mb), and 208 phase 1 BACs (23.2 Mb), were collected from the SGN ftp repository (ftp://ftp.sgn.cornell.edu/tomato_genome/bacs/old/bacs.v548.seq). The 1,250 phase 2 and 3 BACs were assembled into 658 non-redundant contigs with a total length of 117.6 Mb. The BAC contigs were subjected to base error correction using the Illumina data (as outlined in **section 1.18.4**) and screened for residual contamination of *E. coli* sequences (GenBank accessions AC_000091, NC_000913, NC_002655, NC_002695, NC_004431, NC_007946, NC_008253, NC_008563, NC_009800, and NC_009801) through megablast alignments using a 98% identity cut-off. Sequences matching to *E. coli* were excised from the contigs with an additional 100 bp flanking both sides of the matched region. The merged contig sequences were subsequently split into smaller fragments based on the sequence gaps they contained (as a result of the phase 2 BACs), resulting in a total of 1,118 sequences with a total length of 117.5 Mb.

1.17.2 Sanger read clipping and filtering of SBM data

The SBM reads (**section 1.3**) were quality clipped and screened for vector and *E. coli* contamination using Lucy 1.20p[63]. In total, 3,797,957 reads passed the read clipping and filtering (the 'filtered' reads in **Supplementary Table 15**).

1.17.3 Read clipping and filtering of Sanger clone ends

Quality of the tomato clone end sequence reads was assured by processing each read to flag regions of suspected vector origin or low sequencing quality, followed by screening the resulting high-quality sequences for contaminants. Low-quality sequence regions were found by examining the sequence quality scores assigned by Phrap, and vector sequences were found by comparing each read to the sequence of the cloning vector for the appropriate library using Cross_match. In the contaminant screening step, NCBI BLAST was used to compare sequences to sequences from likely contaminants including tomato chloroplast, *Arabidopsis thaliana* mitochondria (when the tomato mitochondrial genome sequence was not yet available), and Lambda phage. Importantly, to retain the option of later re-processing, no original clone end sequence data were discarded during quality control; regions of low quality or suspect origin were recorded as annotations on the raw, original sequence. Duplicate reads (i.e., multiple reads from the same BAC and primer combination) were discarded such that the longest read was maintained. One fosmid clone had identical reads for both ends and was discarded from further processing (**Supplementary Table 16**).

1.17.4 454 read filtering

The 454 reads were grouped per fragment size and processed without removal of the mate-pair linker sequence, if present. Duplicate reads, which most likely correspond to (emulsion) PCR duplicates created during library construction, were removed as follows using two custom Python scripts. First, all homopolymers in the reads were compressed to a length of one (e.g. CAAAATTG would become CATG). The reads were then ordered alphabetically and by descending length. Two reads were considered duplicate if the first 90% of the shorter of the two was an exact substring of the 5' side of the longer read. This was implemented to allow for a variation in read length in duplicated sequences, as two individual beads populated with identical DNA molecules on the picotiter plate do not necessarily yield the same read length. In this manner, groups of duplicate reads were defined and for each group, the longest sequence was retained. Additionally, reads containing less than 50 flows, more than 450 flows or more than one ambiguous base call (N) were discarded. After filtering, all homopolymers were inflated back to their original length (e.g. from CATG back to CAAAATTG). The filtering algorithms were implemented in Python 2.6.

In total, 26% of the 454 read data was discarded (**Supplementary Table 17**). The vast majority of discarded reads were clonal duplicates, whereas the number of reads with ambiguous base calls and the number of reads with aberrant flow counts contributed only marginally to the number of discarded reads. The clonal duplicates were most likely created both through the sample amplification PCR during library preparation and the emulsion PCR in the sequencing protocol. Remarkably, 53% of the sequence data from the 20 kb fragment libraries was discarded, whereas this figure ranged between 21% and 24% for the other fragment sizes, pointing to increased fragment duplication during the preparation of the 20 kb fragment library. Furthermore, there was a considerable variation between libraries created from the same fragment length, with the highest fraction of discarded reads occurring in the libraries from which the most reads were generated. While a single library of a given fragment size did not provide an even, unbiased coverage of the genome, employing multiple libraries of each fragment size alleviated this problem.

1.17.5 SOLiD read trimming and filtering

The SOLiD reads were quality-trimmed by means of an in-house developed program that makes use of the SOLiD quality scores. The longest region in a read where the quality of each 14-base window was above an average score of 12 was selected. Both ends of the selected region were subsequently extended such that each of the 4 terminal bases has a quality score above 14. Reads with trimmed lengths below 35 bp for the 50 bp libraries, or lengths below 20 bp for the 35 bp libraries, or an average quality below 15 were discarded. The accepted reads were then corrected by means of the SOLiD Accuracy Enhancer Tool, which implements a modified version of the spectral alignment error correction algorithm proposed by Pevzner *et al.*[64].

The adjusted reads were aligned against the assembly using PASS[65] and coupled using the pairing option of PASS. This step was essential to produce the data required for the evaluation of the structural correctness of the *de novo* assembly and of the validation of the chloroplast insertions into the nuclear genome, described in **Section 1.22.**

Similar to 454 data, the SOLiD mate-pair libraries turned out to be affected by different levels of read duplication (also referred to as polyclonality). For each library, all the mate-pair reads that mapped on the same genomic positions were identified, and only one representative for each clonal duplicate was retained for subsequent analyses. Since the 4 kb mate-pair library contained 97.4% of duplicated fragments, it was not further considered (see **Supplementary Table 18**).

1.18 Sequence assembly of the reference genome

1.18.1 Summary of the *de novo* assembly

An initial *de novo* assembly of the *S. lycopersicum* 'Heinz 1706' genome was constructed from a 25-fold coverage the pre-processed 454 Titanium and Sanger read datasets, resulting in 3,761 scaffolds spanning 782 Mb with 95% of the assembled scaffold sequence present in only 225 scaffolds. This *de novo* assembly was further scaffolded using approximately 200,000 BAC and fosmid end sequence pairs. Gap filling and improvement of the overall base accuracy were achieved through integration of: 1) high-coverage Illumina and SOLiD data; 2) a second *de novo* assembly produced from the 454 and Sanger reads, and; 3) 117 Mb of high-quality BAC clone sequences. The quality of the *de novo* assembly was further improved through the removal of organellar and bacterial contamination, and structural inconsistencies between the sequence and physical maps were resolved. A mixture of state-of-the-art sequence assembly tools combined with custom developed scripts was employed to produce an initial high quality, near-complete reference genome sequence for tomato.

1.18.2 *De novo* assembly of the 454, SBM and SCE reads

An initial backbone assembly was generated from the filtered 454, SBM and Sanger sequenced clone ends (SCE) using Newbler 2.3-PostRelease-01/11/2010[32] with the "Large or complex genome" option. The 454 data consisted of 12.1-fold coverage in shotgun reads, 6.2-fold coverage in 3 kb mate-pairs, 3.3-fold coverage in 8 kb mate-pairs, and 1.6-fold coverage in 20 kb mate-pairs. The SBM and SCE data, corresponding to 3.5-fold and 0.25-fold coverage, respectively, were assembled as shotgun sequences.

The resulting *de novo* assembly spanned 782 Mb in 3,761 scaffolds with an average length of 208 kb (**Supplementary Table 27**). Out of the 110,872 assembled contigs, 32,899 were assigned to scaffolds. Strikingly, 95% of the assembled scaffold sequence was present in only 225 scaffolds with a minimum length of 625 kb. The total genome size was estimated by both the Newbler and CABOG (see below) assemblers to approximately 900 Mb (884.2 and 902.5 Mb, respectively), and the fold coverage values in this section are based upon this estimate of 900 Mb. As such, the *de novo* assembled scaffolds represent 87% of the full genome sequence.

The assembler successfully incorporated 94.4% of the 454 shotgun reads, 91.8% of the 454 mate-pair reads, 94.9% of the SBM reads and 90.6% of the SCE reads into the assembly (**Supplementary Table 28**). The high fraction of assembled reads from each of the libraries (generated through different protocols and sequencing technologies) indicates a near-completeness of the assembled genome sequence. The majority of the unassembled reads were identified as repetitive by Newbler, whereas a small fraction was deemed to be singletons and outliers. The MboI digested BAC library had a substantially higher percentage of repeat reads than the HindIII and EcoRI BAC libraries. This corresponds well to previous findings[66] where sequence analyses of the SCE sequences revealed that the MboI sequences contained a large fraction of rRNA repeats.

During the scaffolding step of the assembly, approximately 64% of the paired reads from the 454 mate-pair libraries were mapped with both ends to a unique location on the assembly within the expected mate-pair distance (**Supplementary Table 29**). Additionally, 6% of the pairs were reported as false pairs, indicating that these pairs mapped outside of their expected pair distance, or in an incorrect orientation. Another 21% of the pairs mapped to multiple locations with either one or both ends. Moreover, 9% mapped with one end to the assembly (and not with the other), whereas only 0.3% did not map to the assembly at all. There was a positive relation between the fragment size and the fraction of pairs that linked two contigs together into a scaffold, emphasizing the benefit of large-fragment libraries in obtaining a contiguous assembly. Remarkably, the 3 kb fragment libraries displayed a substantial higher fraction of false pairs than the larger fragment libraries. However, more than one third of these were mapped on the same contig just outside the expected pair distance, suggesting that this is the result of a higher variation in fragment length in the 3 kb libraries. There was little variation in the fraction of unmapped reads between the three fragment sizes, suggesting a similar distribution of the libraries over the genome.

1.18.3 Base error correction with SOLiD reads

The per-base error rate of the *de novo* assembly was subsequently reduced through alignment of the 50 bp SOLiD read pairs from the 7 kb and 8 kb libraries. The filtered reads were aligned to the contig sequences with PASS[65] and putative indels and base substitution errors were identified from the alignments. Indels are known to be the most frequent error generated by 454 pyrosequencing chemistry[67] and are mainly located in homopolymer regions. In contrast, mismatches can either be the result of sequencing errors, repeated regions that have been collapsed into incorrect consensus sequences by the assembler, or heterozygous sites in the genome.

Single base substitutions were detected through alignment of the filtered SOLiD reads at 93% identity. Putative errors were corrected if at least three reads confirmed the substitution and if 90% of all aligned reads agreed on the substitution. To detect indel events in homopolymer regions, all filtered SOLiD reads with homopolymers of five colours or longer were considered. For each such read, five variants (-2, -1, 0, +1, +2) were produced that differed in homopolymer length. These reads were then aligned to the contigs without allowing for base mismatches. Homopolymer tracts with less than three aligned reads (or less than two aligned reads if the quality in the assembled contigs was below 20), or more than 40 aligned reads were discarded from further processing. The correct length of each homopolymer tract was then determined from the alignments if at least 80% of the aligned reads agreed on the length.

Using this conservative approach, 42,481 putative errors were corrected in the assembly through alignment of the SOLiD reads, corresponding to one correction every 18 kb. Of these, 10% were deletion events in homopolymers, 63% were insertion events in homopolymers, and 27% were base substitutions. As was expected, there were fewer base substitutions than indel events. Insertions were considerably more numerous than deletions, reflecting a tendency of the Newbler algorithm[32] to underestimate the length of homopolymer tracts.

1.18.4 Base error correction with Illumina reads

A second round of base error correction was performed by integrating the Illumina sequence data into the 454/Sanger hybrid assembly. Because Illumina data are produced in nucleotide space and as such every base is interrogated only once, a different approach was followed to optimize the number of base error corrections while limiting the amount of sequencing errors that could be introduced by this step. Rather than integrating the raw read data, a methodology similar to that of ECINDEL[68] and SRCorr[69] was applied to the assembly as follows. First, 31-mers were extracted from the two Illumina paired-end libraries with the largest fragment sizes and the frequency of each 31-mer in these reads was counted. The two libraries with shorter fragment sizes were not included, as these displayed an aberrant 31-mer distribution (data not shown). The frequency of 31-mers was plotted against their volume in the reads (**Supplementary Fig. 25**), and based on the resulting distribution all 31-mers with a frequency of two or lower were discarded, as these likely represent the majority of sequencing errors. In total, $7.18*10^8$ distinct 31-mers remained after filtering, with a total volume of $2.21*10^{10}$ 31-mers. Assuming a 900 Mb genome size, this corresponds to a 30.7-fold coverage of the genome, which coincides with the peak of the Poisson-shaped distribution in **Supplementary**

Fig. 25. Based on these findings, the 31-mers were considered to represent the vast majority of all correct 31-mers in the tomato genome, and assumed to contain very few 31-mers not present in the tomato genome (i.e., 31-mers containing one or more sequencing errors).

Next, the coverage of each position in the assembly by these 31-mers was determined, such that positions in the assembly that were fully in agreement with the Illumina read data would be covered by 31 overlapping 31-mer, whereas potential errors in the assembly would not be covered by any 31-mer. Positions surrounding putative sequencing errors would have a coverage ranging from 1 at the positions directly adjacent to the putative error in the assembly to 30 at the positions 30 bp away from the site of putative error. Since errors in or adjacent to repetitive sequences are likely to still match some of the 31-mers, such positions would have a coverage higher than 0 (but lower than 31). On the other hand, it is likely that not all regions in the genome were sampled at sufficient depth in the Illumina data[70], implying that some of the error-free positions could have a coverage lower than 31. To correct for this, putative erroneous positions were defined as positions with coverage lower than 11, whereas correct positions were defined by coverage of 21 or higher. The correctness of positions with coverage between these values was considered to be unknown. Prior to correction, 98.3% of all positions were considered correct whereas 0.8% were identified as putatively erroneous.

For each putative error position, all possible single-nucleotide substitution, insertion and deletion events were attempted. If there was a single best-scoring event with a minimum 31-mer coverage of 21 or higher, then the original nucleotide was assumed to be erroneous, and corrected. In total, 84,344 positions were corrected in the assembly, corresponding to one correction in every 9 kb. Of these, 28% were deletions, 56% were insertions and 16% were base substitutions. This resulted in an increase of the fraction of correct positions to 98.6%, with 0.6% remaining erroneous.

1.18.5 Removal of organellar and bacterial contamination

After base error correction of the *de novo* assembly, 77,973 contigs spanning 34.4 Mb that were not included in the scaffolds were separated from the genome assembly. While the majority of these sequences likely represented *bona fide* tomato sequence, their short length would not contribute to subsequent genome analyses. In the remaining scaffolds, residual contamination of *E. coli* (GenBank accessions AC_000091, NC_000913, NC_004431, NC_008563, NC_009800, NC_009801, and NC_010473) and cloning vector (ftp://ftp.ncbi.nih.gov/pub/UniVec/) sequence was identified and removed from the assembly using Cross_match

(http://www.phrap.org/phredphrap/phrap.html) with parameters set to "-penalty -3 -minscore 96 -minmatch 100". Using the same methodology, scaffolds consisting only of plastid or mitochondrial sequence were identified and removed through alignments to the tomato chloroplast and mitochondrial genomes (GenBank accessions NC_007898 and FJ374974, respectively). In total, 17 scaffolds spanning 87 kb were identified as contaminants and removed from the assembly.

1.18.6 Correction of structural inconsistencies

The combination of the complexity and repetitive nature of the tomato genome, and the heterogeneity of sequence data used in the assembly represented a potential source for structural inconsistencies within the assembled scaffolds. To resolve these, potential chimeric scaffolds were identified through comparisons with the genetic map and the WGP physical map. Markers and tag sequences from the WGP map were aligned to the assembly as described in **section 1.19**. Scaffolds that matched to marker sequences from multiple chromosomes were flagged as potentially chimeric. Scaffolds were also flagged when three or more BACs included in the WGP map suggested a link between two scaffolds that contained marker sequences assigned to different chromosomes. The most likely breakpoint in each flagged scaffold was subsequently determined by manual inspection of the position and order of the genetic markers, the aligned sequence tags from the WGP map, and additional alignments to a preliminary *de novo* assembly and the CABOG assembly (see below). In total, 22 breakpoints were identified in 20 scaffolds and subsequently broken.

1.18.7 Integration of an alternative *de novo* assembly

An alternative *de novo* assembly was generated from the pre-processed 454, SBM and SCE data using the CELERA/CABOG assembler version 6.0 beta2 (http://sourceforge.net/apps/mediawiki/wgs-assembler/) with parameters set to "doOverlapTrimming = ; overlapper = mer; unitigger = bog; utgErrorRate = 0.03; bogBadMateDepth = 7; fakeUIDs = 1". The assembly was subsequently screened for bacterial and organellar contamination as outlined above.

The CABOG assembly (**Supplementary Table 30**) displayed a number of superior contig statistics over the Newbler assembly, whereas the scaffold statistics were somewhat inferior. In particular, it showed a remarkably higher contiguity (as reflected by the longer average contig length and lower number of contigs) and a somewhat extended genome coverage, whereas the N95 scaffold index

(i.e., the number of largest scaffolds in the assembly that together comprise 95% of the total sequence length) was more than twofold higher than that of the Newbler assembly. Owing to its longer contig sequences, the CABOG assembly contained a lower fraction of scaffolding gaps than the Newbler assembly (1.1% and 6.9% of 'N' nucleotides in the scaffolds, respectively).

The higher contiguity of the CABOG contigs was exploited to improve the Newbler assembly through gap filling as follows. For each gap in the Newbler assembly, a pair of flanking 2 kb sequences was extracted and aligned to the CABOG assembly using megablast. Pairs of flanking sequences that matched to the same CABOG contig were selected for gap filling when they aligned in the correct orientation and the distance between them was lower than 20 kb. Through this procedure, 3,095 gaps in the Newbler assembly were replaced by the corresponding sequences extracted from the CABOG assembly. This resulted in 4.9 Mb of gaps being replaced by 4.5 Mb of true genome sequence, revealing a tendency of the Newbler assembler to somewhat overestimate the gap size.

1.18.8 Improved scaffolding using BAC and fosmid clone end sequences

The scaffolding of the assembly subsequently was improved through the integration of 135,271 BAC and 64,722 fosmid end sequence pairs using the Bambus software (http://sourceforge.net/apps/mediawiki/amos/index.php?title=Bambus). The hierarchical scaffolding method implemented in Bambus was exploited to prioritize the smaller fosmid clones (30 kb ± 10 kb) over the larger BAC clones (120 kb ± 70 kb). The clone end sequences were mapped on the scaffolds using megablast. The megablast output file then was used to generate the ".contig" file required for Bambus. In total, 3,260 links between scaffolds were detected. Of these, 916 were retained as potential valid links that satisfied the criteria of the library insert size and correct orientation. Links confirmed by less than two clones were rejected to minimize the number of incorrect links due to chimeric BAC or fosmid clones, or repetitive regions in the genome assembly. A total of 664 individual links were retained that together connected 328 scaffolds. As a result, the N95 index of the assembly decreased from 225 to 73 scaffolds while the maximum scaffold span increased from 15 Mb to 42 Mb.

1.18.9 Integration of BAC clone sequences

The non-redundant BAC contig sequences (see **section 1.17.1**) were aligned to the sequence scaffolds using megablast with a 98% identity cut-off and a word size of 200, without masking for low-complexity regions. The alignments were manually curated to remove spurious matches. If a single, linear match was identified between the BAC sequence and the scaffold sequence, and the scaffold sequence did not contain insertions or mismatches longer than 500 bp relative to the BAC sequence, then the scaffold sequence was replaced by the BAC sequence in that region. If the BAC sequence matched to multiple distinct scaffolds, or multiple regions on the same scaffold separated by more than 500 bp, then the BAC sequence was split accordingly and the partial BAC sequences were integrated in a similar fashion.

A total of 116.6 Mb of BAC sequence, representing 99.3% of the non-redundant BAC contig sequences, was integrated into the assembled scaffolds. As a result, 2,597 gaps within the scaffolds were closed, replacing 4.6 Mb of gaps with their corresponding sequence. For 201 of the closed gaps, the corresponding BAC sequence (with a total length of 0.6 Mb) matched to small, single-contig scaffolds from the assembly. These scaffolds were subsequently removed from the assembly.

1.19 Map integration of the reference genome

1.19.1 Summary of map integration

The sequence scaffolds from the error-corrected *de novo* assembly were integrated with two BAC-based physical maps, generated through SNaPshot fingerprinting and Keygene's WGP, respectively. We subsequently anchored the scaffolds to their corresponding chromosomal locations through a high-density genetic map combined with genome-wide BAC-FISH. A custom algorithm was developed in Python2.6 to automate the integration of these data in a systematic and consistent manner. The final integrated assembly consists of twelve chromosomal pseudomolecules that together span 91 scaffolds with a total length of 760 Mb, 594 Mb of which has been oriented. This is supplemented by a virtual chromosome *0* that contains an additional 22 Mb of unanchored sequence.

1.19.2 Assignment of the scaffolds to chromosomes

The sequence scaffolds were assigned to their corresponding chromosomes through alignment of marker sequences from the integrated genetic linkage map constructed using the Tomato-EXPEN 2000 mapping population[45]. The marker sequences were aligned to the scaffold sequences using BLASTN with an E-value cut-off of $1*e^{-1}$, followed by manual curation. A total of 2,082 markers could be located on the scaffold sequences. Scaffolds were then assigned to a chromosome if they spanned at least 10,000 bp and contained at least one marker sequence that was anchored to the genetic map. In total, 89 scaffolds together spanning 758 Mb were assigned to a chromosome, whereas the remaining 3,134 scaffolds (23 Mb in total) were assigned to a virtual chromosome *0*. All data were subsequently processed per chromosome.

1.19.3 Integration of the physical maps

The sequence tags from the BACs in the WGP map were aligned to the scaffold sequences using SOAPaligner v2.20[71], allowing for perfect matches only. Out of the 261,913 sequence tags, 91.6% aligned to the scaffolds with a single match and 6.6% with multiple matches. The remaining 1.8% did not match to the scaffolds. The uniquely matched tags were subsequently clustered per WGP contig. Such clusters were split if the distance between two subsequent tags was larger than 100 kb (corresponding roughly to the length of a BAC insert). Contigs matching to less than two scaffolds, containing less than 50 tags, or having less than 75% of their tags matching to a unique location to the genome sequence, were discarded. The remaining WGP contigs were used to connect sequence scaffolds if at least 10 tags matched to each sequence scaffold, and if these matches were at the outer ends of both the contigs and the scaffolds. The end sequences of the BACs in the SNaPshot map were aligned to the scaffold groups from the previous step using BLASTN. The aligned end sequences were grouped and processed similar to the sequence tags of the WGP map as outlined above. SNaPshot contigs from which at least three end sequences matched to two scaffolds were selected for joining these scaffolds.

In total, three potential scaffold links satisfying the WGP tag alignment criteria were identified, through which two pairs of scaffolds were eventually linked. The third link connected a scaffold from chromosome *2* to a scaffold assigned to chromosome *3*. Since both these scaffolds had been anchored to their corresponding chromosomes with high confidence, the WGP contig was deemed incorrect, likely as a result of a chimeric BAC clone within the contig. Seven potential scaffold links were identified from the SNaPshot map, two of which overlapped with those from the WGP map. In addition to the two scaffold pairs linked through the WGP map, another four scaffold pairs

were linked through the SNaPshot map. Two of these links connected a scaffold from chromosome *0* to a scaffold on chromosomes *9* and *10*, respectively. The remaining link conflicted with another, better link and was not incorporated.

Summarizing, only six pairs of sequence scaffolds were linked through the integration of the WGP and SNaPshot maps. This low number of links can be attributed to the approximately 20-fold higher level of fragmentation in the physical maps compared to the sequence assembly. In total, 95% of the BACs were assembled in 1,674 and 1,217 contigs in the WGP and SNaPshot maps, respectively. In contrast, 95% of the assembled sequence was present in only 73 scaffolds.

1.19.4 Integration of the genetic map and FISH data

After integration of the physical maps, the resulting scaffold groups (with each scaffold group representing one or more scaffolds that are physically connected) were ordered and oriented on the chromosomes using a combination of the genetic map and the FISH data (**Supplementary Table 31**). The alignment of the marker sequences from the genetic map is described above. The BACs that were used in the FISH experiments and for which sequence data were available were located on the scaffolds using BLASTN with an E-value cut-off of $1*e^{-1}$, followed by manual curation. A total of 366 BAC clones had either the entire clone sequence, the clone end sequence(s), or the marker sequence used for the clone isolation available, and could be located on the scaffold sequences in this manner.

The orientation of a scaffold group on the corresponding chromosome was determined as a consensus from both the genetic map and FISH data. If at least one of these two provided an orientation, and it did not conflict with the orientation provided by the other resource, then the scaffold group was oriented accordingly. If neither resource provided an orientation, or if both resources were in disagreement, then the scaffold group remained non-oriented.

For each scaffold group that contained at least four marker sequences, an orientation was derived from the genetic map by considering the placement of all markers compared to the median, and the orientation of each pair of markers on the scaffold compared to their order on the genetic map. Here, the median refers to the median genetic position (in cM) of all markers on a given scaffold group. A scaffold group was oriented if the markers below the median position were separated by at least a quarter of the scaffold's length from the markers above the median, or if at least two-thirds of the marker pairs had an identical orientation relative to each other. Similarly, the orientation of a scaffold was determined from the FISH experiments through the orientation of each pair of BACs.

A scaffold group was oriented if at least two-thirds of the BAC pairs had an identical orientation, and if there were no conflicts between the various experiments.

The order and orientation of scaffold groups on each chromosome were determined from the genetic map and FISH data. When interpretations of the two types of data were in conflict, the order on the genetic map was used, thereby maintaining the general layout of the scaffold groups according to the genetic map. However, most of the conflicts between linkage and FISH mapping were in heterochromatic regions where crossing over is infrequent (**Fig. 1**, **Supplementary Fig. 1, Supplementary Table 13**).

In total, 37 scaffold groups could be oriented based on the genetic map, and the same number of scaffold groups could be oriented by the FISH data. Out of these, 27 scaffold groups had evidence for their orientation from both the genetic map and the FISH data, without conflicts. In summary, there was evidence to orient 47 scaffold groups (representing 53 scaffolds with a total length of 594 Mb) on the chromosomes, whereas 38 scaffolds remained non-oriented. The final integrated assembly (**Supplementary Table 27**) consists of 12 chromosomal pseudomolecules that together measure 760 Mb, and a virtual chromosome *0* containing all unanchored sequence that spans 22 Mb. On average 85% of each chromosome has been assembled (**Supplementary Table 32**) compared to cytogenetic estimates of the chromosome sizes[58], adjusted to the estimated genome size of 900 Mb. The 12 chromosome sequences have a GC content of 34.0%, which corresponds well to that found in previous analyses of genomic tomato sequence[66,72].

1.20 Validation of the reference genome

1.20.1 Summary of genome validation

The structural correctness of the *de novo* assembly was validated independently through the alignment of SOLiD mate-pair sequences and BAC and fosmid end sequence pairs. In total, only 34 putative chimeric contigs were detected through the SOLiD mate-pairs. Less than 0.1% of the aligned BAC and fosmid end sequence pairs aligned inconsistently with the assembly, showing the long-range consistency of the assembly. Alignment of the BAC contigs from the WGP physical map suggested up to 13 potential discrepancies between the final integrated assembly and the physical map, whereas 97.1% of the contigs in the WGP map were collinear with the twelve assembled pseudomolecules. The per-base accuracy of the integrated assembly was determined through the alignment of high-quality BAC sequences, revealing a substitution error rate of 1 per 29.4 kb and an

indel error rate of one per 6.4 kb. In total, 98% of all WGP tags and 98% of all publicly available tomato EST sequences could be aligned to the assembly, demonstrating the near-completeness of the tomato genome assembly.

1.20.2 Read coverage distribution in the *de novo* assembly

Collapsed sequences (i.e., sequence elements that occur in multiple copies in the genome, but are assembled into a single copy sequence) are a common problem in genome assembly and manifest themselves as regions in the assembled sequence with elevated read coverage. The average read coverage of the contigs in the tomato *de novo* assembly was 27.2X and the median coverage was 26X, with 99.0% of the contig positions covered by 1 to 52 reads (i.e., between zero and two times the median coverage). In total, there were 7,791,395 positions in the assembly with a coverage higher than 52, with a highest read coverage of 2,247. Of the positions with read coverage above two times the median, 4,684,818 were found clustered into 17,830 regions of 100 bp or larger, in a total of 8,577 distinct contigs. These findings imply that there may be a substantial number of small collapsed regions in the assembly; however, the extent of these is limited to approximately 5 Mb.

1.20.3 Validation of the structural correctness of the *de novo* assembly

While the Sanger reads from the Selected BAC Mixture (SBM) were generated using the double-barrelled shotgun approach, they were incorporated into the *de novo* assembly as unpaired reads. As a result, the fraction of SBM read pairs found in the assembly at the expected distance relative to each other provides an estimate of the structural correctness of the assembly. In total, 81.5% of the SBM read pairs had both reads assembled correctly into the same scaffold whereas 0.6% of the pairs were incorporated inconsistently within the same scaffold. The remaining 17.9% of pairs did not assemble into the same scaffold. The median span distance (i.e., the sum of the length of both reads plus the distance between them) measured 2,753 bp for pairs assembled within the same contigs, and 3,498 bp for pairs between contigs in the same scaffold. The difference between these values confirms once more the tendency of Newbler to overestimate the gap size between contigs. In total, 73.9% of the read pairs within contigs supported a fragment length of 2,750 – 3,750 bp and 76.3% of the pairs between contigs supported a fragment length of 2,500 – 4,500 bp (**Supplementary Fig. 26**). Strikingly, 8.3% of the pairs within a single contig had a span of 1,000 bp or less, hinting either at an abundance of locally collapsed assemblies or at an aberrant fragment

length distribution in the SBM libraries; however, the former was not supported by the SOLiD analyses described below. Thus, the placement of the SBM read pairs is largely consistent with the genome assembly and provides further evidence for the structural correctness of the assembly.

The SOLiD mate-pair data were used to validate the *de novo* genome assembly generated from the 454 and Sanger reads. The mate-pair reads were aligned on the assembly to detect possible chimeric contigs. Here, chimeric contigs are defined as sequences composed of two non-consecutive genomic fragments that have incorrectly been assembled together by the assembler. On such contigs there are no SOLiD mate-pairs that map across the misassembled region, resulting in a drop of the local physical coverage (LPC). The LPC for a given position in the sequence is determined by counting the number of intra-contig mate-pairs that align to that position and that map at a distance compatible with the library. Putative breakpoints in the chimeric contigs were then identified as regions where the LPC drops to zero. Further evidence of misassembly was obtained by looking at the inter-contig mate-pairs (ICMs), *i.e.*, mate-pairs that were aligned at the edges of the putative breakpoints and that link these edges with other contigs (**Supplementary Fig. 27**). A contig was considered chimeric only when both the above conditions were fulfilled, that is the LPC falling to zero and some ICMs connecting the putative chimeric contig to another contig in the assembly.

The 1 kb and 8 kb SOLiD mate-pair reads were aligned to the assembly with a minimum identity of 90%. To determine the LPC, only the mate-pairs of which both reads were mapped to a unique location in the assembly and mapped inside the same contig were considered. Mate-pairs were considered to be mapped correctly if the distance between the mapped reads was in the interval $\mu \pm 3\sigma$, where μ is the mean value and σ is the standard deviation of the insert length distribution.

Out of the 110,872 contigs in the *de novo* assembly, the correctness of 33,129 contigs was assessed using the 1 kb library, while 14,748 contigs were evaluated with the 8 kb library (**Supplementary Table 33**). The remaining contigs had no mate-pairs mapped to them as they were shorter than the mate-pair fragment length. Moreover, putative chimeric regions present on the contig boundaries could not be detected because the LPC cannot be calculated for these positions. The 1 kb and 8 kb libraries allowed detection of 31 and 13 putative chimeric regions respectively, several of which were confirmed by both libraries. In total, only 34 putative chimeric contigs were detected, confirming the general structural correctness of the *de novo* Newbler assembly in fine detail.

To study the structural correctness of the *de novo* assembly on a higher level, the filtered, paired-end sequences of the BACs and fosmids (which were included as shotgun reads in the *de novo* assembly) were aligned to the scaffolds using BLASTN. In order to ensure unambiguous mapping, only sequences of at least 300 nt that aligned to a unique location with a coverage of 95% or more

and an identity of 99% or better were used. In total, 53.8% of the BAC end sequence pairs and 75.3% of the fosmid end sequence pairs could be aligned to a unique position on the scaffolds with these stringent criteria. Pairs of end sequences that aligned to a single scaffold with incorrect orientation (i.e., both end sequences aligned to the same strand), incorrect direction (i.e., both ends facing outward rather than inward), or at a too large distance from each other (more than 300 kb for the BAC libraries, or more than 60 kb for the fosmid libraries) were considered indicators of potential misassembly. In total, only 44 out of 63,631 aligned pairs of BAC end sequences (0.07%) and 51 out of 42,469 aligned pairs of fosmid end sequences (0.12%) aligned inconsistently with the assembly (**Supplementary Table 34**), again confirming the overall structural correctness.

1.20.4 Per-base accuracy of the assembled scaffolds

The accuracy of a genome sequence has a tremendous impact on all downstream analyses performed on it. Base errors in the contig sequences can result in the incorrect annotation of functional elements such as protein-coding genes, transcription factor binding sites and small RNAs. To estimate the quality of the tomato genome assembly, the per-base accuracy was determined through alignment of the Sanger-sequenced BAC clones to the assembled scaffolds. Moreover, the improvement in accuracy through the integration of the SOLiD and Illumina data was measured by aligning Sanger sequenced BAC clones to both the initial *de novo* assembly and the error-corrected assembly. These BAC sequences were considered to be a gold standard of tomato genomic sequence throughout the assembly process, and were later integrated in the assembly to fill gaps and resolve additional errors.

In total, 117.5 Mb of non-redundant BAC sequences (**section 1.17.1**) were aligned to the scaffold sequences using megablast without filtering for low-complexity regions and with an identity cut-off of 98%. Repetitive matches and matches spanning less than 20 kb were discarded. The BAC contigs that passed these criteria were subjected to a more accurate alignment to the corresponding scaffolds using LASTZ[73] with an identity cut-off of 99%. The alignment scoring matrix was adjusted to a match score of 100 for all pairs of matching nucleotides, a mismatch penalty of 100, a gap creation penalty of 100 and a gap extension penalty of 400. In this manner, all mismatching nucleotide pairs are equally probable, and multiple small gaps invoked by homopolymers of incorrect length are less likely to be wrongly aligned as mismatches. The alignments generated by LASTZ were subsequently processed with custom Python scripts to remove repetitive matches and

truncated to remove regions with a local sequence identity lower than 99%, and regions that were not aligned to either the *de novo* or the error-corrected assembly.

A total of 108.3 Mb of BAC sequence was aligned consistently to both the assemblies. Prior to error correction, there was one indel per 4.0 kb and one substitution error per 30.4 kb, corresponding to one error every 3.6 kb on average. After the integration of the SOLiD and Illumina data, the average error rate was reduced to one error per 5.3 kb, with one indel per 6.4 kb and one substitution error per 29.4 kb. While there was a slight increase in the substitution error rate as a result of false-positive corrections, the overall error rate was reduced substantially. Strikingly, the substitution errors were often found clustered together, with 67.6% of the distances between adjacent substitution errors in a single alignment being below 100 nt. This suggests that many of these represent assembly errors (such as an incorrect consensus sequence for a repeat element that was assembled from reads derived from different repeat copies) rather than independent sequencing errors.

A well-known problem in 454 sequencing is the inaccuracy with which the length of homopolymer tracts is determined. This property of 454 pyrosequencing is the predominant cause for indel errors and increases with the length of the homopolymer tract[32]. In the *de novo* assembly, the error rate of homopolymer tracts shorter than six was found to be below 0.6%, and it increased dramatically with longer tract lengths. For example, more than 16% of the homopolymer tracts of 8 nt had an incorrect length in the *de novo* assembly when compared to the BAC sequences. The error correction with the SOLiD and Illumina data reduced the fraction of homopolymer tracts with incorrect lengths, with a threefold and fourfold reduction in error rate for tracts of lengths seven and eight, respectively (**Supplementary Fig. 28**). While the error rate of longer homopolymer tracts remained substantial even after error correction, such long homopolymer tracts are not often found in tomato genes. Specifically, there are only 643 homopolymer tracts of 9 nt or longer in the annotated CDS features on the tomato genome.

Taken together, these data underline the high base accuracy of the assembled genome sequence. The substitution error rate of 1 per 29.4 kb is somewhat higher than that of a “gold-standard” draft genome[74], however many of these errors cluster together in small regions of the assembly. The indel error rate (one per 6.4 kb) is relatively high and leaves room for improvement, nonetheless the majority of these errors were found in long homopolymer stretches outside of the gene coding regions.

1.20.5 Structural correctness of the final integrated assembly

The structural correctness of the assembled chromosome pseudomolecules was inspected through the alignment of the sequence tags from the WGP physical map (see **section 1.19.3**). For each BAC in the WGP map, the uniquely aligned tags were separated per chromosome and subsequently split into clusters, such that adjacent tags within a cluster were at most 100 kb apart on the chromosome sequence. Clusters consisting of less than three tags were discarded. Out of the 66,084 BACs that were used in the construction of the WGP map, only 2.4% of the BACs were split into multiple clusters of tags, whereas 5.3% did not contain any cluster of three or more tags. The remaining 92.3% (60,960 BACs) were aligned as a single cluster of tags, confirming the overall consistency of the assembled pseudomolecules.

Not only the individual BACs, but also the contigs created from these BACs showed an overall collinearity with the assembled pseudomolecules (**Supplementary Fig. 29**). Of the 66,084 consistently aligned BACs, 52,617 (86%) were assigned to 2,521 contigs in the WGP map, whereas the remaining BACs were singletons. In total, 1,932 of these contigs, containing 24,232 BACs, were collinear with the pseudomolecules. The sequence tags of another 516 contigs (25,730 BACs) aligned to two or more distinct regions in the assembly. In each of these alignments, the BAC contig was aligned such that each section of the contig was collinear with a corresponding region in the assembly, and each connection between distinct regions in the assembly was supported by only a single BAC. This pattern suggests that the discrepant BACs in these contigs are most likely chimeric, and that the assembled pseudomolecules are correct.

The remaining 73 contigs (2,655 BACs) created 96 connections between distinct regions in the assembly that were each supported by multiple BACs, and as such were not in agreement with the pseudomolecules. On closer inspection, 75 of these connected a scaffold from the artificial chromosome *0* to the interior region of a scaffold assigned to one of the other chromosomes, potentially providing a true chromosomal location for these unanchored scaffolds. Out of the remaining 21 conflicting connections, two were the result of the stringent 100 kb cutoff that was used to identify clusters of tags and another seven all concerned the same region.

In conclusion, 92.3% of the BACs and 97.1% of the contigs in the WGP map were collinear with the pseudomolecules, confirming the accuracy of the integrated assembly. From the non-collinear contigs, 13 discrepancies were identified that suggest potential breakpoints between the WGP map and the pseudomolecule assembly.

1.20.6 Completeness of the final integrated assembly

The completeness of the assembly, and in particular that of the non-repetitive, gene-rich regions, was assessed in several ways. Both the Newbler and CABOG assemblers provided an estimated genome size of approximately 900 Mb, out of which 782 and 787 MB was assembled by the two assemblers, respectively. The Newbler assembly included between 91% and 95% of all 454, SBM and SCE reads and covered 93% of the expected genome size. In total, 98.2% of the unique tags from the WGP physical map and 108.3 out of 117.5 Mb (92.2%) of the BAC sequences (that had been primarily selected from the gene-rich euchromatin) could be aligned to the assembly. These findings imply that there is no strong bias towards missing sequences in the WGS approach compared to sequence data derived from the BAC libraries.

Out of the 1,975 markers from the genetic map that yielded consistent sequence data, 1,974 could be aligned to the assembled scaffolds, and additionally all 108 COSII marker sequences could be aligned. Moreover, 98.0% of the 265,234 *S. lycopersicum* ESTs used in the annotation of the genome (**Supplementary section 2**) could be aligned to the 13 pseudomolecule sequences with a coverage and identity of at least 70%. In total, 79.2% of the aligned EST sequences were aligned with a coverage of at least 95% and an identity greater or equal to 98%. These data confirm that nearly the full gene space of tomato is present in the genome assembly.

1.21 The tomato mitochondrial genome and its occurrence in the nuclear genome

A shotgun sequencing strategy was used to produce an assembly of the tomato mitochondrial genome. Highly purified mitochondrial DNA (mtDNA) was isolated from *S. lycopersicum* 'Heinz 1706' etiolated seedlings following published protocols[75,76]. DNA quality controls were performed by PCR amplification of serial dilutions of the extracted mtDNA with primers designed on *actin* (nuclear-), *rubisco rbcL* (chloroplast-) and *coxII* (mitochondrial-specific) genes. Preparations with less than 10% nuclear and chloroplast contamination were subsequently fragmented by sonication and/or by hydrodynamic shearing to a range of 1,000-3,000 bp, purified from agarose gels and ligated into the pMOS-Blue vector following the manufacturer's instructions (Amersham Bioscience). A total of 5,777 sequence reads (average length 750 nt) from 19 independent libraries were produced (Qv≥20), vector masked and trimmed using SeqClean (http://sourceforge.net/projects/seqclean/). Out of these, 5,240 reads were filtered by BLASTN against an in-house DB of all available mitochondrial genomes from the phylum Embriophyta (http://www.mitochondrialgenome.org/BLAST/). A total of 4,154 reads showed 80% of sequence identity to at least one mitochondrial genome. This dataset was assembled with the CLC Assembly

cell large scale sequence analysis and Cap3 sequence assembly software. ORF identification was performed by reciprocal BLASTN comparison against the published tobacco mitochondrial genome[77] and visualized with the annotation tool of the Artemis program. *In silico* analysis for *numts* detection was performed with the Progressive Mauve Alignment software (PMA v2.3.1)[78] using a minimum LCB (locally collinear blocks) weight of 400 to align the mitochondrial genome assembly (v1.5) to the nuclear chromosomes (v2.40). Identified LCB were subsequently analysed by a BLASTN against the nuclear genome assembly (v2.40) with an E-value cut-off of e^{-10}. Individual or clustered *numts* equal to or larger than the size of LCBs detected by the PMA software, as well as hits with maximum E-value are provided in **Supplementary Fig. 30a**. For fluorescence *in situ* hybridisation (FISH), 40 mixed clones of mitochondrial DNA in equimolar amounts were labelled by nick translation with digoxigenin and hybridized to spreads of complete sets of synaptonemal complexes (SCs) on glass slides. The techniques are as described previously[51] and above.

The assembled data comprised a total of 579,717 nt of non-redundant sequence with an average coverage of 9X. The assembly (v1.5) resulted in six scaffolds (SlmtSC_A, _V, _M, _R, _L and _B) and 164 contigs spanning 268,141 and 311,576 nt, respectively and is available at www.mitochondrialgenome.org. This Whole Genome Shotgun project (BioProject ID: 67471) has been deposited at DDBJ/EMBL/GenBank under the accession AFYB00000000. The version described in this paper is the first version, AFYB01000000. The size of the assembly is in agreement with the physical map previously reported[79], and its multipartite organisation is comparable to those reported for the tobacco[77] and rice[80] mitochondrial genomes. The tomato mitochondrial genome carries at least 36 protein-coding genes, three ribosomal RNA genes and 18 tRNA genes. These numbers are similar to those reported for other angiosperm mtDNAs, in which most of the genes encode ribosomal proteins and components of the electron transport chain (complex I to V) (**Supplementary Table 35**). An ORF search resulted in the identification of 30 additional sequences encoding hypothetical proteins.

Bioinformatic analyses of the mitochondrial and nuclear genome assemblies revealed 111 locally collinear blocks (LCB) on the mitochondrial genome to be collinear with the nuclear sequence. A total of 72 of these (~197 kb) were inferred to be nuclear sequences of mitochondrial origin (*numts*[81]). *Numts* varied in number, size, and position, ranging between zero and seven on chromosomes 2 and 5 with the highest number (21) detected on chromosome 11 (**Supplementary Fig. 30a**). Fluorescence *in situ* hybridisation (FISH) of mtDNA generally supported this *in silico*

analysis, except that a distinct signal for a mitochondrial sequence was detected on chromosome *6* (**Supplementary Fig. 30a**).

1.22 Validation of chloroplast insertions into the nuclear genome

A recurrent question in genome projects is whether mitochondrial and plastid sequences assembled into a nuclear genome represent real nuclear insertions or artefacts of cloning and assembly procedures. Scaffolds matching only to mitochondrial or chloroplast sequences had been removed from the assembly at an early stage (see **Section 1.18**); however, scaffolds partially matching these organellar sequences had not been removed. To assess whether these regions correspond to true nuclear sequences we developed an approach that makes use of SOLiD mate-pair sequences.

Mate-pair sequences from the 1 kb fragment library were trimmed, spectral corrected and aligned against the assembly using the PASS program. Aligned reads were filtered to a minimum identity of 90% and paired using PASS_pair from the PASS package. For subsequent analysis, only the mate-pairs of which both reads were uniquely matched to the assembly and mapped inside the same contig were used to compute the LPC index, as described above (see **Section 1.20.3**).

To identify putative plastid insertions into the nuclear genome, a BLASTN similarity search was performed on the assembly (v2.40) using the tomato chloroplast sequence[82]. The results were filtered using an E-value cut-off of e^{-110}, selecting alignments longer than 250 bases that can be validated as true insertion events with our procedure. The detected regions were considered genuine only if they were spanned by mate-pair inserts with an acceptable physical coverage. To verify this, for each matching region the number of correctly mapped mate-pairs from the 1 kb library was counted spanning each window of 600 bases. The rationale behind this strategy is that if such a window contained a misassembly, then there should be no mate-pairs spanning over this window. **Supplementary Fig. 31** shows a validation of this approach on simulated misassemblies of different lengths. All the alignments entirely comprised within an IR were counted only once. In contrast, the chloroplast fragments that incorporate both IR and unique sequence produce two alignments against the nuclear genome of different lengths: the longest one was retained. Of all the 664 nuclear regions matching the chloroplast genome, just one displayed this pattern and therefore must be considered as a contamination, presumably introduced in the assembly during the BAC sequence integration step (ch01: 1839870-1903186) (**Supplementary Table 36**).

The 663 validated chloroplast insertions account for 571,662 bp of nuclear genome sequence of chloroplast origin (**Supplementary Table 36**). The size distribution of these alignments, as well as

the high percentage nucleotide identity, appears to confirm that the plastid DNA integration in the nuclear genome is an ongoing and high-frequency process[83]. The 180 alignments longer than 1,000 bp account for 60.4% of all the sequences of chloroplast origin: the longest alignment is 10,903 bp and has 89.79% nucleotide identity with the chloroplast genome sequence.

1.23 Fine analysis of the chloroplast insertions in the nuclear genome

The validated alignments between the plastid and the nuclear genome represent in general small local blocks of nucleotide matches where BLASTN aligns the sequences as far as they exceed a certain threshold. A long plastid insertion into the nuclear genome is therefore expected to split into different blocks. This can be caused by the presence of gaps in the chromosome assembly (gap between contigs) or by the accumulation of point mutations, deletions or insertions of various origins within the inserted fragment[84].

The alignment results cannot be directly used to count the plastid insertions into the nuclear genome because their fragmentation causes an overestimation of the insertion number. To avoid this effect, the collinearity between chloroplast and nuclear alignments was investigated. One peculiarity of these fragmented pieces is that they are close to each other, both in the donor genome (plastid) and in the acceptor genome (nucleus), and that they maintain the same orientation (which is plus/plus for both genomes, or plus/plus in one genome and minus/minus in the other).

For each chromosome, the alignments were ordered for position and then manually inspected for collinearity. This reduced the number of putative plastid insertions to 492 (**Supplementary Table 37**) and revealed two long collinear regions in chromosomes *2* and *11*, which span 57,102 and 40,729 bp of chloroplast-derived sequence, respectively. Analysis of the genomic distribution of insertions also revealed several hot-spot regions for the integration of plastid DNA. Allowing a maximum gap size of 5 kb in the genomic assembly between nonlinear insertions, 52 of such regions were detected, the longest one, in chromosome *1*, of more than 30 kb and composed of 14 apparently independent plastid insertions.

In order to identify trends in the movement of specific plastid regions toward the nuclear genome, all block alignments were placed in the chloroplast sequence in such a way as to obtain a *per base* insertion value of each plastid base. This value, measuring the number of times each plastid base is integrated in the nuclear genome, can reveal loci that move more frequently than others. The *per base* insertion value was highest around two regions carrying the *ycf* genes (**Supplementary Fig.**

32); this might suggest their preferential retention inside the nuclear genome or alternatively an easier loss of dispensable elements in the chloroplast genome.

1.24 Analysis of the genomic regions flanking mitochondrial and plastid insertions.

In order to identify if there are preferential sites for the organellar insertions into the nuclear genome, the genomic sequences flanking all the insertions have been further investigated. As the length of these possible common sequences could be not foreseen, the analysis was carried out considering 10, 25, 50, 100 and 1000 bases for each flanking region. For each of these classes of sequence length, three different analyses were performed:

First, the nucleotide composition, both on single base level and di-trinucleotide level, was compared to the average genome base distribution. The IUPAC alphabet was also used to better highlight some tendency on word usage, anyway the flanking regions did not reveal specific bias in respect of the average genome composition;

Second, a pattern discovery approach, through clustering of the sequences, was applied. The rationale of the clustering step was to better discriminate the presence of different patterns, each of them possibly in common with only a part of the genomic regions. The clustering was performed with cd-hit (90, 95 and 9% of nucleotide identity)[85] and with ClustalW, at default parameters[86]. The clustering produced a number of clusters equal to the number of the analysed flanking regions, suggesting the absence of a clear pattern , or conversely the presence of very degenerated patterns that could not be easily detected with classical clustering approaches.

Finally, the insertion sites were checked for their proximity to specific regions, essentially repeat elements. Even if the organellar insertions appear to be in general near (usually within 1kb) to some transposon or retrotransposon fragments, the same results can be obtained sampling at random genomic regions placed within 200kb from mitochondrial or plastid insertion points. Therefore, the close proximity of the repeat elements to the insertion sites seems to be given by chance, and reflects the general features of the genomic space, the heterochromatic regions, where these insertions more easily accumulate.

All of these analyses confirm that the insertions of plastid and mitochondrial fragments into the tomato genome are not site-specific.

2 ANNOTATION

2.1 Summary

Annotation of the tomato genome was performed by the iTAG consortium (international Tomato Annotation Group). The iTAG annotation pipeline operates as a distributed, worldwide network of resources and experts (**Supplementary Fig. 33**). It uses SGN (http://solgenomics.net/) as a central data repository and exchange node. The iTAG pipeline performed annotation of repeats and masking of pseudomolecules, mapping of different protein sequence sets, ESTs and full length cDNAs, as well as RNA-Seq reads from Illumina, 454 and SOLiD platforms. Independent *ab initio* predictions were performed with GENEID[87], AUGUSTUS[88] and TWINSCAN[89,90], all specifically trained for tomato and potato. The above listed extrinsic data were integrated using the *a priori* informed gene prediction software EuGene[91]. EuGene prediction, followed by manual expert curation, produced a consensus annotation of 34,727 and 35,004 protein encoding genes for the tomato (iTAG v2.3) and potato nuclear genomes, respectively (**Supplementary Table 38**). Human readable descriptions could be assigned to 78% of tomato proteins. An OrthoMCL[92] clustering including several dicots as well as rice resulted in 8,615 gene families shared among all species, 562 tomato-specific gene families and 8,886 non-clustered tomato genes (singletons). The tomato annotation was evaluated by comparing predicted gene models with full length transcripts that were not included in the training sets used for *ab initio* gene prediction. Manual inspection was carried out to distinguish discrepancies caused by genome sequencing artefacts (and indels) from loci where the prediction failed. 2.25% of the genes, including all genes discussed in **Sections 5.3-5.8**, were manually curated.

As indicated above, the iTAG pipeline was also used for the annotation of the potato genome. The decision not to use the annotation provided with the initial release of the potato genome[93] was taken in order to avoid using results from different pipelines for the comparative analyses between tomato and potato. The potato iTAG annotation was compared to the annotation provided by the Potato Genome Sequencing Consortium (PGSC) at a global level using TAIR10 as an external standard (**Supplementary Fig. 34**) and in more detail, for specific gene families (**Supplementary Table 39**). These comparisons aimed at positioning our annotation results relative to the official potato annotation and the standard TAIR10 *Arabidopsis* annotation. From these comparisons it became clear that the iTAG annotation generated using the iTAG pipeline was superior to the published

PGSC annotation[93]. Although both pipelines predicted roughly comparable numbers of genes (39,031 genes in the PGSC annotation versus 35,004 genes in the iTAG annotation), only 8% of the iTAG potato genes did not give a hit to TAIR10, while 31% of the PGSC potato genes remained orphan compared to TAIR10 (**Supplementary Fig. 34**). Comparing the potato iTAG annotation and the iTAGv2.3 annotation for tomato shows only 3.6% potato specific genes compared to tomato while the reverse reports 12.3% tomato specific genes. The iTAG potato data is available on the FTP site provided for the iTAG tomato genome (see **Section 2.2**).

INFERNAL[94] was used to search for loci of non-coding RNAs. Using specific small-RNA targeted Illumina reads (see **Section 2.9**), 96 tomato and 120 potato conserved miRNAs were annotated and their targets identified.

2.2 Data availability

The sequence assembly, raw data, and annotation information, can be downloaded from:

- SGN website (http://solgenomics.net/)
- MIPS (http://mips.helmholtz-muenchen.de/plant/tomato/index.jsp)
- Genbank (http://ncbi.nlm.nih.gov)

  The whole project is accessible at AEKE00000000, while the individual chromosomes are accessible at CM001064 to CM001075.
- Mitochondrial genome sequence is accessible at www.mitochondrialgenome.org
- For expert/community curation of the annotation please visit http://bioinformatics.psb.ugent.be/genomes/view/Solanum-lycopersicum (mail beg-bogas@psb.vib-ugent.be to request an invitation to set up a personal account; no account is needed to browse the data)
- The FTP and WGP maps are available on:

  ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/physical_mapping/fpc

  ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/physical_mapping/wgp
- The RNA-Seq data are available under accession numbers: SRA049915 (Illumina) GSE33507 (454), SRA050797 (SOLiD) and SRA048144 (third party anonymized Illumina).

2.3 Methods and procedures implemented by iTAG

2.3.1 Masking of the genomic sequences

Prior to gene prediction, special attention was given to annotate all transposable element-related repeats that could give rise to gene models being purely transposable element or gene structures erroneously joined with remnants of those transposable elements. RepeatMasker[43] was used for masking the pseudomolecules with the custom library (see **Section 2.6**) of tomato and potato repeats. Before masking, the library of predicted transposable elements itself, was inspected for potential remnants of inclusions from genuine protein encoding genes, using TAIR9 CDS sequences to hard-mask the tomato-potato transposable element entries. The fraction of the tomato genome masked with the custom library was 63.28%. This masked genome sequence was used for the annotation. No simple repeats, rRNA, tRNA, or microsatellites were masked. The engine used for masking with RepeatMasker was WU-BLAST.

2.3.2 Small RNAs

Non-coding RNAs were annotated by searching against the Rfam[95] database (version 9.1). The search for RNA structure and sequence similarities was performed using INFERNAL[94] (version 1.0, ftp://selab.janelia.org/pub/software/infernal/infernal-1.0.tar.gz) with the Rfam database version 9.1[95]. To reduce the runtime, genomic sequences in the form of pseudomolecules were split in the middle of gaps of size larger than 1000 Ns, based on the contig-AGP file. The sequence coordinates in the rfam_scan,pl (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/9.1/MISC/rfam_scan.pl) output were converted back to the pseudomolecule level and the final output was merged into a single file per chromosome. As Rfam v9.1 was not compatible with INFERNAL 1.0, the special collection of Rfam v9.1 models built for INFERNAL 1.0 was used (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/9.1/infernal-latest.tar.gz).

The parameters used for rfam_scan.pl were: -f gff --nobig --bt 0.000001 (gff output format, skip the large ribosomal RNAs, BLAST E-value cutoff of $1*e^{-6}$). The rfam_scan.pl v0.1 was modified to match INFERNAL 1.0 in default search mode (from global to local). The gff2 output was converted to gff3 by a custom PERL script that also converts Rfam entry names to SOFA v1.93 terms (Sequence Ontology Feature Annotation)[96]. Finally, rfam_scan.pl jobs were submitted in decreasing genomic sequence length order to further reduce runtime. INFERNAL identified 1,853 non-coding RNAs of 90 distinct Rfam families. Almost 48% of all RNAs identified were tRNAs (RF00005).

The annotation of small RNAs, including miRNAs, was further extended taking advantage of RNA-Seq of small RNA libraries from tomato and potato (see **Section 2.4**)

2.3.3 Protein mapping

Protein data from TAIR9 was mapped using GenomeThreader[97] 1.4.3 (GTH) to maximize the contribution of the reference data from *Arabidopsis thaliana*. GTH allows mapping of proteins taking intron-exon boundaries into account. The output of GTH was reformatted into GFF3 and used as highly reliable data. Similarity with curated proteins from SWISSPROT where obtained through BLASTX. The output was passed on to EuGene after reformatting the BLAST output using custom PERL scripts. For the potato annotation, we also mapped the proteins from the curated iTAGv2.3 version of the tomato genome, using GTH (with the same parameters).

2.3.4 Third party *ab initio* predictions

GENEID v1.4 (http://genome.crg.es/software/geneid/) was trained on the largest scaffold (scaffold 2302) with manually inspected gene models from early iTAG runs. Extrinsic data for GENEID included mapped SOLiD reads and intron information extracted from spliced-mappings. The GENEID *ab initio* parameters were obtained by training for coding potential and splice sites and then further modified to optimize the weight given to read data versus the above described gene models. The read mapping positions and putative introns extracted from split-mapped reads were used by GENEID v1.4 to generate gene models having RNA-Seq support. Putative intron-spanning read mappings and gene models were contributed to the iTAG annotation pipeline.

AUGUSTUS[88] was trained on the tomato genome release 1.5 using 298,123 tomato ESTs (GenBank, May 2010). First, PASA[98] was used to construct a training set of more than 5,000 genes from assemblies of EST alignments, that are each likely to contain a full open reading frame. After an initial training of the coding regions using the autoAug.pl pipeline of AUGUSTUS, a training set for UTRs was constructed from EST alignments not correlated to predicted protein coding regions of AUGUSTUS. Subsequently, a UTR model was trained and the parameters optimized. On a disjoint evaluation set of 200 genes, initially withheld from the training set, the *ab initio* version of AUGUSTUS achieved on the protein-coding part of the genes a sensitivity of 92%, 82% and 43% on the base, exon and gene level, respectively. These figures may underestimate the true average accuracy, as the PASA gene set is likely to contain a substantial error rate as well.

Evrogen (www.evrogen.com) was contracted to construct a full length enriched, normalized (SMART) cDNA pool from a complex collection of RNA collected from mixed stages from 5 tissues (*S. lycopersicum* 'Heinz 1706' whole seedling, flowers, root, leaf and fruit). RNA from each tissue pool was used to construct cDNA from each RNA sample pool, which was then pooled followed by a directional normalisation that incorporated adapters carrying SfiI A and Sfi IB sites flanking the cDNAs, so that the cDNA 5' and 3' termini sequences could be identified within the 454 ESTs collection. Two full runs of the 454-FLX sequencer were performed by the University of Florida ICBR genomics core facility on the tomato cDNA pool, which produced ~ 174Mb of sequence represented within >767,856 sequences with average lengths of 226 bp. These sequences were screened through Lucy[63] with parameters: (-range 30 20 20 -alignment 8 12 16) to remove low quality regions and SMART primers, and to discard short (<25 bp) reads. Trimming and cleaning resulted in 675,271 454 tomato EST sequences. The collection of 454 tomato EST sequences combined with 223,441 publicly available Sanger sequenced tomato ESTs were aligned to an early tomato WGS assembly (ver 1.0.3), using PASA[98]. The PASA alignment resulted in a total of 58,314 assemblies, which reduced down to 20,253 after removing assembled transcripts that encode proteins less than 100 aa in length. An additional set of 9,175 assemblies/models were removed that did not contain complete predicted CDS, lacking start and/or stop signals. The predicted tomato protein sequences (query) from the remaining 11,078 models were aligned against the rice and *Arabidopsis* protein sets (subject) with WU-BLASTP (Ver 2.0, B=1, V=1 E=1*e^{-06}). Only those tomato query sequences whose best hit alignment was at least 50% identical to their best hits, and whose length-ratios of subject and query were $\geqslant$0.90, were retained as possible full length tomato proteins (4,574). This collection of proteins was clustered with a custom PERL script that uses WUBLASTP to greedily cluster sequences that share 90% identity over 90% of their lengths and outputs the longest representative of each cluster. The gene models associated with the obtained predicted proteins are retained for training the TWINSCAN_EST *ab initio* gene finder. This whole procedure resulted in 3342 models for training (**Supplementary Table 40**). TWINSCAN_EST[99] is based on TWINSCAN[90,100,101] and incorporates available EST alignments from the genome of interest in addition to integrating a traditional HMM based gene-prediction probability model with information from the alignments between two genomes (the genome being predicted on, and that of a closely related species). Genomic segments containing each of the 3342 predicted models flanked with up to 2 kb of sequence 5' and 3' were cut out of the tomato draft genome assembly. Training TWINSCAN_EST (Version 4.1.2) was performed as detailed in the TWINSCAN documentation (http://mblab.wustl.edu/software/TWINSCAN/) and testing was accomplished using the fourfold

cross validation described[102]. Custom PERL scripts were written to help automate the process. Briefly, each of these 3,342 tomato genomic sequences were WU-BLAST aligned to the *Arabidopsis* genome and the alignments converted to conservation sequences (conseq) (see[90,100]. Likewise, each of the 3342 tomato genomic segments were aligned to a public tomato sanger sequences EST collection with BLAT (http://genome.ucsc.edu/) and the alignments converted to estseq[99]. Training and cross-validation are performed by splitting the 3,342 gene models into 4 approximately equal groups consisting of the tomato genomic segment (FASTA), GTF representation of gene structure (the coordinates of each feature: start codon, first exon, first intron, …, terminal exon, termination codon) and the corresponding conseq and estseq. Three sets are used to train, while the fourth set (the leave out set) is used for prediction and the results compared to the known location of the gene models. This is performed 4 times, each time training on three sets and testing on the fourth. Performance was assessed in terms of the accuracy of predictions and gave rise to measurements of sensitivity and specificity (**Supplementary Table 41**). After fourfold validation to assess accuracy, the final parameter set is derived from training with the complete collection of 3342 models. Gene predictions were made by running TWINSCAN version 4.1.2 on assembly release 2.1 of the tomato genome sequence; see TWINSCAN documentation for details (http://www.mblab.wustl.edu/software/TWINSCAN/). The informant database was the TAIR9 *Arabidopsis* chromosome sequence collection (http://www.arabidopsis.org/). The EST database for estseq generation was a collection of 239,564 sanger sequenced tomato ESTs.

Generation of potato gene models for training followed the same procedure as generation of tomato gene models above with the following changes. PASA was used to align 206,565 potato ESTs to the 12 potato pseudomolecules. PASA alignment resulted in a total of 20, 251 assemblies, which reduced down to 9,740 after removing assembled transcripts that encode proteins less than 100aa in length. 3947 assemblies/models remained after removing those that did not contain predicted CDS start and stop signals. After alignment to known proteins to remove those that do not appear to represent full-length gene models, and final clustering, 1,555 putative potato gene models remained for training. Performance was assessed as above (**Supplementary Table 41**) and genome wide predictions were obtained by using the parameters derived from the full set of 1555 gene models, while the informant database used was the TAIR9 *Arabidopsis* chromosome sequence collection (http://www.arabidopsis.org/) and the EST database for estseq generation was a collection of 206,565 sanger sequenced potato ESTs.

2.3.5 Transcriptome sequencing

Tomato 'Heinz 1706' plants were greenhouse grown under 14 hr / 26 C day and 10 hr / 16 C night photoperiods. Three-week sand grown seedlings were harvested for roots and leaves. Mature plants were harvested for unopened flower buds (buds) and fully open flowers (flowers). Additional flowers were allowed to self pollinate and fruit were harvested at the 1 cM, 2 cM, 3 cM, mature green (MG), breaker (B, early ripening) and 10 days post B (B10, red ripe) stages. Similar leaf, immature fruit, B, and red (B5, B + 5 days) fruit were harvested from *S. pimpinellifolium*. All tissues were frozen in liquid nitrogen, ground to a fine powder with mortar and pestle and stored at -80^{o}C. RNA extraction and strand-specific RNA-Seq library preparations were performed as described[103] with 12 independently bar-coded samples multiplexed and sequenced on one lane of the Illumina HiSeq2000 system. Two independent biological replica samples were prepared from pooled organs of the same age and stage and sequenced for each tissue. Sequencing resulted in 7.4 – 14.0 million reads (48-53 bases long) per replica sample (**Supplementary Table 42**). The reads have been submitted to the NCBI Sequence Read Archive under the accession number SRA049915. For 454 RNA-Seq, plants of *S. lycopersicum* (cv 'MoneyMaker') were grown under long days at controlled temperature. RNA was extracted from root, stem, leaf, flower and fruit at 3 different maturation stages: mature green, breaker and ripe according to published methods[104]. cDNA was produced according to the SMART cDNA PCR synthesis kit (Clontech). cDNA libraries were produced following the 454 Titanium Rapid Library Preparation protocol instructions, then each sample has been loaded on a half picotiter plate to obtain at least 320,000 reads with a median length of 300 bases (**Supplementary Table 43**). The reads have been submitted to the GEO database under the accession numbers GSM828870 to GSM828878.

2.3.6 Mapping and tissue specific expression

Strand-specific Illumina RNA-Seq reads were first aligned to ribosomal RNA sequences using Bowtie[105] to eliminate possible rRNA sequence contamination. The resulting reads were aligned to the tomato genome sequences using TopHat[106]. Following alignment, for each gene model in iTAG annotation v2.3, the count of mapped reads from each sample were derived and normalized to RPKM (reads per kilobase of exon model per million mapped reads[107]. Differentially expressed genes between Heinz mature green and breaker, and mature green and red ripe fruits were identified using the DESeq package[108] and raw p values of multiple tests were corrected using false discovery

rate (FDR)[109]. Summary statistics regarding the alignment per library are presented in **Supplementary Table 42** and correlations between the two biological replicates of each tissue is listed in **Supplementary Table 44**, indicating that the expression values were highly reproducible between replicates (all > 0.97). Normalized expression (RPKM) for all 34,727 tomato genes in iTAG annotation v2.3 in all 14 tissues analysed is presented in **Supplementary Table 1**. The 10 most expressed genes per tissue are summarized in **Supplementary Table 45**. Similarly, the most differentially expressed genes in breaker fruits and leaves of *S. lycopersicum* and *S. pimpinellifolium* are summarized in **Supplementary Table 46**. The mapping of the 454 was done using GenomeThreader[97]. Hereby, duplicated reads were removed using BioPerl version of nrdb (bp_nrdb.pl) to reduce further the number of alignments to be generated. Also reads shorter than 50 nt were discarded. The whole procedure reduced the initial set to a total of 3,351,791 reads that were aligned on the genome using GenomeThreader v 1.4.3[97]. On the resulting 2,851,947 mapped reads, custom filtering was applied to remove reads that were: a) mapped to more than 3 regions or b) mapped in a non-spliced way and to a "lonely" region (no other reads within 1,000 nt range). Only the spliced alignments (1,421,438) were provided to EuGene.

2.3.7 Third party RNA-Seq data

Several international groups, listed in the Acknowledgments section, provided Illumina RNA-Seq data to assist annotation. In total, data from 47 tomato Illumina RNA-Seq anonymized libraries were shared. The reads supporting the iTAG gene models were submitted to the SRA database under the accession numbers SRA048144.1. 61 potato read libraries where also shared by PGSC[93,110]. The data were quality trimmed using the FASTX tools and redundancy removed. This resulted in 329,592,550 reads for tomato and 462,665,432 reads for potato. The mapping was done using Bowtie and TopHat. Default parameters were used (except that up to 3 mismatches and multiple mapping were allowed). For the annotation only spliced reads ("junctions") were used as input for EuGene as those were strand specific and would likely not map on transposable elements.

SOLiD RNA-Seq data were mapped to the assembly using GEM (http://sourceforge.net/projects/gemlibrary/). Of the 269,512,040 sequence reads, we were able to map 207,950,382 (77%), 2.8% of these spanning likely introns. SOLiD RNA-Seq reads were iteratively mapped using the GEM package as follows: 1) all reads were mapped to the assembly with gem-mapper with up to two mismatches 2) unmapped reads were mapped with gem-mapper

with up to three mismatches 3) unmapped reads were mapped with gem-split-mapper with up to three mismatches per split 4) unmapped reads were trimmed to 36 bp and mapped with gem-mapper with up to two mismatches and finally 5) unmapped reads were mapped with gem-split-mapper with up to two mismatches.

Split-mappings were allowed for distances between 30 and 500,000 and for the following consensus intron termini and their reverse complements: GT+AG, CT+AC, GC+AG, CT+GC, ATATC+AN, NT+GATAT, GTATC+AT, AT+GATAC. Reads mapping more than five times in the genome were removed, as were read mappings that did not cluster with any others. This was done by sorting the map positions and then for each mapping, determining whether it had an immediate neighbour within 500 bp. If not, the mapping was discarded.

Expressed Sequence Tags (ESTs) from different cultivated and wild tomato species (*S. lycopersicum*, *S. habrochaites*, *S. pennellii*, *S. lycopersicum* x *S. pimpinellifolium*) as well as from other solanaceous species (*S. tuberosum*, *S. chacoense*, *Capsicum annuum*, *C. chinense*, *Nicotiana tabacum*, *N. benthamiana*, *N. sylvestris*, *N. attenuata*, *N. langsdorffii* x *N. sanderae*, *Petunia hybrida*) were retrieved from the dbEST and from the Nucleotide/mRNA division of GenBank (release 011008). Each EST collection was pre-processed by using the RepeatMasker tool (http://www.repeatmasker.org/) combined with the NCBI's Vector database in order to identify and trim vector contaminations and with RepBase[44] for the masking of simple sequence repeats, low complexity sub-sequences and other DNA repetitive elements. Finally ESTs in each 'cleaned' data set were splice-aligned to the tomato genome sequences using GenomeThreader[97]. Alignments, with 98% identity and 95% length coverage were retained and the output formatted in GFF3 (http://www.sequenceontology.org/gff3.shtml).

2.4 Integration with EuGene

EuGene[91] is an integrative gene finding software that combines its own statistical models (encapsulated in trained Markov models and Support Vector Machine (SVM) based SpliceMachine site prediction) with other *ab initio* predictions, evidence of transcription, splicing and translation based on experimental data (EST, RNA-Seq, cDNA sequences) and similarities to known sequences (protein databanks). EuGene is one of the focal points of the iTAG pipeline, receiving data from many partners and responsible for the production of mRNA/CDS prediction. To facilitate the integration of EuGene inside the iTAG pipeline, all its input/output capabilities have been extended

to allow reading and writing fully compliant GFF3 files. EuGene has been extended to deal with multiple transcript sources (each with specific independent parameters) and to predict alternatively spliced isoforms. Furthermore, internal data-structures and algorithms have been revisited to better cope with large datasets in terms of space or running time.

2.4.1 Training EuGene specifically for tomato and potato

In order to train EuGene, several data sets were prepared, comprising:

- Coding regions, retrieved from BLASTX hits on the genomic sequences (these include full protein sequences and conserved protein domains) and intron sequences extracted from the mapped transcript data (see ESTs and RNA-Seq mapping) were combined to build an Interpolated Markov model (IMM) that will allow discrimination of protein coding regions from non-coding regions.

- Donor and acceptor sites to build SpliceMachine[111] models; these were extracted from the same mapped transcript data. Windows of 402 nt with a canonical splice site (GT or AG) in the middle were extracted (GT donor: 40617 pos. instances; AG acceptor: 41074 pos instances) Negative sets for cross validation were built in the same way, taking care negative instances never were included in any spliced alignment (donor:123079neg, acceptor 123750neg). The number of instances of GC-donors was not high enough to build a dedicated model, but the option to allow splicing over GC-donors was included, allowing EuGene to build gene models with GC-donors suggested by transcript data. The training of SpliceMachine was done using the provided PERL script. Several training rounds are performed sampling different subset of the training data in a ratio of 1000pos/10000neg instances. The best performing SVM models, after cross validation, are included in EuGene.

- Full length, manually checked gene structures were collected (150) in order to let EuGene assign proper weights to the different input data. The training itself was done in several steps aiming at maximal exploitation of the different sets of extrinsic data provided to EuGene. The first step focuses on training the *ab initio* models as those will allow the prediction of genes also when no extrinsic information would be available. Upon satisfactory prediction results from the *ab initio* cycle, *ab initio* parameters are frozen and extrinsic information is provided in successive new training rounds. The order in which the data is provided to the training cycles

depends on the added value and how wide spread the data can be used for the prediction. The most trusted data is incorporated in the last cycle (typically ESTs and RNA-Seq).

2.5 Renaming

Following the final structural gene prediction step, chromosome-based identifiers were assigned to all protein coding gene models and genomic loci. The naming scheme was adopted from the well-established *Arabidopsis thaliana* guidelines (http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp). Using these guidelines, emphasis was given to the stability of locus identifiers on specific chromosomes e.g. in case of structural changes to the gene model or a splice variant. Although the numbering of the locus identifiers (by sorting the identifiers by ascending numbers) reflects the order along the chromosome in most cases at the moment, this does not hold true in some special cases such as subsequent adding of ncRNA loci etc. In the future, possible genome sequence rearrangements can also break that order. Nevertheless we expect only local reversals due to the mature nature of the tomato genome sequence/assembly. The first digit separated from the locus identifier by a dot identifies the splice variant. The second digit denotes the model version number that will be incremented as soon as structural changes (e.g. by manual curation) were performed on that particular model. A tracking history of changes will be kept and provided by the participating tomato genome databases and repositories (SGN, BOGAS, MIPS). To indicate different genetic element types, protein coding loci identifiers carry 'g', transposable elements 't' and non-coding RNAs 'r' letters in the name.

2.6 Functional annotation of protein coding genes in the iTAG pipeline

2.6.1 Assignment of human readable descriptions (AHRD) and PhyloFUN Gene Ontology annotation

Human readable descriptions were assigned to 78% of tomato proteins, while 22% received a description of “unknown protein”. 42% or 14,565 annotations by AHRD fulfilled all three quality criteria: a) Bit score of the BLAST result is >50 and e-value is $<e^{-10}$. b) Overlap of the BLAST result is >60%. c) Top token score from lexical analysis is >0.5 (see below). 10% or 3,371 annotations additionally matched a GO term assigned by PhyloFUN. 3% or 1,070 proteins were

manually curated by annotators, and the AHRD description was replaced by the manually assigned description.

Using PhyloFUN (see below) and Interpro2GO we could assign 39192 GO terms to 19662 or 57% of 34727 tomato proteins (**Supplementary Table 47**). 84,95 or 24% were assigned GO terms by both tools, 2085 or 6% only by PhyloFUN, while 9,082 or 26% only by Interpro2GO. 106 unique GO terms were assigned by both tools, 1,454 were assigned by PhyloFUN only and 548 were assigned by Interpro2GO only. While Interpro2GO is more sensitive and can annotate more proteins, the PhyloFUN pipeline is more specific and can annotate more specific GO terms that are further down the GO hierarchy.

2.6.2 InterProScan

To infer functions for the protein-coding genes, we used InterProScanversion 4.5 (ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/RELEASE/4.5/iprscan_v4.5.tar.gz)[112] to scan protein sequences against the protein signatures from InterPro[113] (version 22.0). InterPro integrates protein families, domains and functional sites from different databases: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. InterProScan integrates the searching algorithms of all these databases. InterProScan identified 240,027 protein domains of 13,752 distinct domain types. 87% of the genes (30,148 out of 34,727 genes in total) have been assigned with at least one domain. The top 20 Pfam and Superfamily domains are plotted in **Supplementary Fig. 35**.

2.6.3 Automated assignment of Human Readable Descriptions (AHRD)

AHRD uses similarity searches and lexical analysis for Automatic assignment of Human Readable Descriptions to protein sequences. It utilizes (1) BLASTP[114] search results against the Swissprot[115], TAIR[116] and TrEMBL[115] databases, (2) domain search results from InterProScan[112] and (3) gene ontology (GO) terms[117] predicted by PhyloFUN (see below). The 200 top-scoring BLAST results (based on e-values) from each database search are chosen. These are scored based on alignment scores, expected quality of descriptions per database and a lexical scoring of individual “words” computed from their frequency in the descriptions of top-scoring BLAST results. Filters and corrections include a black list of uninformative words, e.g. ”hypothetical protein” or “similarity to”, a cutoff for declaring a query protein as “unknown” and a score for co-occurrence of words.

Additionally, predicted GO terms (see below) are used to preferentially select protein descriptions that use standard terminology as found in GO term descriptions. The highest scoring description is assigned to the query and the database accession of the hit protein added to enable evidence tracking. If InterProScan results are available, the domain names are extracted and appended to the description line. In case of multiple InterPro matches the InterPro Parent-Child-Tree is used to reduce the number of reported domains by selecting the most specific child and excluding its parents. The following database releases and software versions were used: NCBI BLASTP 2.2.21 and 2.2.24; Swissprot and TrEMBL release 2010_06; TAIR9 protein database; InterProScan version 4.5; InterPro database version 22. The descriptions contain, in brackets between the transferred human readable description and InterPro domain results, if any, the keyword AHRD and version, followed by a four-character quality code and the database accession of the BLAST hit from which the human readable description was transferred. This is included to make it obvious that the description was assigned by an automated procedure and to allow tracing evidence and evaluating the reliability of the description. The quality code is composed of four characters, each being one of "-" (criterion not fulfilled) or "*" (criterion fulfilled). The criteria are, in order: a) Bit score of the BLAST result is >50 and e-value is $<e^{-10}$; b) overlap of the BLAST result is >60%; c) top token score from lexical analysis is >0.5; d) words found in the description are also found in annotated gene ontology terms.

2.6.4 Automated assignment of human readable descriptions to gene families

In order to automatically assign human readable descriptions to gene families derived from OrthoMCL clustering, InterProScan[112] results for all family members were utilized. InterPro database version 30.0 was used. The most frequent InterPro match of type family is used as the protein family's description if the frequency is greater than or equal to 0.5. In cases where the above threshold is not met, the most frequent InterPro match of type family is assigned, combined with the most frequent InterPro match of another type. The description line consists of one to several lines, consisting of the frequency of the InterPro match, the AHRD score, the InterPro match ID and type (e.g. Domain, Family, etc.), and finally the description of the InterPro match.

2.6.5 PhyloFUN pipeline

BLASTP searches against a selected set of whole proteomes with experimentally verified GO annotations were performed and phylogenetic trees computed as described in the supplement[118]. SIFTER[119] was used to transfer GO terms within the phylogenetic tree. We report all GO assignments that receive a Sifter-score above 0.4. Whole proteomes were retrieved from Uniprot[115], GO annotations from Uniprot, TAIR[116] and GOA[120]. The species included were *Agrobacterium tumefaciens, Anaplasma phagocytophilum, Arabidopsis thaliana, Bacillus anthracis* Ames, *Bos taurus, Caenorhabditis elegans, Campylobacter jejuni, Carboxydothermus hydrogenoformans, Clostridium perfringens* ATCC, *Colwellia psychrerythraea* H, *Coxiella burnetii* RSA , *Danio rerio, Dehalococcoides ethenogenes , Dictyostelium discoideum, Drosophila melanogaster, Ehrlichia chaffeensis* (strain Arkansas), *Escherichia coli* (strain K), *Gallus gallus, Geobacter sulfurreducens, Homo sapiens, Hyphomonas neptunium, Leishmania major, Listeria monocytogenes* (serovar b, strain F), *Magnaporthe grisea, Methylococcus capsulatus, Mus musculus, Neorickettsia sennetsu, Oryza sativa, Plasmodium falciparum, Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas syringae, Pseudomonas syringae* tomato, *Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Shewanella oneidensis* and *Silicibacter pomeroyi.* In parallel, GO annotation was performed using Interpro2GO[121]. Results were merged with the SIFTER annotations and duplicates removed.

2.6.6 Expert curation

To allow the community to access the data for expert curation of the automatic gene prediction and function description, the BOGAS system (http://bioinformatics.psb.ugent.be/webtools/bogas/) was set up to host the Tomato genome. The system allows a registered user to directly work and modify the data via a web interface. The system keeps a history of all changes made to the data and has a build-in alert system to follow specific genes. The platform displays multiple pre-computed analyses to help expert annotators to evaluate the quality of the gene models and validate the functional description. Some fields are to be updated textually, while gene structures can be manipulated through the graphical interface provided by GenomeView (http://genomeview.org). Upon modification applied to gene models, the system automatically updates all the pre-computed data allowing further curation.

From the originally 36,287 predicted genes, ~1000 genes were reclassified as transposable elements that escaped the masking. A remaining ~500 genes were reclassified as pseudo-genes or discarded in favour of better gene models built by the expert. The current status reports 34,727 predicted

genes, from which 1,346 genes underwent an expert intervention (3.8%) whether it was to complement/correct the functional description or correct erratic/incomplete gene models.

2.6.7 Localisation of genes with phenotyped mutants

Tomato cloned genes with a phenotyped mutant were mapped on the assembly and searched within the tomato annotation: precise chromosomal localisations could be identified for most of them (**Supplementary Table 14**).

2.7 Genomic distribution of small RNAs

We compared the genomic distribution of small RNAs in the genomes of *Arabidopsis*, tomato and potato (**Supplementary Fig. 2**). *Arabidopsis* has previously been reported to have a pericentric localisation of small RNAs that correlates well with repeat density[122]. We performed the same analysis on the tomato and potato genomes, in each case using the normalized abundances divided by the number of matching genomic locations ("hits-normalized abundances" or HNA) (**Supplementary Fig. 2**). The *Arabidopsis* distribution was consistent with prior reports, while the tomato and potato distributions showed a different distribution, with higher abundances in distal euchromatic regions and reduced abundances in pericentric regions. **Supplementary Fig. 2** shows data for only one library, but the same distribution was observed for each of the three tomato and potato libraries. A comparison of the small RNA data with the genome annotation and RNA-Seq data demonstrated a direct relationship on every tomato chromosome between the abundance of small RNAs (expressed in hits-normalized abundances (HNA)) and those of genes and RNA-Seq data (**Fig. 1**, **Supplementary Figures 1** and **2**).

We hypothesized that a small RNA density distribution different from that of *Arabidopsis* might be attributed to a different distribution of high and low-copy repeats. In order to examine this hypothesis, we focused on a region of chromosome 3 in which a sharp boundary was observed between high and low densities of small RNAs (**Supplementary Fig. 36**). In this region, we observed that the high HNA small RNA density primarily derives from non-annotated, intergenic sequences which were not identified by RepeatMasker and which are relatively low copy (as determined by a k-mer analysis). In contrast, the regions with an apparently low-density of small RNAs were associated with very high copy-number retrotransposons, identified by RepeatMasker predominantly as Gypsy-class LTR retrotransposons (**Supplementary Fig. 36**). If the small RNAs

were not normalized for the number of genome-matched locations ('hits'), the total abundance of small RNAs associated with these high-copy LTR retroelements was much higher than the small RNA-producing loci in the genic regions. Thus, it is likely that small RNA-producing loci in both genic and high-copy LTRs are heterochromatic. However, the tomato and potato genomes are different from that of *Arabidopsis* because of their segregation of high-copy repeats into discreet, non-genic subgenomic regions, resulting in the observed difference in the distribution of small RNA densities.

Note: For the small RNA genome distribution images (**Supplementary Fig. 2**), the *Arabidopsis* data used were from GenBank's GEO record GSM280228. The tomato data were from the SLY3 (breaker-stage fruit) library and the potato data were from the STU3 (stolon) library, both of which were also used for the miRNA analysis (see following Section). In each genome image, the sum of hits-normalized small RNA abundances was calculated and displayed for bins of 5 kb.

2.8 Mapping of sRNA reads to promoters of protein coding genes

Supplementary Fig. 37 shows the expression profile during fruit development of sRNAs that map to promoter regions of protein coding genes. The promoter regions were divided into 100 bp fragments and the read number of sRNAs mapping to each 100 bp fragment were combined. It is striking how specifically the sRNAs map to short regions in the promoter. Usually only a 100-200 bp region produces a significant amount of sRNAs and this is extremely reproducible across the ten samples. The other interesting feature of promoter mapping sRNAs is the dynamic expression profile during fruit development. The top two panels show promoters where the sRNAs are expressed at a high level in the flowers and developing green fruits and after the mature green stage (T6) the level of sRNAs decreases dramatically. On the other hand, the bottom two panels show promoters where sRNAs are expressed at a very low level in flowers and developing green fruits but after the mature green stage there is a significant increase in the level of sRNAs. It is worth pointing out that the majority of these promoter mapping sRNAs are 24mers (blue colour), which are known to mediate methylation or de-methylation of DNA. Therefore, they have the potential to impact gene expression in either way. In addition to these, the bottom right panel shows strong expression of a group of sRNAs in flowers that are mainly 22mers. The function of 22mer sRNAs is largely unknown. They are produced by Dicer Like-2 but mainly from viral RNAs. It is remarkable that different size-class sRNAs show opposite expression profiles on two different regions of a putative promoter sequence. The exact function of sRNAs mapping to promoter regions remains to

be determined. Nevertheless, the tomato genome allowed us to examine and characterize a potential layer of gene expression regulation by sRNAs through the promoter and that seems to be extensive. This opens up a new avenue of research that will address the biogenesis, regulation and function of these promoter mapping sRNAs.

Supplementary Fig. 38a shows the distribution of offset fold change (OFC) computed for differentially expressed promoter mapping sRNAs. We identified 2,687 promoters with sRNAs that are differentially expressed during flower-fruit transition and fruit development, but the temporal distribution of differentially expressed sRNAs is not random. Large numbers of promoters produce sRNAs that are differentially expressed at key developmental transitions, such as from flower to fruit or from fruit growth to fruit ripening, but very few promoters produce sRNAs that are differentially expressed during fruit growth or ripening. **Supplementary Fig. 38b** demonstrates that there is a correlation between mRNA expression and production of sRNAs from the corresponding promoter. The top panel shows the expected Gaussian distribution of correlations between mRNA and accumulation of sRNAs mapping to promoters of all genes on the Affymetrix whole genome array[123] that could be confidently mapped to the tomato genome. The distribution of correlations is different for the genes which produced differentially expressed promoter mapping sRNAs.

Note: The genome mapping reads were normalized using the "per total" method, with a fixed total of 1 million reads[124]. The expression levels of mRNAs (Tomato Affymetrix Chip) were processed using the Affy package[125] in Bioconductor[126]. The method used for background correction was RMA and the normalisation was done using quantile normalisation[127]. The differentially expressed series (mRNA/sRNA) were determined using the offset fold change on expression levels in linear scale[123]. The offset fold change was computed on the maximum and minimum expression levels in the series, using offset=20. OFC_max_min=(max+offset)/(min+offset). Small RNA and Affymetrix chip datasets used here were described elsewhere[123].

2.9 Identification, mapping and expression analysis of conserved miRNAs in tomato and potato

We deployed a genome-independent approach to identify and quantify known miRNA sequences in different tomato and potato tissues, followed by mapping of their precursors on the two respective genomes. For both analyses, we used three small RNA libraries constructed from *S. lycopersicum* 'Heinz 1706' (SLY1, SLY2 and SLY3 from leaf, pre-anthesis flower and fruit at the breaker stage,

respectively) and three small RNA libraries from *S. tuberosum* (STU1, STU2, STU3 from fully grown leaf (Kennebec cultivar), mature flower (mixture of Kennebec, Katahdin and Pentland Ivory cultivar) and stolon (Defender cultivar). Small RNA libraries were sequenced with Illumina technology yielding approximately 4 million reads per library.

For the identification of conserved miRNAs, we downloaded 3,941 mature plant miRNAs from miRBase (release 17.0, April 2011). Identical miRNA sequences identified in different species or duplicated loci in a genome were collapsed, resulting in a non-redundant list consisting of 1,772 unique miRNAs. Sequences belonging to the same miRNA family were further analysed by multiple alignment using ClustalW (www.clustal.org) and classified in subgroups (S01, S02, S03, …) to distinguish *bona fide* mature miRNAs from misannotated miRNA* forms or sequences generated from different regions of the same precursor. This non-redundant library was then used to screen the tomato and potato small RNA libraries. All the small RNA reads in the range of 20 to 24 nt in size and represented by at least 2 reads in a library were aligned to the 1,772 unique miRNAs derived from miRBase. For the screening, a maximum of 3 mismatches was allowed and up to 2 nt overhanging nucleotides at the 5' and/or 3' end. Alignments were performed using SeqMap[128]. The output was filtered and reformatted with custom PERL scripts, classifying the identified miRNAs according to miRBase. A second round of alignments was performed to score the amount of sequence divergence from miRBase entries to ensure correct sorting between very similar miRNA families. Customized PERL scripts were used to create HTM heatmaps summarizing the information for all the miRNA families at once, showing either the sum of abundances of all the variants (**Supplementary HTM Table 1)** or the abundance of the most frequent variant for each miRNA family (**Supplementary HTM Table 2)** across the six small RNA libraries. Heatmaps were also created for each miRNA family or subgroup, showing the abundance of each identified miRNA variant in the investigated tissues (**Supplementary HTM Tables 3-269**). The expression pattern of the 33 most significant and abundant miRNAs found to be conserved in tomato or potato is displayed in **Supplementary Fig. 39**. The HTM Tables (**Supplementary HTM Tables 1-269**) depicting the heatmaps can be opened with a web browser. Library labels and abundances can be visualized by mousing over the heatmap cells. In **Supplementary HTM Tables 3-269**, green cells show highly annotated, mature miRNAs, while sequences in white cells indicate close matches to the annotated miRNA. Abundances of each variant are indicated as TPM (transcripts per million) in the heatmap. Sz = size in nucleotides of the small RNA.

Our analyses confirm that miR156, miR166 and miR168, known to be conserved across all the higher plants investigated so far[129], were highly expressed in tomato. Among the most abundant

miRNAs, we also found the Solanaceae-specific miRNA family miR5300. In contrast, another miRNA first identified in tomato, miR5301[123], was significantly expressed above background level (~ 10 TPM) only in potato. In addition to miR5301, also miR479, miR482, miR1919, miR2089, miR2118 and miR5300 appeared to be differentially expressed between the two species (**Supplementary Fig. 39**). miR482, miR2089 and miR2118 were predicted to target TIR-NBS-LRR genes, suggesting differential silencing of resistance genes in the two solanaceous species. The miR894 family shows high similarity to ribosomal RNA and was not included in the analyses. All these miRNAs were preferentially or exclusively expressed in one or more potato tissues. Moreover, *in silico* target prediction, performed using the CleaveLand v.1 pipeline described in Addo-Quaye *et al.*[130], revealed a clear prevalence of NBS-LRR resistance genes among the target genes predicted for miR482, miR2089 and miR2118 (**Supplementary Tables 48-49** for tomato and potato respectively). This prevalence was more pronounced in potato than in tomato.

For the identification of known miRNA loci, we aligned the miRNAs from the non-redundant library discussed earlier to the chromosomes of the two genomes, allowing up to three mismatches in the alignment, using SeqMap[128]. From every locus, we extracted two ~200-nt regions surrounding each aligned miRNA (from -30 to +160 and from -160 to +30 nucleotides relative to the putative miRNA start and end coordinates, respectively). Minimum energy RNA secondary structures were predicted for each region using the RNAFOLD program of the VIENNA RNA 1.8.4 package (http://www.tbi.univie.ac.at/~ivo/RNA/) using default settings. In addition, small RNAs from the different sequenced libraries were mapped on these regions, allowing no mismatches, in order to pre-select putative miRNA loci that showed evidence of expression in the three plant tissues analysed. We evaluated RNA structure and small RNA alignment in all the regions based on: s) dominance of plus-stranded small RNAs; b) position of most abundant small RNAs relative to the predicted miRNA coordinates; c) prevalence of 20-22 nt small RNAs in the predicted miRNA locus; d) position of putative miRNA across the stem-loop structure and; e) absence of oversize (≥3 nt) bulges in the miRNA/miRNA* alignment. After reduction of overlapping loci to a non-redundant set and removal of stem-loop structures with the wrong orientation compared to miRNAs registered in miRBase, we manually inspected the remaining loci for further evaluation according to the miRNA annotation criteria proposed previously[131]. Stringency was relaxed when small RNA expression data strongly indicated the presence of miRNA loci that did not meet the whole set of criteria.

Following the above procedure, we identified 96 and 120 conserved miRNA loci in tomato and potato respectively (**Supplementary Fig. 40** and **Supplementary Table 50**). Mapping data

(**Supplementary Tables 51-52**) were generally consistent with the pattern of presence/absence suggested by the heatmap in **Supplementary Fig. 39**; only the precursors of miR858, miR894, miR2089, miR2911 and miR5054 could not be validated in either genomes, while miR479 and miR5301 were mapped in the opposite arms of miR171 and miR482 stem-loop precursors respectively. It should also be noted that the approach we used was strongly based on small RNA expression data. Therefore, differences in miRNA expression between species and tissues may have an effect on the number of genomic loci that could be identified with high confidence, especially if a given miRNA is preferentially expressed in organs, developmental stages or environmental conditions not covered in this study.

A comparison between plant sequenced genomes using a subset of conserved miRNA families (**Supplementary Table 53**) indicated that the number of mapped loci is consistent with that of some model and non-model species (e.g. *Arabidopsis*, *Brachypodium*, soybean, strawberry, cacao) but roughly half as much as in some monocot crops (rice, sorghum and maize) and other eudicots surveyed thus far (e.g. grape and poplar). Considering that the several miR5303 loci found in potato are balanced by an excess of miR395 genes in tomato, the larger number of *loci* mapped in potato is primarily explained by an increased number of genes encoding other miRNA families (miR160, miR171, miR172, miR390, miR399, miR482 and miR1919) as well as by the presence of miRNA families that have been lost or are strongly reduced in expression in the tomato species/tissues analysed (miR319, miR397, miR403, miR530, miR1436, miR2111 and miR2118). The number of mapped loci is largely compatible between tomato and potato for a subset of miRNAs that are ubiquitous or widely conserved in plants (miR156, miR160, miR162, miR164, miR166, miR167, miR168, miR169, miR172). Predicted targets for these miRNAs were identified in the two genomes using the CleaveLand pipeline (**Supplementary Tables 54-55**). Among putative targets that will require experimental validation are gene functions that are known to be regulated via the same miRNA pathways in different species. Therefore, our observations confirm that the basal network of miRNA-mediated gene regulation common to several if not all land plant species is generally well conserved in the Solanaceae.

2.10 Transposon and repetitive sequence detection and annotation

Full length LTR retrotransposons where detected *de novo* with the program LTR-STRUC[132] on the tomato v2.40 assembly and on the potato pseudomolecules (version 1). The candidates from the LTR-STRUC search were subjected to a quality check (at least one LTR specific inner protein

domain, <30% simple sequence content, typical LTR dotplot picture) resulting in 1,647 intact LTR-retrotransposons for tomato and 1,309 for potato. The intact LTR-retrotransposons where assigned to the gypsy (RT,RH:IN) or copia (IN,RT,RH) subgroup by the order of their inner protein domains. A large amount remained unclassified for now, because not all inner domains could be identified with the automated procedure.

Additional full length LTR elements were found by homology, which lead in total to 4,052 still intact elements for tomato and 2,173 for potato. A cluster analyses (⩾80% identity over ⩾80% sequence length) of these sequences showed, that both species share common LTR-retrotransposons: 10% of the tomato LTR elements also occur in potato and 5% of the potato elements have a counterpart in tomato.

The insertion events of LTR-retrotransposons where dated by the sequence divergence between left and right solo LTR, which stem from identical copies created during the transposition event[133] (**Supplementary Fig. 41)**. In contrast to *Sorghum* (C), where a continuous rise towards recent insertions is observed, the two Solanaceae (A,B) have a different overall age pattern, with reduced young elements and a shift towards older amplification peaks. The average insertion age for *Sorghum* is 0.8 my, for the two *Solanum* species it is 2.8 my. In addition tomato has fewer full length elements (~4.000 vs. ~11.000) and fewer large sized family clusters (5 vs. 18 clusters ⩾ 50 members) than *Sorghum*. These results fit to the lower amount of repetitive sequences seen in the k-mer analyses (**Supplementary Fig. 42**). Both tomato and potato have a similar differential age distribution of the two LTR-retrotransposon subgroups gypsy and copia (A, B). Copia elements are younger (m50, 1.7 my, 1.2 my) with a downward slope towards older insertions. Gypsy elements are older (m50, 2.2 my, 3.6 my), with past maximum activities and few recent insertions.

Transposons where annotated by the WU-BLAST version of RepeatMasker[43] against the 'dicots' section of mipsREdat (REdat_v8.9_Eudico, 40 Mb, 8,193 entries). This transposon library is connected to a repeat classification scheme (mips_REcat) and contains a collection of known transposons as well as *de novo* detected LTR-retrotransposons from tomato (1,647) and potato (1,309). The RepeatMasker output was subjected to two post-processing filter steps: a) removal of low confidence hits (length <50 bp, score ⩾255) and b) cleaning of overlapping annotations in a priority based approach, where higher score hits where assigned first and overlapping lower score hits either shortened or if the overlap exceeded 80% of their length removed. The obtained transposon annotations from tomato and potato are summarized and compared in **Supplementary Table 56**. The three letter transposon type code of this table is in accordance with an official

transposon classification[134]. Retrotransposons constitute the biggest portion of the tomato (62%, 460 Mb) and potato (53%, 311 Mb) genome. DNA transposons account for only ~1% in both genomes. Their number is to be seen as lower limit, since they were only identified by homology to known transposons, without additional *de novo* detection. Overall the main and subtype proportions are similar, there are no extreme discrepancies between the two genomes. The observed variations are summarized as x-fold bar chart. The most obvious difference between tomato and potato is the 9% lower transposon content of potato, which is caused by a higher amount of LTR-retrotransposons in tomato. This may be due to a real difference and/or caused by missing highly repetitive regions in the assembly of the potato pseudomolecules. Another large difference is the distinctly higher amount of copia LTR-retrotransposons in tomato.

Tandem repeat sequences where detected with the program Tandem Repeats Finder[135] under default parameters. The tandem repeats where classified by their monomer length into microsatellites (2-9 bp), minisatellites (10-99) and satellites (⩾100 bp). Overlapping annotations where joined and classified as hybrid type, if they contained more than one of the 3 classes. **Supplementary Table 57** shows the tandem repeat composition of the tomato genome.

K-mer frequencies are a repeat library-independent and thus unbiased method to access the repetitive portion of a genome. The program Tallymer[136] from the program suite GENOMETOOLS (http://genometools.org/) was used to calculate the frequency for each 16mer in the tomato and potato sequences (**Supplementary Fig. 42**).

Heat-maps and stacked bar-charts are used to visualize and compare chromosomal structures. The heatmap data were created by going along the chromosome in a sliding window (0.5 Mb window size with 0.1 Mb shift) and determining for each window the number and percent bp coverage of the respective element type, like genes or LTR-retrotransposons. The density values where corrected for the number of Ns per window, if the N content exceeded 60% the value is set to null and drawn in white colour. The number value was extrapolated to number per Mb to facilitate comparisons. The heatmaps where created from the obtained density values using the python pylab module in combination with the jet colour map (low to high values from blue to red).

2.11 Genome compositional features of tomato compared to other plant species

The average GC content of the *S. lycopersicum* genome is 34% (s.d.= ±2.29) (**Supplementary Table 58**). Chromosome *0* shows a higher GC content than the other chromosomes, which have

similar compositional properties.. The GC distribution was also compared to that of two other dicots, *Arabidopsis thaliana*[137] and *Vitis vinifera*[138], and one monocot, *Oryza sativa (japonica)*[139] respectively. Close average GC values are evident for dicots (**Supplementary Table 58**), whereas rice shows a remarkable higher value, confirming previous experimental investigations in which the nuclear genomes of Poaceae were found GC-richer than those of dicots[140-142]. A similar difference between eudicots and monocots is found when considering the GC content of the genes as well as the one of exons and introns (**Supplementary Table 58**). The lower GC level of introns compared to exons is a general feature of eukaryotic genes. The *Arabidopsis thaliana*[137] TAIR9 genome release, the *Vitis vinifera* PN40024[138] 12X coverage genome release and *Oryza sativa (japonica*) genome release 6.1[139] were downloaded from the TAIR website (http://www.arabidopsis.org/), Genoscope (http://www.genoscope.cns.fr/spip/) and Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/) respectively, together with their respective annotations. Sequences were analysed using a non overlapping window of 100 kb. The GC content (GC%) of these sequences was calculated using an in-house developed script, which is also suitable for window based compositional analysis.

3 THE GENOME OF *SOLANUM PIMPINELLIFOLIUM*

3.1 Plant material

Sixth generation inbred seedlings of *S. pimpinellifolium* (accession LA1589) were germinated and grown in the dark for two weeks and harvested in bulk for DNA extraction using a standard CTAB protocol with modifications for nuclear enrichment[143]. Illumina sequencing libraries of insert size 300-400 bp were prepared from 10 µg of genomic DNA according to the manufacturer's instructions. For the genotyping of *S. pimpinellifolium* accessions, at least four seedlings were pooled for a single CTAB DNA extraction for each accession. DNA was diluted to 20 ng/µl and used for PCR.

3.2 *De novo* assembly of *S. pimpinellifolium*

A *de novo* assembly of the wild tomato genome was generated from 552 million 101 bp paired-end Illumina reads comprising over 55 Gb of raw DNA. A pre-screen that removes reads with ambiguous bases (a requirement for assembly) yielded 389 million reads comprising 39 Gb, representing approximately 40X read depth of the genome (**Supplementary Table 59**). The assembly was generated with ABySS (version 1.1.2), which is an open-source, parallel sequence assembler[144]. The algorithm first builds a distributed de Bruijn graph over all input k-mers of a specified length. The graph is then decomposed into unique paths, each representing a *de novo* contig. A k-mer length of 50 was chosen based on initial observations of coverage distributions at varying k-mer values in the input data. A second stage of the assembler uses mate-pair information to build larger scaffolds over the contigs and to resolve ambiguities. The assembly ran in 3 days on a dedicated 32 CPU high-memory machine. The final assembly yielded 739 Mb of DNA in ~627,000 contigs with an N50 contig length of 5,100 bp. The assembly includes over 136,000 contigs 1 kb or longer and over 13,000 contigs 10 kb or longer (**Supplementary Fig. 43**).

To assess the validity of the assembly, the 34,724 coding sequences from the available gene models annotated on the domesticated *S. lycopersicum* 'Heinz 1706' genome (release ITAG 2.3) were aligned to the *de novo* assembled contigs by BLAT[145]. 34,092 genes (98.2%) mapped to the *de novo* contigs with at least 5% hit coverage (≥95% identity), indicating that most tomato coding regions are "touched" by the assembly. Applying more stringent criteria, the analysis indicates that 17,308

'Heinz 1706' genes (49.8%) map to a *de novo* contig at full coverage (≥98% identity; **Supplementary Fig. 44**). Of these fully covered genes, 5,819 (33.6%) align to 2,654 contigs that harbour two or more such genes, providing gene co-linearity and genomic context for further investigation.

3.3 Polymorphism detection and putative *S. pimpinellifolium* introgressions into 'Heinz 1706'

A total of 389M high quality paired-end Illumina reads, each 101bp in length, used to identify variation between the *S. pimpinellifolium* and *S. lycopersicum* genomes. A two-stage approach was used to map reads to the *S. lycopersicum* 'Heinz 1706' assembled chromosomes (release 2.40). Bowtie[105] was first used to map the reads (parameters: " -a --best --strata -S -m 1") retaining alignment results for reads that map uniquely. The remaining unmapped reads were then put through Novoalign (www.novocraft.com), a sensitive aligner that accepts longer indels in the reads[146] (k-mer length = 14, stepsize = 2, default options otherwise). Over 68% of the reads mapped uniquely, 12% mapped to multiple locations and 20% remain unmapped. The unmapped fraction consists mostly of poor quality reads as well as genome segments missing from the *S. lycopersicum* reference, such as transposons and other repetitive elements.

Using the SAMtools package[147], alignment results were merged and indexed as BAM files. Coverage and variant discovery were done by converting the BAM alignments into pileup files. Polymorphic sites were identified at positions where there was at least 5X coverage, the non-reference allele was detected in over 80% of the reads mapping to the locus, and the non-reference allele had an average Phred score ≥ 15 (**Supplementary Tables 5** and **6**). *S. lycopersicum* gene models were imported into an EnsEMBL database scheme[148] and the EnsEMBL API was used to annotate the variation. The EnsEMBL API was also used to produce FASTA dumps of each CDS model, while the corresponding *S. pimpinellifolium* model was constructed using the information from the pileup files. The Nei-Gojobori[149] estimation of *Ka/Ks* for each *S. lycopersicum* - *S. pimpinellifolium* gene pair was then computed using Yn00 in the PAML package[150].

Putative *S. pimpinellifolium* introgressions into 'Heinz 1706' were detected by comparing intraspecific genetic variations within domesticated tomatoes to interspecific divergence between domesticated and wild tomatoes. Specifically, regions with high nucleotide diversity between domesticated tomatoes but low divergence between *S. pimpinellifolium* and 'Heinz 1706' indicate

potential admixture between 'Heinz 1706' and *S. pimpinellifolium*. To detect putative introgressions of *S. pimpinellifolium* into the genome of 'Heinz 1706', we queried the spatial distribution of SNPs between the two genomes and identified 50kb genomic windows showing low SNP densities (lower than one standard deviation below the genome-wide average SNP density). We set a threshold of three or more consecutive 50kb windows showing low SNP density for calling a putative introgression. We then utilized Illumina mRNA-sequencing transcriptome data from another processing tomato cultivar called 'M82' to detect regions of higher intraspecific variation relative to the variation between *S. pimpinellifolium* and 'Heinz 1706'. The 'M82' transcriptome data set consisted of 341,347,384 high quality paired-end 50bp reads, which are is available for download from the Solanaceae Genomics Network hosted at the Boyce Thompson Institute (ftp://ftp.solgenomics.net/transcript_sequences/by_species/Solanum_lycopersicum/libraries/illumina/LippmanZ/). The reads were mapped to the predicted CDSs of 'Heinz 1706' using the BWA read mapping program with default parameters[151], and nucleotide variations were called from the BAM format alignments using the samtools/bcftools pipeline with default parameter settings[147]. The resulted vcf file was used to generate reconstructed CDSs of 'Heinz 1706' using the FastaAlternateReferenceMaker application implemented in the GATK package[152]. A total of 19,611 CDSs were reconstructed with sufficient coverage to calculate gene-by-gene nucleotide diversity values (π) between 'Heinz 1706' and 'M82' CDSs after ClustalW realignment. Finally, average nucleotide diversity of genes within the genomic intervals of low SNP density was calculated. In **Supplementary Table 9**, intervals of low SNP density that also showed higher than average nucleotide diversity between 'M82' and 'Heinz 1706' CDSs are highlighted in yellow, and represent the most supported putative introgression of *S. pimpinellifolium* into the genome of 'Heinz 1706'. Interestingly, we detect large introgressions on both chromosomes 9 and 11, and both chromosomes have been implicated in the breeding of disease resistance loci into 'Heinz 1706' using *S. pimpinellifolium* germplasm[29]. Several other introgressions are detected throughout the genome, including ~1Mb introgressions on chromosomes *4* and *12*.

3.4 Diversity and Gene Ontology (GO) term enrichment analysis

Genes showing high *Ka/Ks* value relative to *S. lycopersicum* 'Heinz 1706' (≥1, an indication of possible positive selection; **Supplementary Fig. 45**) and genes identified as having lost or gained stop codons were subjected to GO term enrichment tests, including the hypergeometric test in the

GOstats package in R[153], and Fisher's exact test in Blast2GO[154,155]. The biological functions of enriched genes were summarized from both types of tests. Based on the reconstructed coding sequences and SNP calls, 862 gene showed *Ka/Ks*≥ 1, 380 genes lost stop codons, and 854 genes gained stop codons in *S. pimpinellifolium*. Genes showing high *Ka/Ks* values are enriched with genes involved in cell death, apoptosis, proteolysis and defence response processes ($P \leq 0.01$). Genes that have lost stop codons are enriched for functions involved in cell death, apoptosis and transcription ($P \leq 0.01$). Genes with stop codons gained are enriched for functions involved in proteolysis and other metabolic processes ($P \leq 0.01$) (**Supplementary Table 7**).

3.5 Identification of unique genomic regions in *S. lycopersicum* 'Heinz 1706' and *S. pimpinellifolium*

Regions present in the *S. lycopersicum* 'Heinz 1706' genome but absent in the *S. pimpinellifolium* genome were identified as follows. First, 31-mers were extracted from the Illumina reads of both species (see also "Base error correction with Illumina reads"), and the frequency of each 31-mer in these reads of each species was counted. All 31-mers with a frequency of two or lower were assumed to represent sequencing errors and discarded. Next, the coverage of each base in the *S. lycopersicum* assembly by both these 31-mer sets was determined. For each of the two 31-mer sets, the coverage of each base in the assembly ranged from 0 (meaning that no 31-mers covering this position exist in the corresponding Illumina data set) to 31 (meaning that all possible 31-mers covering this position exist). This coverage was then normalized to account for small sequence polymorphisms, such that a coverage of 0 to 11 would be 0; 11 to 21 would be 1, and 21 to 31 would be 2. Unique regions in the *S. lycopersicum* genome were then defined as contiguous regions of 1,000 nt or more where the normalized coverage of *S. pimpinellifolium* 31-mers equalled 0, and the normalized coverage of *S. lycopersicum* 31-mers equalled 2. In total, there were 3,423 such regions with a total length of 7.3 Mb. While the majority of these regions were distributed evenly over the genome, two megabase-sized regions on chromosomes *1* and *10* displayed a clear abundance of sequence unique to *S. lycopersicum*. We verified these findings by PCR amplifying 40 randomly selected putative unique fragments greater than 5 kb in 'Heinz 1706', of which 36 could not be amplified in *S. pimpinellifolium* LA1589 (90% primers available upon request). We further tested for this deletion in 37 other accessions of *S. pimpinellifolium* and detected intra-specific polymorphism (**Supplementary Fig. 7**).

We also attempted to identify unique regions in *S. pimpinellifolium* by assembling *de novo* all unmapped Illumina reads. This identified ~ 4 Mb of sequence (after filtering out contigs that align well to the 'Heinz 1706' genome or have good Heinz k-mer coverage). Among 855 contigs of ≥1 kb, only 1 Mb remains, and the largest contig is only 3.6 kb. Of these 855 contigs, 624 have BLAST hits to the nt database; however, most sequences are low-identity, low-coverage matches to *Solanum* species, perhaps representing highly diverged repetitive sequences. These putative *S. pimpinellifolium* unique sequences were not investigated further.

3.6 Identification of *S. pimpinellifolium* diversity in domesticated germplasm

To identify SNPs in *S. lycopersicum* varieties, EST sequences derived from *S. lycopersicum* varieties including Micro-Tom, R11-12, R11-13, Rio Grande PtoR, TA492, TA496, TA496 E6203, and West Virginia 106 were retrieved from the MiBASE (http://www.pgb.kazusa.or.jp/mibase/download/all_Microtom_EST.seq.gz)[156], the Kaftom (http://www.pgb.kazusa.or.jp/kaftom/download/HTC13227.zip)[157], and the SGN ftp site (ftp://ftp.solgenomics.net/est_sequences/species/Tomato/tomato_species_2008_10_21.seq)[158]. To clean up the EST sequences, the following base calls were removed; (i) quality value less than 30, (ii) 2-nt repeat more than 20-nt long, (iii) CC- or GG-repeats more than 10-nt long, and (iv) bases in the downstream and upstream of polyA and polyT, respectively. Additionally, the following low-quality EST entries were removed; (i) a ratio of a single nucleotide occupied more than half of the entire length, and (ii) polyA or polyT appeared more than twice in a single EST entry. The EST sequences from the respective varieties were assembled to make consensus sequences using the Phrap program with the parameter settings of penalty -5, minmatch 20, and minscore 100 (http://www.phrap.org/phredphrapconsed.html). The 'Heinz 1706' scaffold that has the highest similarity to each of the consensus sequence was selected using the BLASTN program by using the *S._lycopersicum*_chromosomes 2.40 dataset (http://solgenomics.net/genomes/Solanum_lycopersicum/genome_data.pl). The consensus sequences of which 95% length was covered by the scaffold, and of which identity to the scaffold was more than 98%, were aligned to the scaffold using the est2genome program[159]. According to the alignment, SNPs were identified by the following criteria; (i) candidate nucleotide was covered by more than 3 ESTs, and (ii) the percentage of the number of ESTs with consistent mismatch was more than 75% of total ESTs at that position. According to this alignment, 5,393,697 nt of the

'Heinz 1706' scaffold were covered by *S. lycopersicum* ESTs, and 3,831 SNPs were identified between ESTs from *S. lycopersicum* varieties and the 'Heinz 1706' genome scaffolds (**Supplementary Table 60**).

All detected intra-specific polymorphisms (SNPs) were overlaid with *S. pimpinellifolium* SNPs to investigate the overlap of SNP position and base change. The exact same base changes at the same positions were considered the overlapped SNPs, which were plotted with intra-specific SNPs in 1 Mbp window along the chromosome to see the spatial distribution (**Supplementary Fig. 8)**.

3.7 Cytogenetics

S. pimpinellifolium is thought to be the wild ancestor of *S. lycopersicum*[160]. This is supported by comparing synaptic patterns from hybrids between tomato and other species in the tomato clade[161]. Tomato bivalents are synapsed homologously along their lengths (with the exception of the NOR on the short arm of chromosome *2*). However, tomato hybrids show a variety of synaptic irregularities (**Supplementary Fig. 46**). Mismatched kinetochores are commonly observed in all tomato hybrids. The mismatches indicate non-homologous synapsis that could be due to pericentric inversions. Other irregularities include paracentric inversion loops, a reciprocal translocation, and differences in chromosome length that result in mismatched ends and asynapsed buckles. These synaptic irregularities demonstrate a history of chromosomal rearrangements in different branches of the tomato phylogenetic tree. Significantly, the *S. pimpinellifolium* x *S. lycopersicum* hybrid shows the fewest synaptic irregularities, as would be expected if *S. pimpinellifolium* is cultivated tomato's nearest relative (**Supplementary Fig. 46**). Spreads of synaptonemal complexes (SCs ~ pachytene chromosomes) were prepared from tomato and tomato hybrids as described previously[51,161]. Briefly, cell walls were digested from primary microsporocytes to make protoplasts that were then hypotonically burst on plastic-coated glass slides. SC spreads were air dried before DNase digestion and staining with phosphotungstic acid. The plastic film with SC spreads was transferred to a copper grid for photography in a transmission electron microscope.

4 COMPARATIVE GENOME ANALYSES

4.1 Comparison between the tomato and potato genome sequences

All pairs of corresponding tomato and potato chromosomes were aligned using LASTZ[73] with parameters set to "--gfextend --chain --gapped --identity=80..100 --matchcount=1000". Repetitive overlapping matches were removed using custom Python scripts such that the largest match of each pair of repetitive matches remained. Sequence polymorphism and indel frequencies between the aligned sequence regions were derived from the LASTZ alignments using custom Python scripts. To avoid overestimation of small indels due to sequencing errors, indels of 1 and 2 bp were not considered if the distance between them was less than 30 bp or the identity of 10 bp of flanking sequence on each side was smaller than 95%. In total, 77,896,027 bp was aligned between the two genomes in which there were 6,780,607 mismatches (1 mismatch per 12 bp) and 749,966 indels of up to 300 bp (1 indel per 110 bp). The majority of aligned regions were confined to the gene rich regions in the distal regions of the chromosome arms.

There are 6.2% (2,229) proximally duplicated genes in tomato (4.4% shared with potato, 0.5% uniquely retained, and 1.3% uniquely gained); versus 11.6% (4,047) in potato (10.2% shared with tomato, 3.1% uniquely retained, and 1.0% uniquely gained). Proximal duplication (PD) in tomato is, however, somewhat less frequent than in other well-studied genomes, for example, 8-15% in *Arabidopsis* and 7-14% in rice[162]. The maximum size of a single cluster is 21 for proximal duplication shared by tomato-potato and 25 for proximal duplication unique to tomato; while the respective numbers for potato are 32 and 9 (**Supplementary Fig. 47**), compared with a maximum of 12 in *Arabidopsis* and 9 in rice across several parameter sets.

Uniquely retained tomato proximal duplications are especially enriched for GOslim terms cell organisation and biogenesis, nucleic acid binding, and Plant Ontology structure terms central cell, endosperm, filament, floral bud, flower abscission zone, fruit, hypophysis, lateral root tip, leaf primordium, nucellus, ovule, root meristem, root tip, root vascular system, shoot apical meristem, stigma, stipule, stomatal complex, trichome **(Supplementary Fig. 48).**

Uniquely gained tomato proximal duplications are enriched for GOslim terms ER, cell organisation and biogenesis, structural molecule activity; and PO structure terms anther, callus, central cell, endosperm, filament, floral bud, gynoecium, juvenile leaf, receptacle, stigma, synergid, trichome **(Supplementary Fig. 49)**.

4.2 Partition of the tomato genome into three subgenomes following the *Solanum* triplication

Homology search between annotated tomato (35,802 loci) and grape genes (26,346 loci, 2010 March release, http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/) was done by BLASTP (E-value threshold 1e-10, top 20 hits). An in-house Python script was used to merge tandem repeats within local rank distance of 10 genes, and to retain only top BLAST hits by applying a C-score threshold of 0.8 (C-score(A, B) = score(A, B)/max(best score of A, best score of B))[163]. Filtered homologous gene pairs (23,577) were screened with Quota-Align[164] (https://github.com/tanghaibao/quota-alignment) using quota ratio 1:3 (because a region in the grape genome sharing the *Solanum* triplication is expected to have three best matching orthologous regions in the tomato genome) to identify triplicated regions in tomato following "T".

Tomato triplicated blocks identified were chained into three subgenomes using dynamic programming. The main criteria are that the chained triplicated blocks are: 1) non-overlapping in the tomato genome; 2) have no more than 10% overlap between their orthologous grape regions (results were similar using 0% overlap in grape); 3) maximize coverage of the tomato genome (annotated gene space). A list of tomato genes in the three subgenomes and their orthologous grape genes (if still detectable) are in **Supplementary Table 61.** The distributions of synonymous substitution rates (*Ks*) corresponding to the *Solanum* triplication and tomato-grape divergence are plotted in **Supplementary Fig. 10b**.

4.3 Confirmation of the *Solanum* triplication in potato

The existence of the triplication "T" in potato is already strongly, albeit indirectly supported as follows:

1. "T" is well established in tomato by synteny and Ks analyses. The synteny redundancy levels of "T" paralogs are similar to those from the γ triplication in grape[138], much higher than those retained after two rounds of WGDs in Arabidopsis[165,166] ;

2. the potato genome is in high collinearity with that of tomato;

3. the divergence time of tomato and potato (about 7 million years) is well after the confidence interval of "T" (53-91 million years).

We did also analyse the potato genome directly. A dot plot comparing the potato and grape genomes (**Supplementary Fig. 11**) shows that syntenic regions with redundancy level of 1 in grape (therefore not γ regions) have redundancy levels of 3 in the potato genome. Statistically 27.8% of the coding proportion of grape genome has one orthologous region in potato, 38.1% have two, and 14.5% have three, collectively accounting for 68% of the potato gene space and 80% of the grape gene space, using the published potato annotation[93].

Independent Ks analysis consistently assigns most of the grape paralogs to the "T" age group, clearly distinct from those of γ. The corresponding Ks distribution for "T" paralogs in tomato is provided as **Supplementary Fig. 10b**.

Further, we did a careful examination of Supplementary Figure 6b of the potato genome paper[93] which aligns syntenic regions between grape, *Arabidopsis*, poplar, and potato. We noticed that this figure missed the third syntenic region on potato chromosome *8* that is also produced in "T", besides the two regions aligned on their figure. This example of one-to-three correspondence in the grape and potato genomes can be viewed clearly in **Supplementary Fig. 11**. A detailed corrected multiple alignment is also provided in **Supplementary Table 62**.

4.4 Three-way comparison of the *S. lycopersicum*, *S. pimpinellifolium* and *S. tuberosum* genomes to reveal domestication signatures

We determined 18,809 ortholog groups between *S. lycopersicum* (Sl), *S. pimpinellifolium* (Sp) and *S. tuberosum* (St) by MCscan in order to find genes changed by various factors such as domestication. **Supplementary Fig. 50** shows that the omega values of the most Sl-St pairs are naturally higher than Sl-Sp pairs in the ortholog groups while the number of ortholog groups whose Sl-Sp pair omega value were higher than Sl-St are 221. Additionally, discrete distribution of omega values of Sl-Sp ortholog pairs reflects a few substitutions between the pairs. **Supplementary Table 63** shows functional category and its changing of the 221 groups based on similarity with *Arabidopsis* (TAIR10; e-value ≤ 1e-5) by GO slims (Version 1.2; http://www.geneontology.org/GO.slims.shtml). In the table, only 15 groups show changing of functional category even though some of the changes are obvious such as Solyc04g072790-Sopim04g072790-Sotub04g027070. Moreover, among the 221 groups, only three Sl-Sp pairs were passed the Fisher's exact test because of the too few substitutions. Thus, the results reflect that the two tomato species are too close to find differences supported by statistical evidence. We conclude

that the *S. pimpinellifolium – S. lycopersicum* comparison requires sequencing of multiple genotypes from both species and comparison to an outgroup closer than *S. tuberosum* to reveal statistically significant differences.

4.5 Comparative COS maps of Solanaceae genomes

Over the past two decades the family Solanaceae has been the subject of extensive pairwise comparative mapping studies that allowed depiction of the syntenic relationships among major solanaceous crops (potato, eggplant and pepper) relative to tomato[167]. More recently, the identification of a large set of single-copy conserved ortholog (COSII) markers has greatly enhanced the power of comparative mapping within the family Solanaceae and across plant families—at least with the relatively closely related plant family Rubiaceae (coffee)[168]. COSII genetic maps have been constructed for potato, eggplant, pepper, two diploid *Nicotiana* species as well as coffee, and pairwise comparisons have been reported between each species and tomato (**Supplementary Fig. 14**)[169-172]. A schematic depiction of the main chromosomal rearrangements identified by these COSII comparative mapping efforts conducted in the genomes of potato, eggplant, pepper and *Nicotiana* with respect to the tomato genome is shown in **Fig. 1b**. These synteny studies have highlighted a high level of genetic conservation among solanaceous species; inversions being the most common cause of genome rearrangements—with translocations occurring at a lower frequency. As expected, moving out from the core tomato reference genome, across the phylogenetic tree, we observe a generally increasing extent of genome rearrangements[167]. Therefore, while tomato and potato, close taxonomic relatives in sister clades, differ by only a minimum of 9 inversions (also see **Supplementary Fig. 14**), the synteny between tomato and pepper is disrupted by a minimum of 19 inversions, 6 translocations and numerous putative single gene transpositions[170]. Shifting to the plant family Rubiaceae reveals even more extreme rearrangements (e.g. inversions and translocations); yet, as indicated by the high-resolution COSII comparative study conducted between tomato and coffee, it is still possible to decipher synteny across most of the two genomes[169]. These COSII synteny maps, therefore, provide the infrastructure by which the high quality tomato genome sequence can be used as a reference for the euasterids, and be directly utilized by researchers working on virtually all solanaceous species and the closely allied coffee (Rubiaceae in the order Gentianales). This will facilitate both applied and basic research also for those important crops for which a whole genome sequence is not yet available.

In **Fig. 1b** we provide a schematic representation of the COSII-based comparative maps of potato, eggplant, pepper and *Nicotiana* with respect to the tomato genome. Deductions concerning the syntenic relationships of the potato and tomato genomes were based on 141 COSII markers of known map position in tomato (http://solgenomics.net/) which were mapped in the diploid potato mapping population F1840, for which detailed RFLP linkage maps have been constructed[173,174]. DNA polymorphisms were detected by single strand conformation analysis as described by Bormann *et al.*[175] and mapped using the software package MAPRF (E. Ritter, NEIKER, Apdo 46, E-01080 Vitoria, Spain). 33 additional anchors were obtained *in silico* by sequence comparison of potato and tomato markers. The details of the comparative maps between tomato and potato are shown in **Supplementary Fig. 14**. Putative centromere positions of potato chromosomes, indicated by white dots in **Fig. 1b**, are based on marker clustering on genetic maps of potato as well as on cytogenetic positions of centromeres for the potato map[173,176-178]. The COSII-based mapping study identified a minimum of eight macro-inversions based on inverted order of at least three markers: the five known paracentric inversions[46] between tomato and potato on chromosomes *5, 9, 10, 11* and *12* were confirmed; two additional inverted chromosome segments were identified on the long arm of chromosome *2*, and another one was detected on the long arm of chromosome *12*. There are a number of further, putative small rearrangements based on inverted order of two markers, which could also be the result of mapping artefacts (**Supplementary Fig. 14**). Moreover, a ninth inversion has been identified on the short arm of chromosome *6* by BAC-FISH mapping[59,179,180]. One of the small two-marker rearrangements has been observed on the short arm of chromosome *6*, bordered by markers At3g25120 and GP79 (**Fig. 1b and Supplementary Fig. 14**).

The bordering markers for the 8 macro-inversions are as follows:

Chromosome *2* (short arm): bordering marker loci: At2g04700 and At4g33985

Chromosome *2* (long arm): bordering marker loci: At5g67370 and At5g40530

Chromosome *5* (long arm): bordering marker loci: At1g60440 and T1640

Chromosome *9* (short arm): bordering marker loci: At2g27090 and At2g38025-a

Chromosome *10* (long arm): bordering marker loci: At1g67740 and At3g09740

Chromosome *11* (short arm): bordering marker loci: At5g09880 and CT182

Chromosome *12* (short arm): bordering marker loci: At2g32950 and At5g42740

Chromosome *12* (long arm): bordering marker loci: At4g33360 and At3g17000

For the pairwise comparisons between tomato and eggplant, pepper and *Nicotiana*, a modified tomato genetic map was prepared for which the framework was based predominantly on the COSII markers[167,170-172]. The approximate centromeric locations on the tomato map are based on the comparison with maps of previous studies[170-172,181]. In all these comparative studies the degree of synteny was assessed using synteny marker pairs (SMPs), each defined as a pair of synteny markers that are adjacent to each other (regardless of other non-synteny markers) on both maps, as well as conserved syntenic segments (CSSs), regions within which gene/marker order is well preserved.

The degree of synteny between the eggplant and tomato genomes was based on 289 orthologous markers (110 COSII markers and 179 tomato-derived markers) or "synteny markers"[172]. For the pepper and tomato genomes the synteny was deduced using 299 "synteny markers" including 263 COSII markers[170]; for this comparison, the chromosomes P1-wild and P8-wild were used as they represent the condition shared by most *Capsicum* species except for cultivated *C. annuum* (a derived condition). Approximate centromere positions of eggplant and pepper chromosomes are based on their synteny with tomato[170,172].

For the tobacco and tomato genomes, a direct comparison of the genetic maps is complicated by the tetraploidy of tobacco. Therefore, to overcome this problem, the tobacco and tomato genomes have been compared through the bridge of diploid *Nicotiana* maps[171]. Two genetic maps for the diploid *Nicotiana* species, *N. tomentosiformis* (Tmf population) and *N. acuminata* (Acn population) were constructed, each representing one of the two putative diploid ancestral tobacco genomes – *N. tomentosiformis* and *N. sylvestris*, respectively[171]. For the Tmf population 268 synteny markers (262 COSIIs and 6 tomato-derived markers), which provided a good (93%) coverage of the tomato genetic map, were combined with 221 tobacco SSR markers to generate a relatively dense genetic map, comprising twelve linkage groups. For the Acn population 134 synteny markers (133 COSIIs and 1 tomato-derived marker), covering 55% of the tomato genetic map, were combined with 174 tobacco SSR markers to generate a genetic map, comprised of 308 markers and twelve linkage groups. The complex syntenic relationships with tomato did not allow determination of the centromere positions for the *Nicotiana* chromosomes[167].

5 TOMATO GENE FAMILY ANALYSES

5.1 Detection of gene families from tomato and the Solanaceae using OrthoMCL

To define gene family clusters among different plant species we used the OrthoMCL software[182] in its version 1.4. In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an e-value cut-off of 1e-05. On the resulting similarity matrix, OrthoMCL performs a Markov clustering algorithm to define the cluster structure. For the clustering algorithm we used an inflation value (-I) of 1.5 (OrthoMCL default). The input datasets from various plant species are described in **Supplementary Table 64**. Splice variants were removed from the data set (the longest protein sequence prediction was kept) and data sets were filtered for internal stop codons and incompatible reading frames. All results are available on the European Tomato Genome Database TomDB at MIPS (http://mips.helmholtz-muenchen.de/plant/tomato/index.jsp**)** and SGN (http://solgenomics.net/) and are available as bulk download on request.

Tomato, *Vitis*, *Arabidopsis* and rice:
A total of 119,876 sequences from *Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa* and *Solanum lycopersicum* were clustered into 17,487 gene families. 8,899 clusters contained sequences from all four genomes, 1,700 from eudicots (tomato, *Vitis*, *Arabidopsis*), 607 from fleshy fruit-bearing plants (tomato, *Vitis*) and 1,274 were specific to tomato. Of the 34,771 protein-coding genes predicted for tomato, 25,006 were clustered in a total of 13,821 groups. The 1,274 tomato-specific clusters contained 5,575 genes of which 2,992 have at least one InterPro domain. The tomato sequences which did not fall into any cluster (singletons) made up a total of 9,765 genes of which 4,417 have at least one InterPro domain (**Supplementary Fig. 51**).

Tomato, *Vitis*, *Arabidopsis*, rice, papaya, cucumber, soybean and apple:
A total of 272,572 sequences from *Arabidopsis thaliana, Vitis vinifera, Oryza sativa, Solanum lycopersicum, Carica papaya, Cucumis sativus, Glycine max* and *Malus* × *domestica* were clustered into 25,722 gene families. 6,906 clusters contained sequences from all eight genomes and 1,114 clusters were specific to tomato. Of the 34,771 protein-coding genes predicted for tomato, 25,313

were clustered in a total of 14,361 groups. The 1,114 tomato-specific clusters contained 4,925 genes of which 2,529 have at least one InterPro domain. Singletons made up a total of 9,458 genes of which 4,172 have at least one InterPro domain.

Tomato, potato, *Vitis*, *Arabidopsis* and rice:

A total of 154,880 sequences from *Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa*, *Solanum lycopersicum* and *S. tuberosum* were clustered into 23,208 gene-groups ("families"). Of the 34,771 protein-coding genes predicted for tomato, 25,885 were clustered in a total of 18,783 (≥2 members), gene-groups. From the 18,783 gene-groups 8,615 contained sequences from all four genomes, 1,727 from eudicots (tomato, *Vitis*, *Arabidopsis*, potato), 58 from fleshy fruit-bearing plants (tomato, *Vitis*) and 562 clusters were specific to tomato. 5,165 groups were identified as Solanaceae specific (tomato, potato) with 6,197 and 6,723 tomato and potato genes, respectively. The 562 Tomato-specific clusters contained 1,939 genes of which 650 have at least one InterPro domain. Singletons make up a total of 8,886 genes of which 3,465 have at least one InterPro domain (**Supplementary Fig. 5**). OrthoMCL results were used to identify gene clusters unique to tomato or to the Solanaceae (tomato + potato) and revealed clusters uniquely absent in tomato or in both species (**Supplementary Tables 65-66**). Interestingly, many gene family expansions observed here involve clusters of different transcription factors (bHLH95, PHD zinc finger-containing proteins) and genes involved in hormone homeostasis (GH3 family proteins) (**Supplementary Table 67**).

5.2 Phylogenetic analyses

Evolutionary analyses were performed using gene sequences provided by manual curation from datasets. Some gene families were enriched by performing additional BLASTP searches against gene sets from the genomes of tomato, potato, grape, *Arabidopsis* and rice. All hit sequences with an e-value < $1*e^{-10}$ were included in the analysis. Gene families that were mixed based on pure sequence similarity were corrected to previous definitions of these gene families. PHYML[183] and MEGA[184] were used to reconstruct phylogenetic trees based on CDS and protein sequences. The most probable tree for each gene family was analysed using PAML[150] to infer positive selection at whole-CDS level and at each site in the aligned sequences. MCscan was used to infer gene

collinearity[165], which was taken as credible evidence for gene paralogy and orthology, related to polyploidisations and lineage divergence.

5.3 Ascorbate biosynthesis

Because the first two steps of the ascorbate (vitamin C) L-galactose pathway are catalysed by the hexokinase and phosphoglucose isomerase of glycolysis, the eight enzymatic reactions shown in **Supplementary Fig. 52** are considered as the vitamin C-specific synthesis pathway. Most genes show high expression in all tissues analysed, with the exception of *PMM3* and *GPP3* that show no detectable expression in any of the tissues analysed (**Supplementary Table 68**). Genes encoding *PMM*, *GMP* and *GME* are highly expressed in all tissues tested except flowers. Most genes show high expression levels in fruits, consistent with the hypothesis that the fruit regulates its own vitamin C synthesis and accumulation[185] rather than being dependent on transport from leaves as previously reported[186].

Several genes from the first six steps of the L-ascorbic acid biosynthetic pathway in tomato (*PMI*, *PMM*, *GMP*, *GPP*) are higher in number than the other species analysed (**Supplementary Fig. 52b**). *GDH* and *GalLDH*, coding for the two last enzymes of the pathway, are single copy and limited in expression (**Supplementary Table 68**), although they are highly efficient enzymes[187,188]. Ascorbate is transformed to the short-lived radical monodehydroascorbate (MDHA) after oxidation. MDHA is converted to ascorbic acid by the action of MDHA reductase (MDHAR) or via dehydroascorbate (DHA), after MDHA dissociation, by DHA reductase (DHAR) (**Supplementary Fig. 52a**). These two enzymes are necessary for recycling of ascorbate in its active reduced form and for maintenance of the intracellular redox state[189,190]. The genes coding for these enzymes are duplicated (*DHAR*) or triplicated (*MDHAR*) in tomato and highly expressed in all tissues analysed (**Supplementary Fig. 52b**).

5.4 Cytochrome P450s

Cytochrome P450 sequences including pseudogenes were detected in tomato genome sequence assembly version 2.3 by BLAST searching with known members from the eleven CYP clans. Sequences from the 50 tomato P450 families were used in the searches. Genes were assembled by manual inspection of BLAST output and comparison to the closest known annotated genes from

tomato or other plants. Expert curation by D. Nelson (University of Tennessee) corrected erroneous gene models from the automated pipeline gene assemblies as needed. Neighbour-joining phylogenetic trees were created from CLUSTALW alignments manually edited to remove large insertions or N- and C-terminal extensions. Motifs such as the I-helix, EXXR and heme-binding region were checked for proper alignment. The trees were computed using the Phylip package programs PROTDIST and NEIGHBOR. Trees were drawn with FIGTREE v 1.3.1 and labelled in Illustrator.

Interrogation of the tomato genome sequence reveals 272 cytochrome P450s. These have been compared to *Arabidopsis* P450s in two Neighbour-joining trees (**Supplementary Fig. 53**). Cytochrome P450s in angiosperms are distributed in 10 CYP clans or gene clades, nine of which can be found in tomato (tomato lacks the CYP727 clan). The CYP51, CYP74, CYP97, CYP710, CYP711 and CYP727 clans each contain one CYP family with 1-7 genes. These families carry out very specific biochemical tasks, such as sterol biosynthesis (CYP51, CYP710), carotenoid hydroxylations (CYP97), branch-inhibiting hormone synthesis (CYP711), oxylipin metabolism (allene oxide synthase, hydroperoxide lyase, divinyl ether synthase) (CYP74). CYP727 function remains unknown. The other four clans CYP71, CYP72, CYP85 and CYP86 contain from 4-26 CYP families that perform the bulk of cytochrome P450 reactions in angiosperms. There are 59 defined angiosperm CYP families with a predominant core subset of 52 families[191,192]. Tomato has 272 cytochrome P450 genes and 183 pseudogenes (**Supplementary Tables 69-70**), which reside in 50 CYP families. Only CYP709 and CYP727 are missing from the core family set. CYP709 is missing from the papaya genome and CYP727 is missing in *Arabidopsis* suggesting these are not essential CYP families.

CYP71 is the largest plant CYP family. Tomato has 11 CYP71 subfamilies with 49 genes including 18 CYP71D subfamily genes alone. These, plus the CYP71AT and CYP71AX subfamilies are likely to accomplish specific functions in tomato as they do in other plants[193-196]. Tomato CYP71AX7 is the second most highly expressed P450 genes in breaker fruit, Heinz_B10 and Pimp_B5, though its function remains to be defined **(Supplementary Table 71)**.

Two other large CYP families in tomato are CYP72 (25 genes) and CYP76 (22 genes). CYP72A1 is the enzyme secologanin synthase[197]. CYP72A51 is among the most highly expressed P450s (top 4%) in tomato fruits at ripe stage **(Supplementary Table 71)**. CYP76B10 is geraniol 10-hydroxylase involved in terpenoid indole alkaloid biosynthesis[198]. CYP76M7 is an ent-cassadiene C11 alpha-hydroxylase in diterpenoid biosynthesis in rice[199]. Four other tomato CYP families

include 11-12 genes each (CYP81, CYP94, CYP706 and CYP736). CYP94A24 is by far the most highly expressed P450 gene in ripe fruit in both *S. lycopersicum* and *S. pimpinellifolium* **(Supplementary Table 71)**.

Comparison of tomato-potato P450s revealed gene losses in tomato (**Supplementary Table 72**). Twenty eight potato genes have pseudogene orthologs in tomato. Twelve are in the large CYP71 and CYP72 families. It is important to note that a small percentage of ESTs are missing from both genomes and thus some would be expected to be missing though present in the actual genome. Nevertheless, there are tomato pseudogenes that identify the complete loss of six CYP subfamilies: CYP71BM, CYP80N, CYP82V, CYP86G, CYP94K and CYP716Q. CYP80N1 from potato belongs to a largely basal eudicot associated CYP family associated with alkaloid biosynthesis[191]. Loss of this gene in the tomato may be related to the production of an attractive fruit dispersed by animals and then domesticated by humans. However, tomato is known to make the alkaloid α-tomatine in foliage and immature fruits and, in some cherry tomato strains, in ripe fruits, without toxic effects[200]. CYP82E4 is a nicotine demethylase in tobacco[201] so the reduction of CYP82E tomato subfamily genes (one in tomato versus three in potato) may also affect alkaloid content. The unique CYP82E11 gene in tomato has near zero expression in all tissues in **Supplementary Table 71**. The remaining missing subfamilies have no ties to known pathways.

The tomato and potato genomes have 272 and 399 annotated cytochrome P450s, respectively. Comparison identified 285 ortholog pairs with some being pseudogenes. Many of the excess potato P450 genes are caused by tandem duplication in gene clusters. Twenty-four of the potato orthologs have a highly similar tandem duplicated neighbour not present in tomato. Another cause of excess potato P450s is gene loss in tomato, which in some cases can be validated by comparison with eggplant and tobacco. The potato genes CYP78A80, CYP80N1, CYP82E12, CYP82E13, CYP94C31, CYP704A76 and CYP716Q3 are absent from tomato and represent gene losses. There are only 21 tomato genes that do not have potato orthologs. None of these is the sole member of its subfamily.

5.5 Carotenoid biosynthesis

Carotenoid biosynthesis takes place in the plastids (chloroplasts and chromoplasts), where glyceraldehyde-3-P and pyruvate are converted to GGPP via the MEP pathway. GGPP is converted into the main fruit carotenes, lycopene and β-carotene, and to downstream xanthophylls, through

the action of several dedicated biosynthetic genes[202]. A series of apocarotenoids, including ABA are generated from carotenoids through the action of cleavage dioxygenases (NCEDs, CCDs) (**Supplementary Fig. 54**). Whole-genome analysis showed an expansion, in *Solanum* with respect to *Arabidopsis*, of genes encoding several biosynthetic steps, such as *PSY*, encoding the first dedicated step in lycopene biosynthesis, and *LCY-b/CYC-b*, mediating the conversion of lycopene into β-carotene (**Supplementary Fig. 54**). The different paralogs show different tissue-specific expression (**Supplementary Table 73**). For example, the tomato *PSY1* and, to a lesser extent, *PSY2* genes are expressed in fruits; however, only *Psy1* controls fruit pigmentation (**Fig. 3**). The *Psy1-Psy2* duplication appears to coincide with the *Solanum* triplication (**Supplementary Fig. 16**) as are several duplications in the *IPP*, *GGPS*, and *CHY* families. The eudicot-common triplication may have contributed to the expansion to the *PSY*, *GGPS* and *CYP97* families.

5.6 Transcriptional and hormonal regulation of fruit ripening

Six ethylene receptor genes have been previously reported in tomato (5 in *Arabidopsis*), though the genome sequence reveals a seventh (Solyc05g055070). Three genes corresponding to the *NEVER-RIPE*(*LeERT3*), *LeETR4* and *LeETR6* genes have fruit associated effects in mutant or transgenic repression lines[203-205] indicating an expansion of the receptor family to mediate ethylene signalling in the maturing fruit organ. Tomato harbours three *CTR1* genes (one in *Arabidopsis*)[206,207] Suggesting that multiple aspects of ethylene signalling have been expanded in tomato. Rice, which does not produce a fleshy fruit, contains only one of each gene while the grape genome contains four receptors and two *CTR1*s, consistent with the absence of a ripening ethylene requirement in its non-climacteric fruit. There are 11 tomato ACC synthase (*ACS*) ethylene biosynthesis genes[208]. *ACS2* is the most similar to *ACS4* with three additional closely related genes in the same OrthoMCL cluster (ORTHOMCL532), together with only three *Arabidopsis*, two rice, and one grape gene. In summary, it appears that both critical ethylene synthesis and perception gene families have expanded in tomato and are manifest in the unique tissues of the ripening fruit.

Many ripening processes, including ethylene synthesis, are regulated in tomato by the *RIN* and *TAGL1* MADS-box[209,210], *CNR-SQUAMOSA* promoter binding protein[211], *LeHB-1*[212] and *SlAP2a*[213,214] transcription factors. Analysis of the tomato genome indicated a total of 2,459 loci that could be annotated as transcription or transcriptional regulators (**Supplementary Table 74**) of which 223 show expression patterns that change substantially between mature unripe and ripening

fruit (**Supplementary Table 75**) and thus potentially influencing ripening processes. OrthoMCL analysis of the *CNR* ripening regulator (Solyc02g077920) indicates additional loci in tomato (Solyc10g009080) and grape, as compared to rice and *Arabidopsis*. Similarly, phylogenetic analysis of the *RIN* gene[215] shows it to be member of the *SEPALLATA* clade of MADS-box genes, of which there are four in *Arabidopsis* with two additional genes in tomato including *RIN* (Solyc05g012020) and a closely-related presumed pseudogene (Solyc04g005320) for which there is minimal transcriptional support. Comparative OrthoMCL analysis of *Arabidopsis*, rice and grape indicates no close *RIN* homologs suggesting an expansion of the family to facilitate ripening apparently only in climacteric fruit. This is consistent with recent data showing that *RIN* binds the promoters of a number of ripening-related and ethylene-responsive genes[216,217]. Distinct from the *SEPALLATA* clade is the *AGAMOUS* (*AG*) clade of MADS-box genes, represented by four genes in both tomato and *Arabidopsis*[209]. In contrast to the cases of CNR and the *RIN/ SEPALLATA* MADS-box clade, where gene family expansion has mediated a unique evolutionary adaptation (ripening), *AG* clade genes have functionally differentiated between *Arabidopsis* and tomato to promote essentially equivalent functions in the context of distinct fruit types.

5.7 Cell wall remodelling in ripening tomato fruits

Plant cell walls are highly complex matrices comprising a variety of polysaccharides and several types of structural glycoproteins. In tomato, as is typical with the primary walls of dicotyledons, the principal features include a semi-rigid composite of cellulose microfibrils and the hemicellulose xyloglucan, which is embedded in a gel-like matrix of differenct classes of pectins. The precise roles of most of the polysaccharide components are still unknown and the functions of the most of the glycoprotein components are also poorly understood[218].

The tomato genome sequence contains more than 700 gene models annotated as having cell wall-related functions. Of these, around 90 have been characterized to varying degrees, and named sequences appear in GenBank (**Supplementary Table 76**), A few general conclusions can be drawn with respect to those genes that are predominantly expressed in the fruit (**Supplementary Table 12**).

The transition from the mature green stage to the onset of ripening (breaker stage) is marked by a substantial decrease in the expression of at least four cellulose synthase genes and also those encoding endo-1,4-β-glucanases (designated glycosyl hydrolase family 9, GH9;

http://www.cazy.org). Fifteen genes annotated as encoding xyloglucan endotransglucosylase/hydrolases (XTHs)[219] of GH16 show predominantly fruit development and / or ripening-related expression, of which four do not have closely related forms in *Arabidopsis*, grape or rice (**Supplementary Table 12**). There are also indications of gene duplications within the tomato *XTH* family (**Fig. 4a**) and comparative analysis of gene number and content with potato reveals changes especially in the *XTH* gene families; for example, *SlXTH10* shows evidence of purifying selection (**Supplementary Table 12**). These data suggest that xyloglucan modifying enzymes may play a more important role in fruit development and softening than previously suspected.

Several genes were identified that were mainly, or exclusively, expressed in fruit tissue (**Supplementary Table 12**) such as a polygalacturonase (PG; Solyc10g080210). As previously reported, several pectin esterases show fruit-related expression and our analysis confirms that they are all down-regulated between the mature green and red ripe stages. In contrast, a PG-2a (Solyc10g080210) and a member of the pectate lyase gene family (Solyc03g111690) showed a dramatic increase in expression from breaker onward (**Supplementary Table 12**). An unexpected observation was the presence of five fruit-related fasciclin-like arabinogalactan proteins (FLAs) that showed a substantial decrease in transcript abundance after the mature green stage. These FLAs could have an important role in the softening process as they have been proposed to function in cell wall cross-linking or as pectin plasticizers[220].

The cell wall, transcription factor and ethylene annotated genes were filtered based on their expression patterns. Those selected from their respective functional category lists (**Supplementary Tables 74, 76**) were chosen on the basis of a) expression levels of at least 10 RPKM (transcription and ethylene genes) or 100 RPKM (cell wall genes) in developing and/or ripening fruits and b) differential expression between mature green and breaker or mature green and red ripe stages (> two-fold change and adjusted p value < 0.01). These differentially expressed cell wall and transcription/ethylene genes are listed in **Supplementary Tables 12** and **75**, respectively.

5.8 Resistance-like proteins

We selected from the PRG database[221] previously described R-genes. Seventy-eight genes including 33 from members of the Solanaceae were chosen to build HMM motifs reflective of R-genes. Specifically, 43 CNL, 13 TNL, 13 RLP, and 8 RLK sequences were used to create 140 HMM

profiles specific for the typical resistance gene classes and their corresponding conserved domains. The initial output from scanning the tomato genome revealed 1300 putative R-genes. We then inspected them for the presence of specific R-protein domains using InterProScan and the InterPro database[113] to select for 693 higher probability R-gene candidates. Predicted R-proteins with high similarity for sequence and structure as compared to know R-proteins were retrieved from the tomato, potato *Arabidopsis*, grape and rice genomes. Each sequence was assigned to appropriate known R-gene classes with more unique sequences annotated based on their conserved domains. Evolutionary history within selected OrthoMCL groups was inferred by using the Maximum Likelihood method based on the Whelan and Goldman + Freq. model[222].

In order to scan the tomato genome for putative R-genes a prediction pipeline was developed, based on HMM profiling and InterProScan domain analysis[113]. Using this approach we predicted 266 putative resistance proteins containing a NBS or a TIR domain. Most fall in previously described R-protein classes (**Supplementary Table 77**). Among the eudicot genomes, tomato and potato have the highest fraction of CNL proteins that are generally among the most highly represented in monocot genomes[223]. Furthermore, the tomato, grape and potato genomes with respect to *Arabidopsis*, showed a contraction in the TNL (Toll/interleukin-Nucleotide binding site-Leucine rich repeat) subclass of R-genes. Rice harbours three times as many CNL genes (402) as tomato (118), while *Arabidopsis* and grape have much smaller CNL gene families. To obtain an overview of genes putatively involved in resistance process proteins belonging to RLP and RLK classes were also recorded. Several differences in the distribution of single transmembrane receptor class among eudicot species were observed.

Phylogenetic analysis of selected OrthoMCL groups demonstrates that allelic diversities vary remarkably among R-gene families **(Supplementary Figures 55** and **56**). Together these observations are consistent with the notion that the R-gene component of plant genomes is large, diverse and under pressure for high rates of change resulting in substantively different gene repertoires even in closely related species such as tomato and potato.

5.9 Representation of evolutionary trees

Initial tree(s) for the heuristic search were obtained automatically as follows: if the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used, otherwise the BIONJ method with MCL distance matrix was used. A

discrete Gamma distribution was used to model evolutionary rate differences among sites (five categories). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Bootstrap analysis was conducted to test robustness of tree topology. Evolutionary analyses were conducted in MEGA5[224] . The trees for the *CNR*, *ETR*, *ACS*, *MADS-RIN* and *PHY* gene families are presented in **Supplementary Fig. 17**.

6 REFERENCES

29 Ozminkowski, R. Pedigree of variety Heinz 1706. Rep Tomato Genet Coop 54, 26 (2004).

30 Zhang, H.-B., Zhao, X., Ding, X., Paterson, A. H. & Wing, R. A. Preparation of megabase-size DNA from plant nuclei. The Plant Journal 7, 175-184 (1995).

31 Budiman, M. A., Mao, L., Wood, T. C. & Wing, R. A. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res 10, 129-136 (2000).

32 Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380 (2005).

33 Wiley, G. et al. Methods for generating shotgun and mixed shotgun/paired-end libraries for the 454 DNA sequencer. Curr Protoc Hum Genet Chapter 18 (2009).

34 Roe, B. A. Shotgun library construction for DNA sequencing. Methods Mol Biol 255, 171-187 (2004).

35 Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8, 4321-4325 (1980).

36 van Oeveren, J. et al. Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res 21, 618-625 (2011).

37 Vos, P. et al. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23, 4407-4414 (1995).

38 Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci 13, 523-535 (1997).

39 Applied-Biosystems. (ed Applied Biosystems) 50 (2010).

40 Kim, H. et al. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza. Genome Biol 9, R45 (2008).

41 Luo, M. C. et al. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82, 378-389 (2003).

42 Nelson, W. & Soderlund, C. Integrating sequence with FPC fingerprint maps. Nucleic Acids Res 37, e36 (2009).

43 Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker (2009).

44 Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462-467 (2005).

45 Shirasawa, K. et al. An interspecific linkage map of SSR and intronic polymorphism markers in tomato. Theor Appl Genet 121, 731-739 (2010).

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

47 Soderlund, C., Nelson, W., Shoemaker, A. & Paterson, A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res 16, 1159-1168 (2006).

48 Chang, S. B., Anderson, L. K., Sherman, J. D., Royer, S. M. & Stack, S. M. Predicting and testing physical locations of genetically mapped loci on tomato pachytene chromosome 1. Genetics 176, 2131-2138 (2007).

49 Peterson, D. G., Lapitan, N. L. & Stack, S. M. Localization of single- and low-copy sequences on tomato synaptonemal complex spreads using fluorescence in situ hybridization (FISH). Genetics 152, 427-439 (1999).

50 Stack, S. M. et al. Role of fluorescence in situ hybridization in sequencing the tomato genome. Cytogenet Genome Res 124, 339-350 (2009).

51 Stack, S. M. & Anderson, L. K. Electron microscopic immunogold localization of recombination-related proteins in spreads of synaptonemal complexes from tomato microsporocytes. Methods Mol Biol 558, 147-169 (2009).

52 Peterson, D. G., Pearson, W. R. & Stack, S. M. Characterization of the tomato (Lycopersicon esculentum) genome using in vitro and in situ DNA reassociation. Genome 41, 346-356 (1998).

53 Zwick, M. S. et al. A rapid procedure for the isolation of C0t-1 DNA from plants. Genome 40, 138-142 (1997).

54 Reeves, A. MicroMeasure: a new computer program for the collection and analysis of cytogenetic data. Genome 44, 439-443 (2001).

55 Sherman, J. D. & Stack, S. M. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. V. Tomato (Lycopersicon esculentum) karyotype and idiogram. Genome 35, 354-359 (1992).

56 Sherman, J. D. & Stack, S. M. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (Lycopersicon esculentum). Genetics 141, 683-708 (1995).

57 Szinay, D. et al. High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. Plant J 56, 627-637 (2008).

58 Chang, S.-B. et al. FISH mapping and molecular organization of the major repetitive sequences of tomato. Chromosome Research 16, 919-933 (2008).

59 Tang, X. et al. Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. Genetics 180, 1319-1328 (2008).

60 Kocsis, E., Trus, B. L., Steer, C. J., Bisher, M. E. & Steven, A. C. Image averaging of flexible fibrous macromolecules: the clathrin triskelion has an elastic proximal segment. J Struct Biol 107, 6-14 (1991).

61 Peterson, D. G., Stack, S. M., Price, H. J. & Johnston, J. S. DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes. Genome 39, 77-82 (1996).

62 Mueller, L. A. et al. A snapshot of the emerging tomato genome sequence. The Plant Genome 2, 78-92 (2009).

63 Chou, H. H. & Holmes, M. H. DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093-1104 (2001).

64 Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98, 9748-9753 (2001).

65 Campagna, D. et al. PASS: a program to align short sequences. Bioinformatics 25, 967-968 (2009).

66 Datema, E. et al. Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. BMC Plant Biol 8, 34 (2008).

67 Shendure, J. & Ji, H. Next-generation DNA sequencing. Nat Biotechnol 26, 1135-1145 (2008).

68 Chaisson, M., Pevzner, P. & Tang, H. Fragment assembly with short reads. Bioinformatics 20, 2067-2074 (2004).

69 Wong, T. K., Lam, T. W., Chan, P. Y. & Yiu, S. M. Correcting short reads with high error rates for improved sequencing result. Int J Bioinform Res Appl 5, 224-237 (2009).

70 Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6, 291-295 (2009).

71 Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967 (2009).

72 Peters, S. A. et al. Solanum lycopersicum cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. Plant J 58, 857-869 (2009).

73 Schwartz, S. et al. Human-mouse alignments with BLASTZ. Genome Res 13, 103-107 (2003).

74 Chain, P. S. et al. Genomics. Genome project standards in a new era of sequencing. Science 326, 236-237 (2009).

75 Gianniny, C., Stoeva, P., Cheely, A. & Dimaculangan, D. RAPD analysis of mtDNA from tomato flowers free of nuclear DNA artifacts. Biotechniques 36, 772-774, 776 (2004).

76 Scotti, N., Cardi, T. & Marechaldrouard, L. Mitochondrial DNA and RNA isolation from small amounts of potato tissue. Plant Molecular Biology Reporter 19, 67-67 (2001).

77 Sugiyama, Y. et al. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Molecular Genetics and Genomics 272, 603-615 (2004).

78 Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147 (2010).

79 Shikanai, T., Kaneko, H., Nakata, S., Harada, K. & Watanabe, K. Mitochondrial genome structure of a cytoplasmic hybrid between tomato and wild potato. Plant Cell Reports 17, 832-836 (1998).

80 Tian, X., Zheng, J., Hu, S. & Yu, J. The rice mitochondrial genomes and their variations. Plant Physiol 140, 401-410 (2006).

81 Lopez, J. V., Yuhki, N., Masuda, R., Modi, W. & O'Brien, S. J. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol 39, 174-190 (1994).

82 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 (2006).

83 Stegemann, S., Hartmann, S., Ruf, S. & Bock, R. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A 100, 8828-8833 (2003).

84 Sheppard, A. E. & Timmis, J. N. Instability of plastid DNA in the nuclear genome. PLoS Genet 5, e1000323 (2009).

85 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006).

86 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680 (1994).

87 Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. Curr Protoc Bioinformatics Chapter 4, Unit 4 (2007).

88 Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62 (2006).

89 Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. J Comput Biol 13, 379-393 (2006).

90 Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1, S140-148 (2001).

91 Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics 3, 87-97 (2008).

92 Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383 (2007).

93 Xu, X. et al. Genome sequence and analysis of the tuber crop potato. Nature 475, 189-195 (2011).

94 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335-1337 (2009).

95 Gardner, P. P. et al. Rfam: updates to the RNA families database. Nucleic Acids Res 37, D136-140 (2009).

96 Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6, R44 (2005).

97 Gremme, G., Brendel, V., Sparks, M. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology 47, 965-978 (2005).

98 Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-5666 (2003).

99 Wei, C. & Brent, M. R. Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 7, 327 (2006).

100 Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res 13, 46-54 (2003).

101 Tenney, A. E. et al. Gene prediction and verification in a compact genome with numerous small introns. Genome Res 14, 2330-2335 (2004).

102 Brown, R. H., Gross, S. S. & Brent, M. R. Begin at the beginning: predicting genes with 5' UTRs. Genome Res 15, 742-747 (2005).

103 Zhong, S. et al. High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc 2011, 940-949 (2011).

104 López-Gómez, R. & Gomez-Lim, M. A. A method for extraction of intact RNA from fruits rich in polysaccharides using ripe mango mesocarp. HortScience 27, 2 (1992).

105 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).

106 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).

107 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628 (2008).

108 Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106 (2010).

109 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57, 12 (1995).

110 Massa, A. N. et al. The Transcriptome of the Reference Potato Genome Solanum tuberosum Group Phureja Clone DM1-3 516R44. PLoS One 6, e26801 (2011).

111 Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21, 1332-1338 (2005).

112 Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848 (2001).

113 Hunter, S. et al. InterPro: the integrative protein signature database. Nucleic Acids Res 37, D211-215 (2009).

114 Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402 (1997).

115 Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31, 365-370 (2003).

116 Huala, E. et al. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29, 102-105 (2001).

117 Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-29 (2000).

118 Jocker, A., Hoffmann, F., Groscurth, A. & Schoof, H. Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). Bioinformatics 24, 2393-2394 (2008).

119 Engelhardt, B. E., Jordan, M. I., Muratore, K. E. & Brenner, S. E. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1, e45 (2005).

120 Camon, E. et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32, D262-266 (2004).

121 Camon, E. B. et al. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6 Suppl 1, S17 (2005).

122 Lu, C. et al. Elucidation of the small RNA component of the transcriptome. Science 309, 1567-1569 (2005).

123 Mohorianu, I. et al. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. Plant J 67, 232-246 (2011).

124 Fahlgren, N. et al. Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA 15, 992-1002 (2009).

125 Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307-315 (2004).

126 Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5, R80 (2004).

127 Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249-264 (2003).

128 Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24, 2395-2396 (2008).

129 Cuperus, J. T., Fahlgren, N. & Carrington, J. C. Evolution and functional diversification of MIRNA genes. Plant Cell 23, 431-442 (2011).

130 Addo-Quaye, C., Miller, W. & Axtell, M. J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics 25, 130-131 (2009).

131 Meyers, B. C. et al. Criteria for annotation of plant MicroRNAs. Plant Cell 20, 3186-3190 (2008).

132 McCarthy, E. M. & McDonald, J. F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19, 362-367 (2003).

133 SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. Nat Genet 20, 43-45 (1998).

134 Wicker, T. et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973-982 (2007).

135 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-580 (1999).

136 Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9, 517 (2008).

137 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796-815 (2000).

138 Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463-467 (2007).

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

140 Carels, N., Hatey, P., Jabbari, K. & Bernardi, G. Compositional properties of homologous coding sequences from plants. J Mol Evol 46, 45-53 (1998).

141 Matassi, G., Montero, L. M., Salinas, J. & Bernardi, G. The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res 17, 5273-5290 (1989).

142 Salinas, J., Matassi, G., Montero, L. M. & Bernardi, G. Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res 16, 4269-4285 (1988).

143 Gendrel, A. V., Lippman, Z., Martienssen, R. & Colot, V. Profiling histone modification patterns in plants using genomic tiling microarrays. Nat Methods 2, 213-218 (2005).

144 Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. Genome Res 19, 1117-1123 (2009).

145 Kent, W. J. BLAT-the BLAST-like alignment tool. Genome Res 12, 656-664 (2002).

146 Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11, 473-483 (2011).

147 Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).

148 Flicek, P. et al. Ensembl's 10th year. Nucleic Acids Res 38, D557-562 (2011).

149 Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-426 (1986).

150 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591 (2007).

151 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595 (2010).

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

153 Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. Bioinformatics 23, 257-258 (2007).

154 Bluthgen, N. et al. Biological profiling of gene groups utilizing Gene Ontology. Genome Inform 16, 106-115 (2005).

155 Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676 (2005).

156 Yano, K. et al. MiBASE：A database of a miniature tomato cultivar Micro-Tom. Plant Biotechnology 23, 195-198 (2006).

157 Aoki, K. et al. Large-scale analysis of full-length cDNAs from the tomato (Solanum lycopersicum) cultivar Micro-Tom, a reference system for the Solanaceae genomics. BMC Genomics 11, 210 (2010).

158 Mueller, L. A. et al. The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol 138, 1310-1317 (2005).

159 Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477-478 (1997).

160 Ranc, N., Munos, S., Santoni, S. & Causse, M. A clarified position for Solanum lycopersicum var. cerasiforme in the evolutionary history of tomatoes (solanaceae). BMC Plant Biol 8, 130 (2008).

161 Anderson, L. K., Covey, P. A., Larsen, L. R., Bedinger, P. & Stack, S. M. Structural differences in chromosomes distinguish species in the tomato clade. Cytogenet Genome Res 129, 24-34 (2010).

162 Rizzon, C., Ponger, L. & Gaut, B. S. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. PLoS Comput Biol 2, e115 (2006).

163 Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86-94 (2007).

164 Tang, H. et al. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12, 102 (2011).

165 Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res 18, 1944-1954 (2008).

166 Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422, 433-438 (2003).

167 Wu, F. & Tanksley, S. D. Chromosomal evolution in the plant family Solanaceae. BMC Genomics 11, 182 (2010).

168 Wu, F., Mueller, L. A., Crouzillat, D., Petiard, V. & Tanksley, S. D. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174, 1407-1420 (2006).

169 Lefebvre-Pautigny, F. et al. High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. Tree Genetics & Genomes 6, 565-577 (2010).

170 Wu, F. et al. A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. Theor Appl Genet 118, 1279-1293 (2009).

171 Wu, F. et al. COSII genetic maps of two diploid Nicotiana species provide a detailed picture of synteny with tomato and insights into chromosome evolution in tetraploid N. tabacum. Theor Appl Genet 120, 809-827 (2010).

172 Wu, F., Eannetta, N. T., Xu, Y. & Tanksley, S. D. A detailed synteny map of the eggplant genome based on conserved ortholog set II (COSII) markers. Theor Appl Genet 118, 927-935 (2009).

173 Gebhardt, C. et al. Comparative mapping between potato (Solanum tuberosum) and Arabidopsis thaliana reveals structurally conserved domains and ancient duplications in the potato genome. Plant J 34, 529-541 (2003).

174 Rickert, A. M. et al. First-generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. Plant Biotechnol J 1, 399-410 (2003).

175 Bormann, C. A. et al. Tagging quantitative trait loci for maturity-corrected late blight resistance in tetraploid potato with PCR-based candidate gene markers. Mol Plant Microbe Interact 17, 1126-1138 (2004).

176 Achenbach, U. C., Tang, X., Ballvora, A., de Jong, H. & Gebhardt, C. Comparison of the chromosome maps around a resistance hot spot on chromosome 5 of potato and tomato using BAC-FISH painting. Genome 53, 103-110 (2010).

177 Dong, F. et al. Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Theor Appl Genet 101, 1001-1007 (2000).

178 van Os, H. et al. Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173, 1075-1087 (2006).

179 Iovene, M., Wielgus, S. M., Simon, P. W., Buell, C. R. & Jiang, J. Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato. Genetics 180, 1307-1317 (2008).

180 Lou, Q., Iovene, M., Spooner, D. M., Buell, C. R. & Jiang, J. Evolution of chromosome 6 of Solanum species revealed by comparative fluorescence in situ hybridization mapping. Chromosoma 119, 435-442 (2010).

181 Frary, A. et al. Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111, 291-312 (2005).

182 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13, 2178-2189 (2003).

183 Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33, W557-559 (2005).

184 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599 (2007).

185 Gautier, H., Massot, C., Stevens, R., Serino, S. & Genard, M. Regulation of tomato fruit ascorbate content is more highly dependent on fruit irradiance than leaf irradiance. Ann Bot 103, 495-504 (2009).

186 Franceschi, V. R. & Tarlyn, N. M. L-Ascorbic acid is accumulated in source leaf phloem and transported to sink tissues in plants. Plant Physiol 130, 649-656 (2002).

187 Alhagdow, M. et al. Silencing of the mitochondrial ascorbate synthesizing enzyme L-galactono-1,4-lactone dehydrogenase affects plant and fruit development in tomato. Plant Physiol 145, 1408-1422 (2007).

188 Gatzek, S., Wheeler, G. L. & Smirnoff, N. Antisense suppression of l-galactose dehydrogenase in Arabidopsis thaliana provides evidence for its role in ascorbate synthesis and reveals light modulated l-galactose synthesis. Plant J 30, 541-553 (2002).

189 Noctor, G. & Foyer, C. H. ASCORBATE AND GLUTATHIONE: Keeping Active Oxygen Under Control. Annu Rev Plant Physiol Plant Mol Biol 49, 249-279 (1998).

190 Smirnoff, N. L-ascorbic acid biosynthesis. Vitam Horm 61, 241-266 (2001).

191 Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. Plant J 66, 194-211 (2011).

192 Nelson, D. R., Ming, R., Alam, M. & Schuler, M. A. Comparison of cytochrome P450 genes from six plant genomes. Tropical Plant Biology 1, 216-235 (2008).

193 Szczyglowski, K., Hamburger, D., Kapranov, P. & de Bruijn, F. J. Construction of a Lotus japonicus late nodulin expressed sequence tag library and identification of novel nodule-specific genes. Plant Physiol 114, 1335-1346 (1997).

194 Schroeder, G. et al. Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: Tabersonine 16-hydroxylase. FEBS letters 458, 97-102 (1999).

195 Turner, G. W. & Croteau, R. Organization of monoterpene biosynthesis in Mentha. Immunocytochemical localizations of geranyl diphosphate synthase, limonene-6-

hydroxylase, isopiperitenol dehydrogenase, and pulegone reductase. Plant Physiol 136, 4215-4227 (2004).

196 Ralston, L. et al. Cloning, heterologous expression, and functional characterization of 5-epi-aristolochene-1,3-dihydroxylase from tobacco (Nicotiana tabacum). Arch Biochem Biophys 393, 222-235 (2001).

197 Irmler, S. et al. Indole alkaloid biosynthesis in Catharanthus roseus: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. Plant J 24, 797-804 (2000).

198 Collu, G. et al. Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. FEBS Lett 508, 215-220 (2001).

199 Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. Plant Cell 21, 3315-3325 (2009).

200 Rick, C. M., Uhlig, J. W. & Jones, A. D. High alpha-tomatine content in ripe fruit of Andean Lycopersicon esculentum var. cerasiforme: developmental and genetic aspects. Proc Natl Acad Sci U S A 91, 12877-12881 (1994).

201 Lewis, R. S., Bowen, S. W., Keogh, M. R. & Dewey, R. E. Three nicotine demethylase genes mediate nornicotine biosynthesis in Nicotiana tabacum L.: functional characterization of the CYP82E10 gene. Phytochemistry 71, 1988-1998 (2010).

202 Giuliano, G., Tavazza, R., Diretto, G., Beyer, P. & Taylor, M. A. Metabolic engineering of carotenoid biosynthesis in plants. Trends Biotechnol 26, 139-145 (2008).

203 Kevany, B. M., Tieman, D. M., Taylor, M. G., Cin, V. D. & Klee, H. J. Ethylene receptor degradation controls the timing of ripening in tomato fruit. Plant J 51, 458-467 (2007).

204 Tieman, D. M., Taylor, M. G., Ciardi, J. A. & Klee, H. J. The tomato ethylene receptors NR and LeETR4 are negative regulators of ethylene response and exhibit functional compensation within a multigene family. Proc Natl Acad Sci U S A 97, 5663-5668 (2000).

205 Wilkinson, J. Q., Lanahan, M. B., Yen, H. C., Giovannoni, J. J. & Klee, H. J. An ethylene-inducible component of signal transduction encoded by never-ripe. Science 270, 1807-1809 (1995).

206 Adams-Phillips, L. et al. Evidence that CTR1-mediated ethylene signal transduction in tomato is encoded by a multigene family whose members display distinct regulatory features. Plant Mol Biol 54, 387-404 (2004).

207 Leclercq, J. et al. LeCTR1, a tomato CTR1-like gene, demonstrates ethylene signaling ability in Arabidopsis and novel expression patterns in tomato. Plant Physiol 130, 1132-1142 (2002).

208 Barry, C. S., Llop-Tous, M. I. & Grierson, D. The regulation of 1-aminocyclopropane-1-carboxylic acid synthase gene expression during the transition from system-1 to system-2 ethylene synthesis in tomato. Plant Physiol 123, 979-986 (2000).

209 Vrebalov, J. et al. Fleshy fruit expansion and ripening are regulated by the Tomato SHATTERPROOF gene TAGL1. Plant Cell 21, 3041-3062 (2009).

210 Vrebalov, J. et al. A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. Science 296, 343-346 (2002).

211 Manning, K. et al. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet 38, 948-952 (2006).

212 Lin, Z. et al. A tomato HD-Zip homeobox protein, LeHB-1, plays an important role in floral organogenesis and ripening. Plant J 55, 301-310 (2008).

213 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 (2010).

214 Karlova, R. et al. Transcriptome and metabolite profiling show that APETALA2a is a major regulator of tomato fruit ripening. Plant Cell 23, 923-941 (2011).

215 Hileman, L. C. et al. Molecular and phylogenetic analyses of the MADS-box gene family in tomato. Mol Biol Evol 23, 2245-2258 (2006).

216 Martel, C., Vrebalov, J., Tafelmeyer, P. & Giovannoni, J. J. The tomato MADS-box transcription factor RIPENING INHIBITOR interacts with promoters involved in numerous ripening processes in a COLORLESS NONRIPENING-dependent manner. Plant Physiol 157, 1568-1579 (2011).

217 Fujisawa, M., Nakano, T. & Ito, Y. Identification of potential target genes for the tomato fruit-ripening regulator RIN by chromatin immunoprecipitation. BMC Plant Biol 11, 26 (2011).

218 McCann, M. & Rose, J. K. C. Blueprints for building plant cell walls. Plant Physiol 153, 365 (2010).

219 Saladié, M., Rose, J. K. C., Cosgrove, D. J. & Catala, C. Characterization of a new xyloglucan endotransglucosylase/hydrolase (XTH) from ripening tomato fruit and implications for the diverse modes of enzymic action. Plant J 47, 282-295 (2006).

220 Ellis, M., Egelund, J., Schultz, C. J. & Bacic, A. Arabinogalactan-proteins: key regulators at the cell surface? Plant Physiol 153, 403-419 (2010).

221 Sanseverino, W. et al. PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res 38, D814-821 (2010).

222 Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18, 691-699 (2001).

223 Goff, S. A. et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296, 92-100 (2002).

224 Tamura, K. et al. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony methods. Mol Biol Evol (2011).

225 Navarro, C. et al. Control of flowering and storage organ formation in potato by FLOWERING LOCUS T. Nature 478, 119-122 (2011).

226 Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. & Araki, T. A pair of related genes with antagonistic roles in mediating flowering signals. Science 286, 1960-1962 (1999).

227 Yamaguchi, A., Kobayashi, Y., Goto, K., Abe, M. & Araki, T. TWIN SISTER OF FT (TSF) acts as a floral pathway integrator redundantly with FT. Plant Cell Physiol 46, 1175-1189 (2005).

228 Xi, W., Liu, C., Hou, X. & Yu, H. MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in Arabidopsis. Plant Cell 22, 1733-1748 (2010).

229 Yoo, S. J. et al. BROTHER OF FT AND TFL1 (BFT) has TFL1-like activity and functions redundantly with TFL1 in inflorescence meristem development in Arabidopsis. Plant J 63, 241-253 (2010).

230 Lifschitz, E. et al. The tomato FT ortholog triggers systemic signals that regulate growth and flowering and substitute for diverse environmental stimuli. Proc Natl Acad Sci U S A 103, 6398-6403 (2006).

231 Pnueli, L. et al. The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. Development 125, 1979-1989 (1998).

7 AUTHOR CONTRIBUTIONS

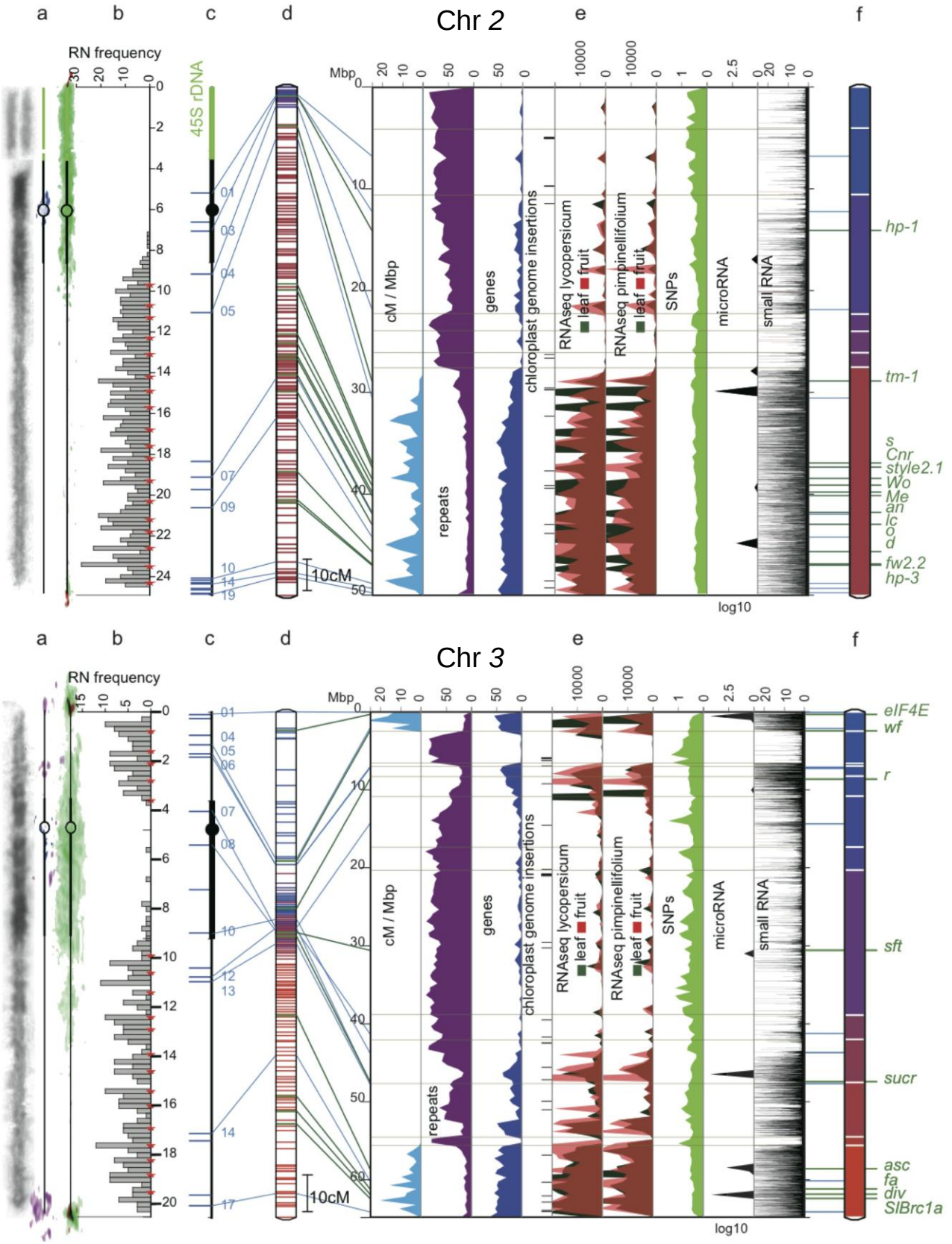
AAL, ABA, ABE, ABG, ABO, AFI, AGR, AHA, AHU, AJA, AJM, AJO, AKM, AKS, AKT, AMO, APA, ARI, ASI, ATR, AUS, AVE, BAR, BAW, BCK, BCM, BDK, BFQ, BMW, BTC, CBL, CCL, CCO, CDE, CGE, CGH, CHR, CKL, CLI, CLR, CMQ, CNI, CNO, CPE, CRI, CYL, DCA, DCH, DDW, DHK, DIC, DKU, DLF, DOZ, DRN, DSH, DSZ, DTO, DWB, DZA, EAS, ECA, EDA, EDP, EFA, EGS, FCA, FDP, FEA, FFU, FRE, FYF, FYU, FZN, GBS, GDB, GEB, GFA, GGI, GHE, GIG, GJB, GLI, GPE, GUL, GVA,GWA, HBE, HBT, HDJ, HEB, HFU, HGU, HHI, HLA, HLJ, HSC, HSL, IFI, IMO, IYT, JAL, JCA, JDS, JDW, JEK, JFA, JFR, JGA, JHE, JHZ, JJG, JKN, JKP, JKR, JLE, JLG, JLI, JLL, JMA, JMC, JPK, JRG, JRO, JTV, JUW, JVB, JVE, JVH, JVO, JWA, JYI, JZH, KAO, KCO, KGA, KJI, KKL, KMA, KML, KQW, KSH, KYA, LAM, LAS, LDA, LDP, LGE, LHY, LKA, LLO, LMA, LOG, LPX, MAB, MAF, MAV, MBO, MCC, MCO, MDA, MDB, MDF, MDR, MEG, MEL, MHM, MHU, MKP, MLC, MOR, MPH, MPI, MRE, MSI, MSP, MVS, MWI, MZO, NDA, NEA, NKS, NME, NVI, PCH, PDH, PFA, PFR, PJM, PKH, PKS, PPE, PRI, PTE, QFE, RAD, RAW, RBO, RBR, RCR, RDI, RFE, RGO, RGU, RHS, RKU, RLC, RMB, RMC, RMF, RMK, RSC, RVH, RWI, RZI, SAP, SBC, SBH, SCH, SDT, SGA, SGR, SHH, SHJ, SHL, SHU, SIK, SIS, SIY, SJO, SKA, SKE, SKP, SMA, SMK, SMR, SMS, SNE, SOS, SPA, SPL, SSA, SSM, SST, STO, SUR, SVA, SVY, SXH, SYK, SZH, SZU, TAL, TDA, THE, THL, TKA, TMO, TRS, TSC, TWT, TYO, UGO, VDO, VGU, VYF, WBB, WRM, WSA, WYA, XLI, XPA, XWA, XXU, XXW, YBA, YBX, YDU, YGU, YHL, YLI, YNA, YPW, YRE, YSH, YXU, YYU, YZH, ZBL, ZFE, ZJI, ZJL, ZYA, ZYE, ZZH were involved in data generation and/or analysis. AAL, AGR, AHP, AKT, AVE, BAR, CCO, CYL, DCH, DIC, DRN, DSZ, DWA,DZA, ECA, EDA, EDP, EGS, FEA, GBS, GEB, GHE, GIG, GJB, GVA, HDJ, HHI, HSC, IFI, JFR, JJG, JKR, JLE, JLG, JMC, JPK, JTV, JVE, KJI, KMA, LAM, LKA, MAB, MBO, MCO, MKP, MLC, MPI, MRE, MSP, MVE, MZO, NKS, RAW, RMK, RVH, RWI, SAP, SDT, SGR, SIK, SIY, SKN, SMR, SMS, SPA, SSA, SSM, TDA, TMO, TRS, TSC, WBB, YVP, YXU, ZBL, ZFE wrote the manuscript. AGR, AHU, AJA, AKS, AKT, ALA, AVE, BAR, BCM, BHA, CGH, CRO, CYL, DCH, DIC, DTO, DWA, DZA, FEA, FZN, GBS, GDB, GEB, GIG, GJB, GVA, GYJ, HDJ, HFU, HQL, HSC, JCP, JDW, JJG, JPK, JRO, JRZ, JVE, JWG, KMA, LAM, LFR, MAB, MBO, MCA, MPR, MSP, MVE, MVH, MVS, PJG, PKH, RAD, RGU, RGV, RHS, RMK, RVH, RWI, SAB, SDT, SHU, SMR, SMS, SPL, SSA, STA, TDA, TSC, VYF,WJS, WRM, XPA, YBX, YDC, YDU, YOX, YSL, YVP, YWA, ZBL, ZKC designed experiments, supervised data generation/analysis and managed subprojects/tasks.

doi:10.1038/nature11119

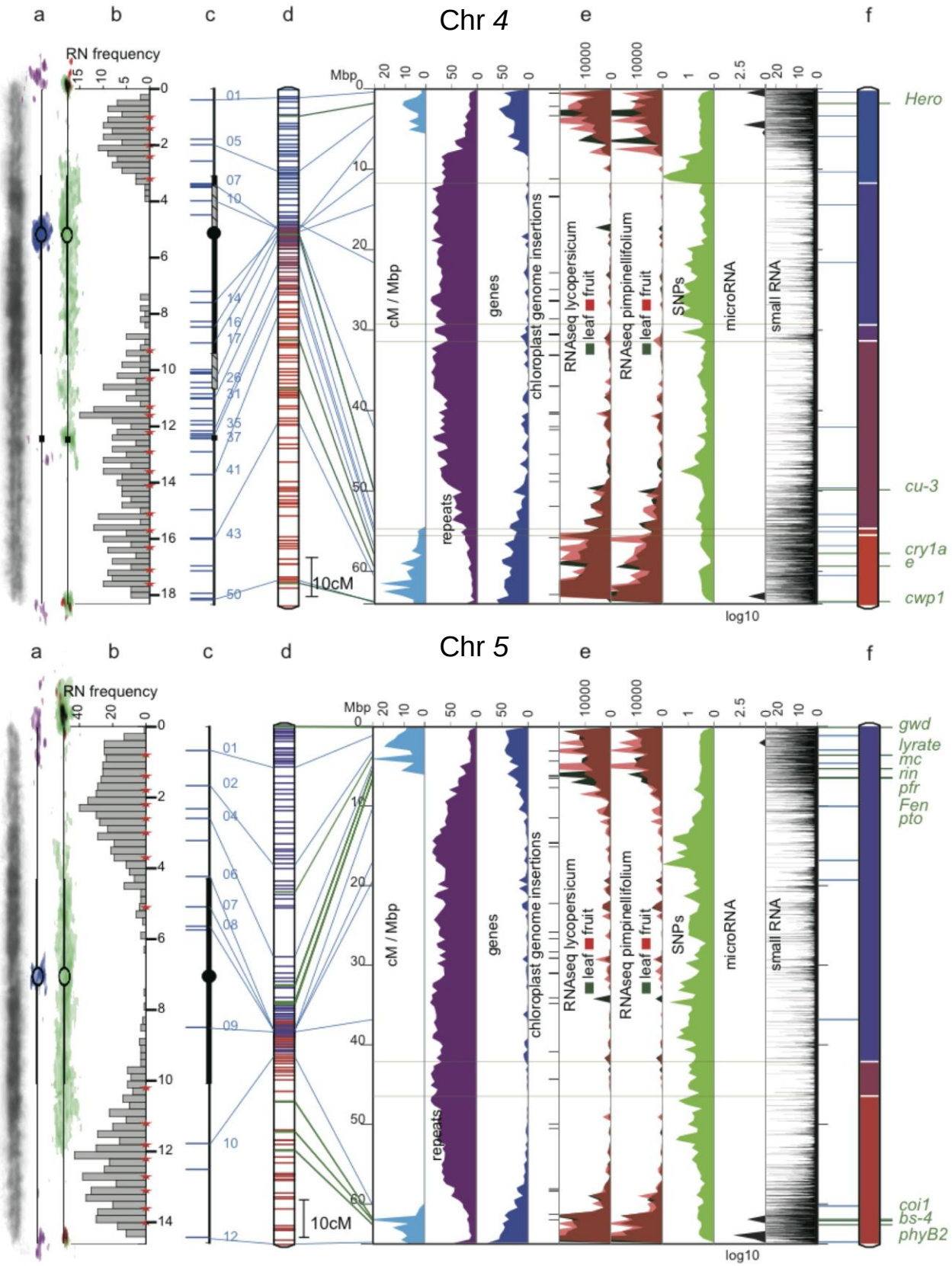
The tomato genome sequence provides insights into fleshy fruit evolution

The Tomato Genome Consortium (TGC)

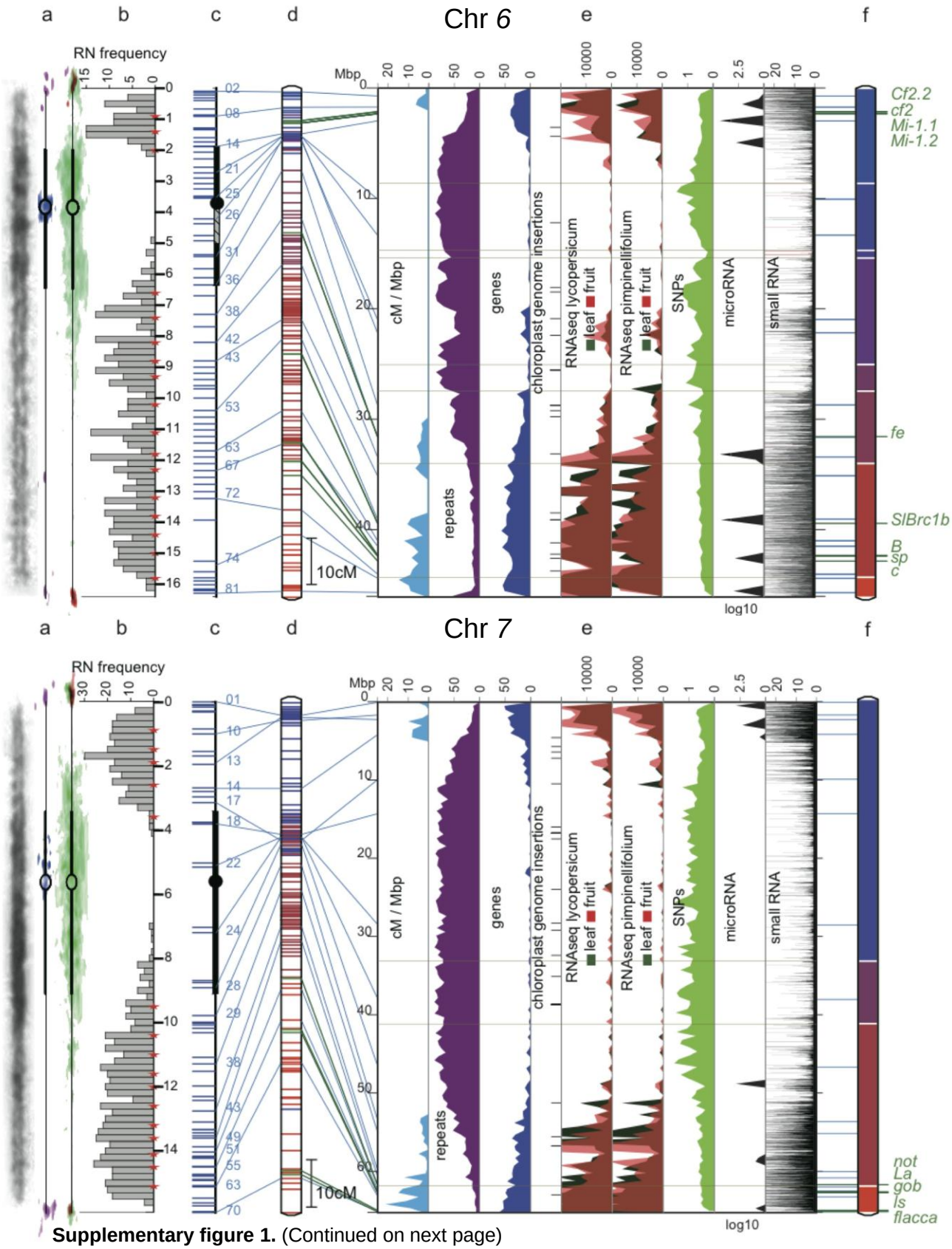
Supplementary figures and legends



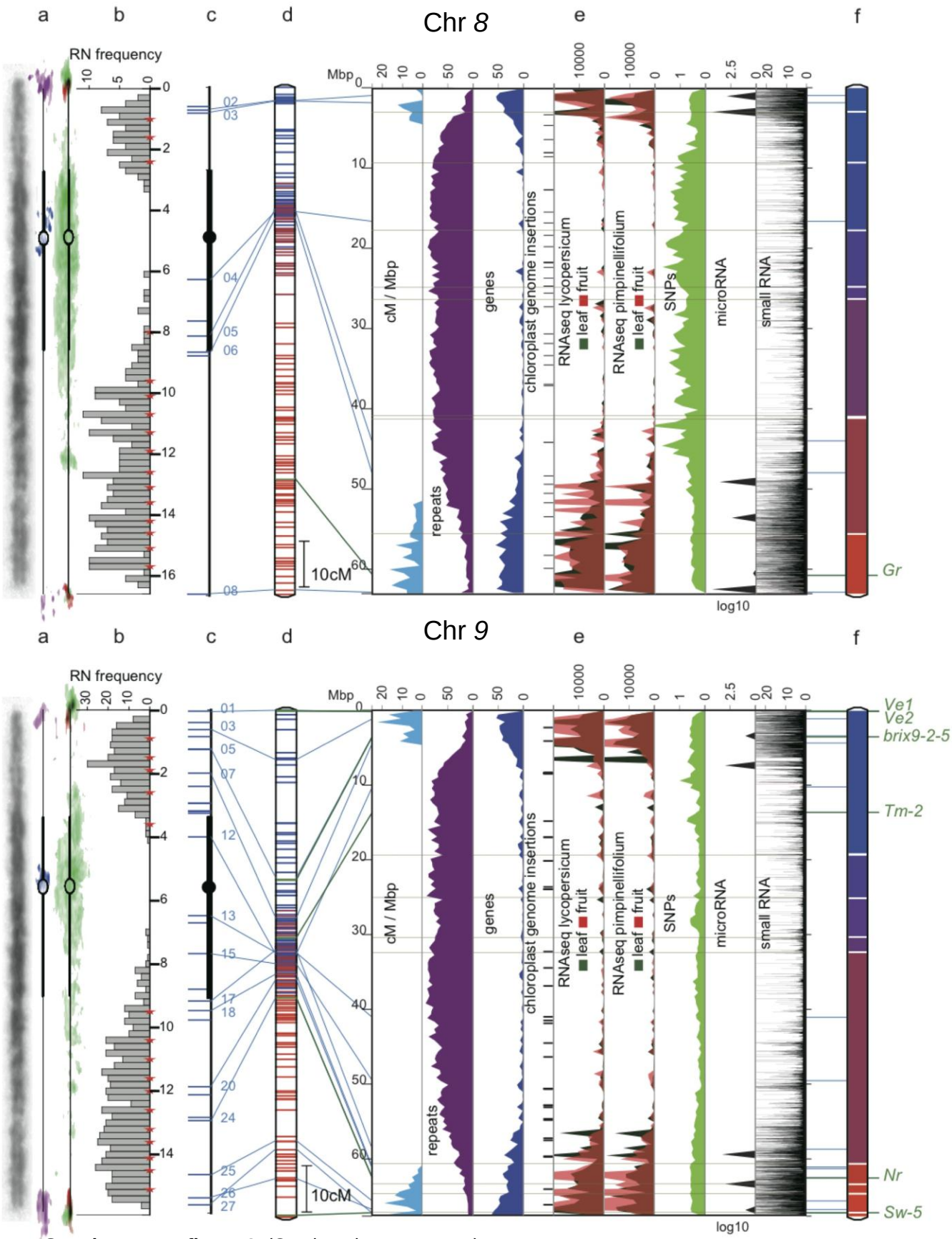
Supplementary figure 1. (Continued on next page)



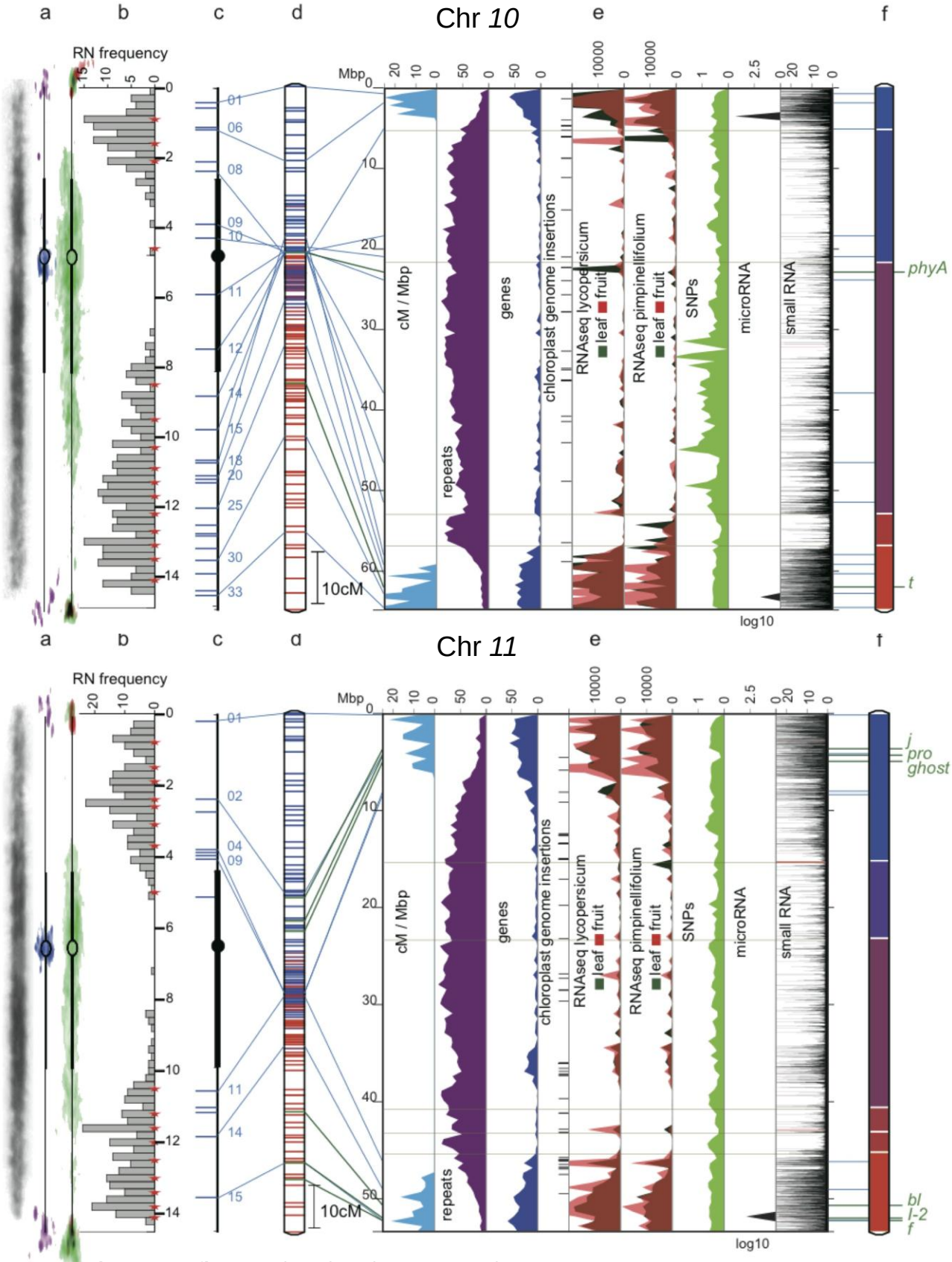
Supplementary figure 1. (Continued on next page)



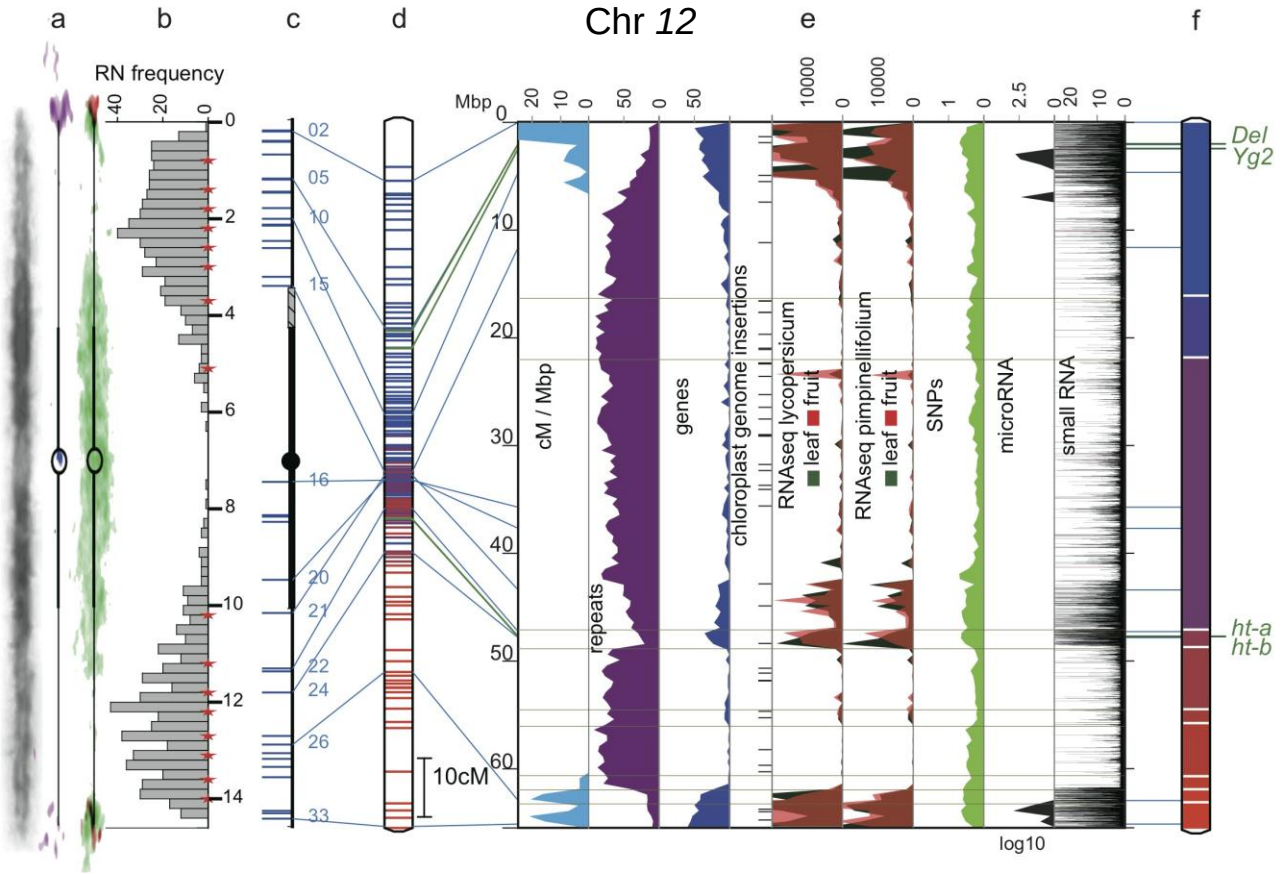
Supplementary figure 1. (Continued on next page)



Supplementary figure 1. (Continued on next page)

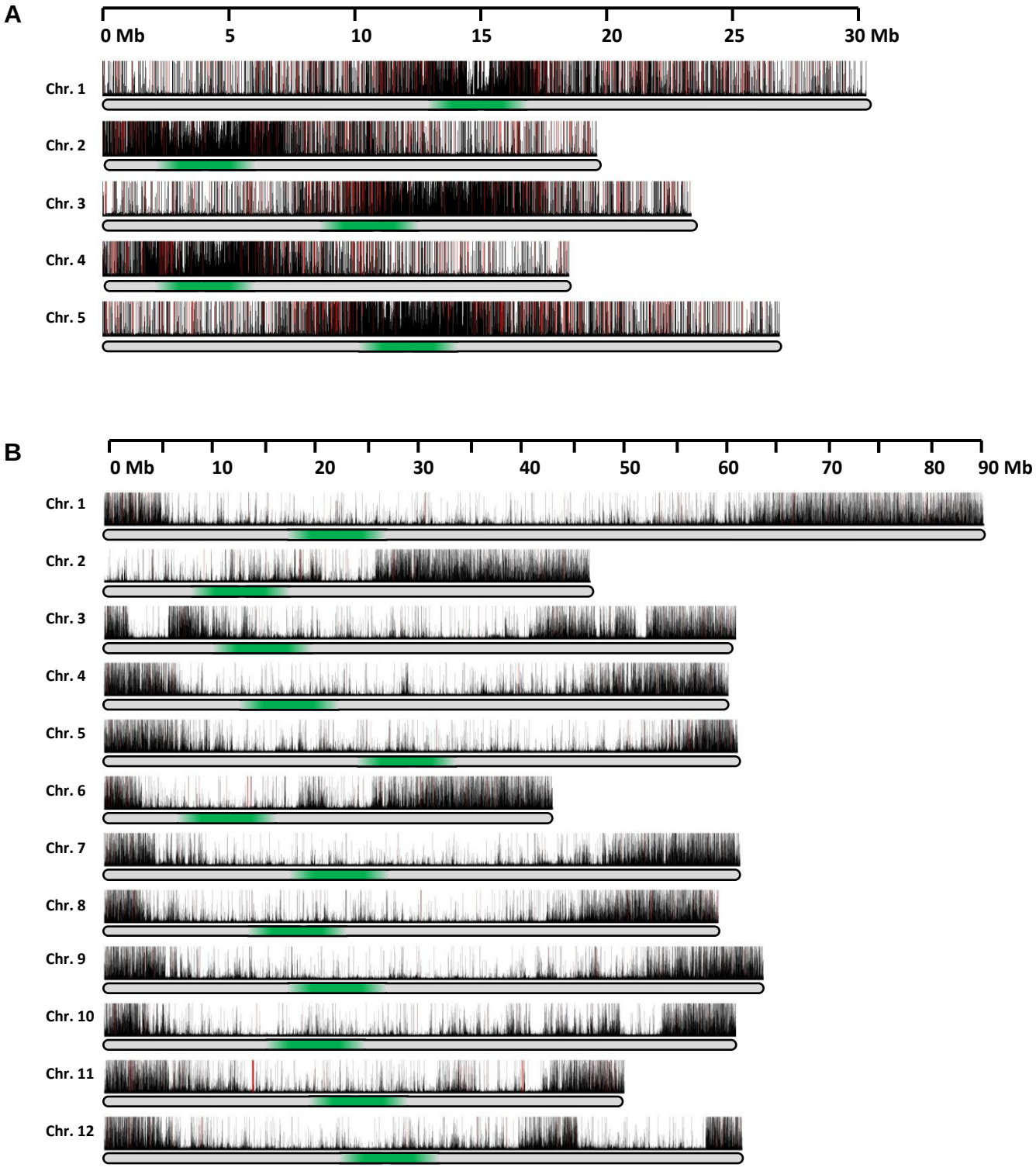


Supplementary figure 1. (Continued on next page)

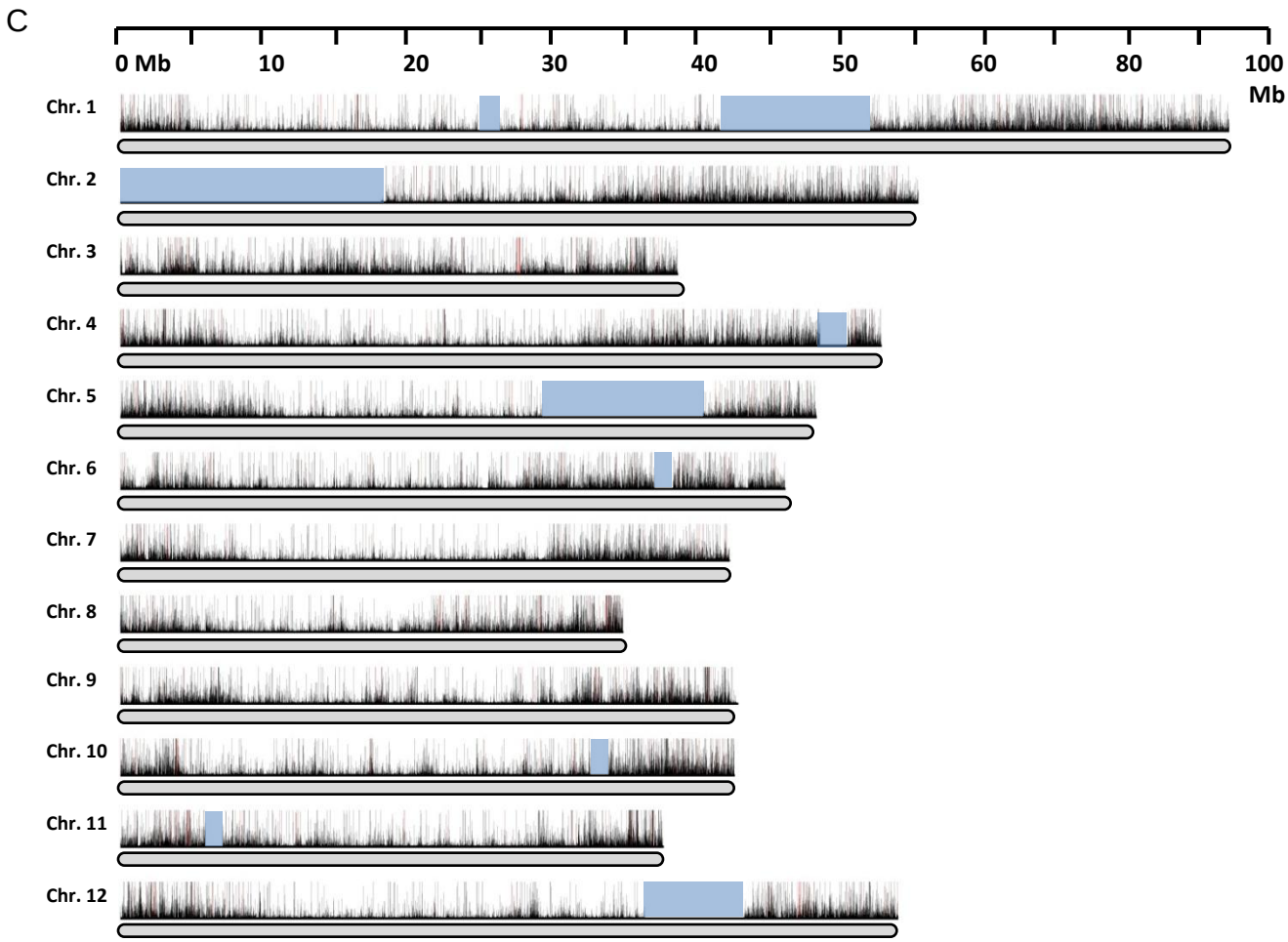


Supplementary figure 1. (Continued on next page)

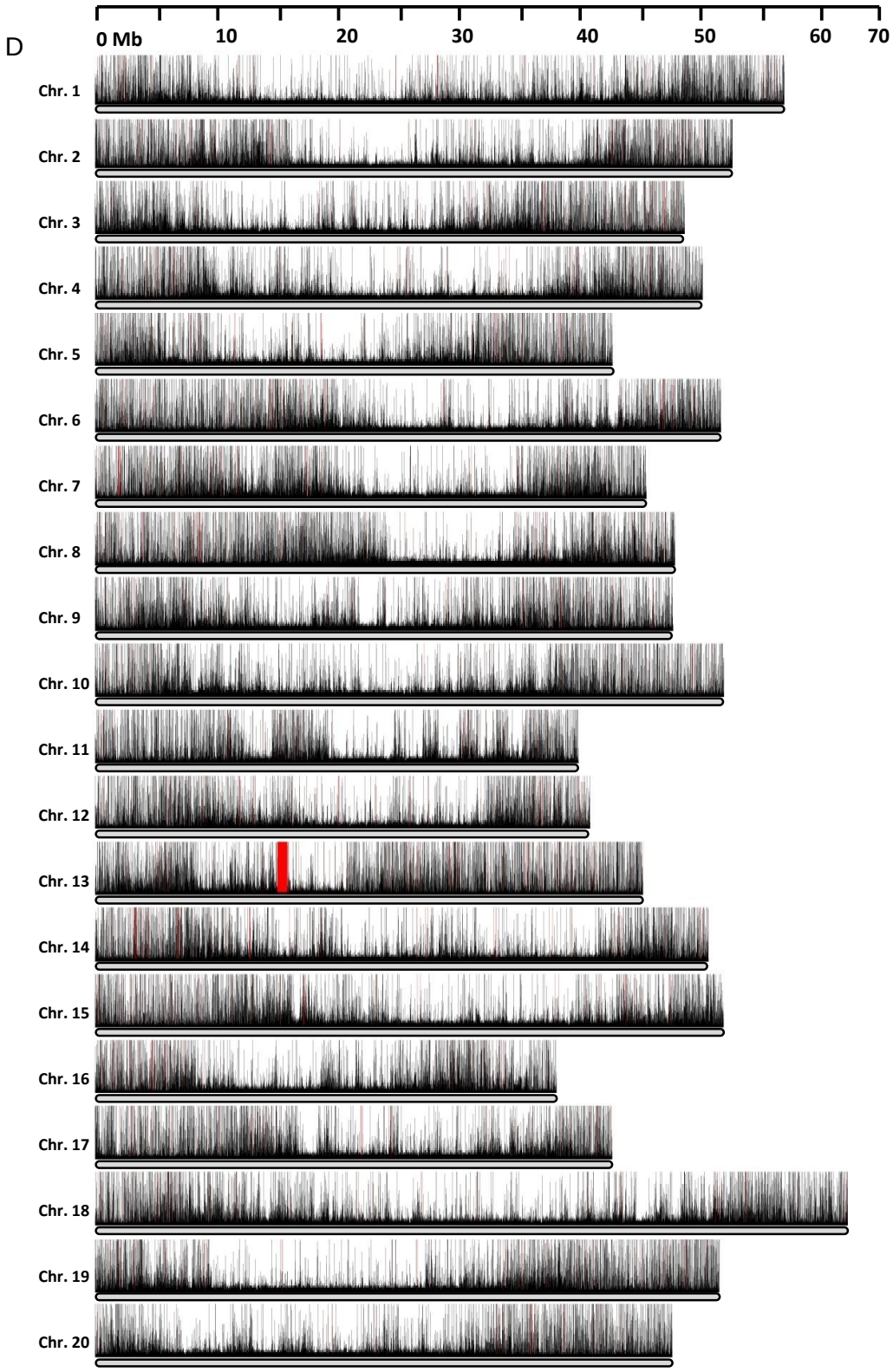
Supplementary Figure 1. Multi-dimensional topography of tomato chromosomes 2-12.
(a) Left: contrast-reversed, DAPI-stained pachytene chromosome prepared by synaptonemal complex (SC) spreading (**Supplementary Section 1.13**). Large blocks of darkly stained pericentric heterochromatin and distal lighter euchromatin flank the centromere; centre and right: FISH signals for repeat sequences on diagrammatic pachytene chromosomes (centromere: empty circle, pericentric heterochromatin: thick lines; distal euchromatin: thin lines): TGR1 is purple, TGR4 is blue, telomere repeat is red, and Cot 100 DNA (including most repeats) is green.
(b) Frequency distribution of recombination nodules (RNs). Each RN represents a crossover. The distances between adjacent red stars mark 5 cM intervals starting from the end of the short arm (top). Most RNs (and map distance) occur in distal euchromatin. Scale is in micrometers.
(c) FISH-based locations of selected BACs (horizontal blue lines on the left). Representative BACs with mapped sequences are connected by blue lines to their corresponding locations on (d), (e) and (f).
(d) Kazusa F2-2000 linkage map (**Supplementary Section 1.10**). Blue lines to the left connect linkage map markers on the (c) BAC-FISH map, (e) heat maps and (f) DNA pseudomolecule. Green lines to the right link the positions of tomato genes identified by map-based cloning (**Supplementary Table 14**) to their heat map locations (e) and the pseudomolecule (f).
(e) From left to right: linkage map distance (cM/Mb, turquoise) is concentrated in euchromatin; repeated sequences (% nucleotides/500 kb, purple) are much more common in pericentromeric heterochromatin; genes (% nucleotides/500 kb, blue) are much more common in euchromatin; chloroplast insertions; Illumina RNA-Seq reads from leaves and breaker fruits of *S. lycopersicum* and *S. pimpinellifolium* (number of reads/500kb, green and red, respectively) map mostly to euchromatin; microRNA (Transcripts per million/500 kb, black) genes occur in few, evenly distributed peaks; small RNAs (thin horizontal black lines) are primarily derived from euchromatin (sum of hits-normalized abundances (HNA) abundance divided by genome matches, normalized to transcripts per four million (TP4M) for all small RNAs, summed in 10 kb bins). Horizontal grey lines represent gaps in the pseudomolecule (f).
(f) DNA pseudomolecule. Unsequenced gaps are indicated by white horizontal lines. Tomato genes identified by map-based cloning (**Supplementary Table 14**) are indicated on the right.



Supplementary figure 2. (Continued on next page)

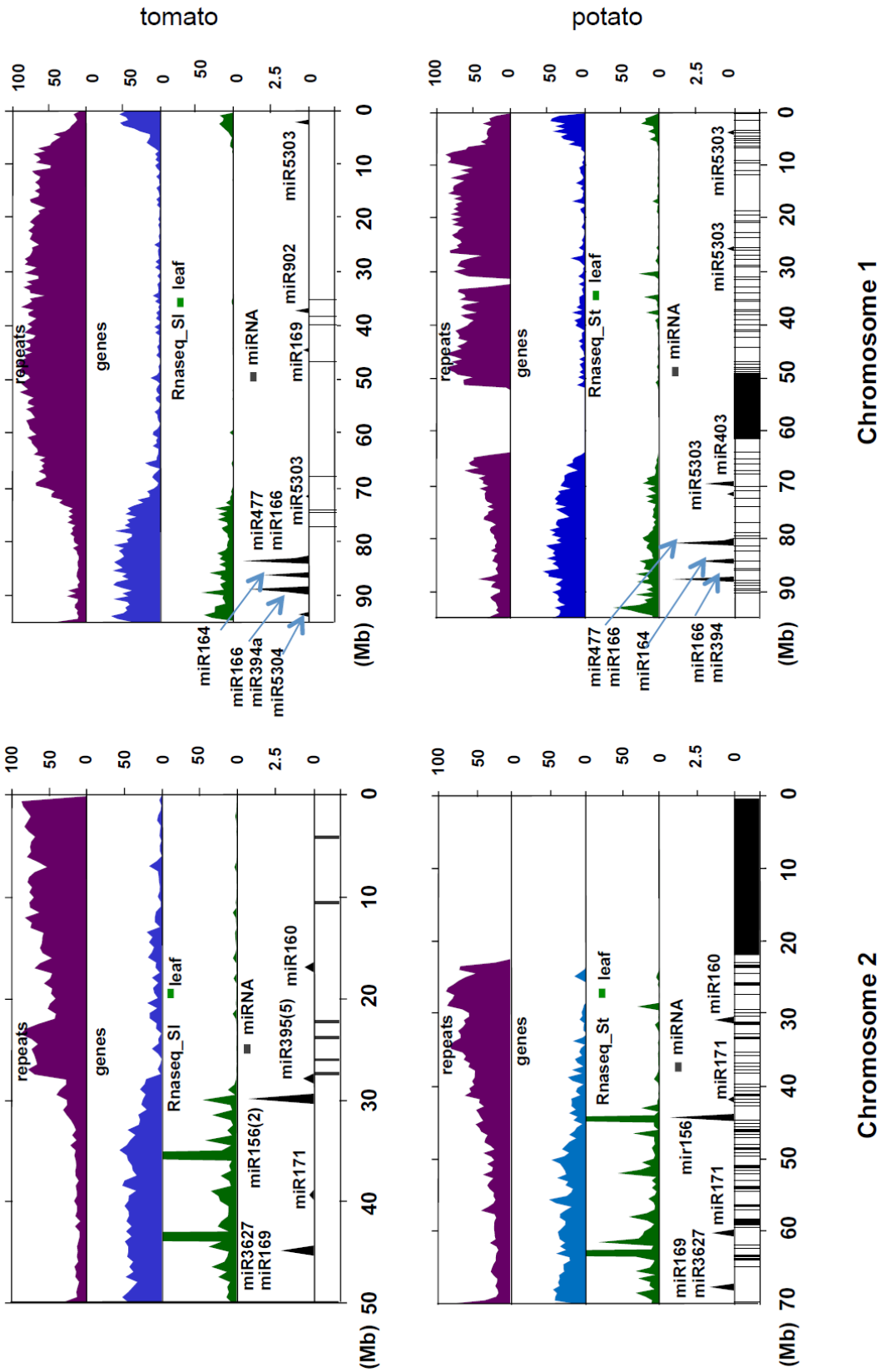


Supplementary figure 2. (Continued on next page)

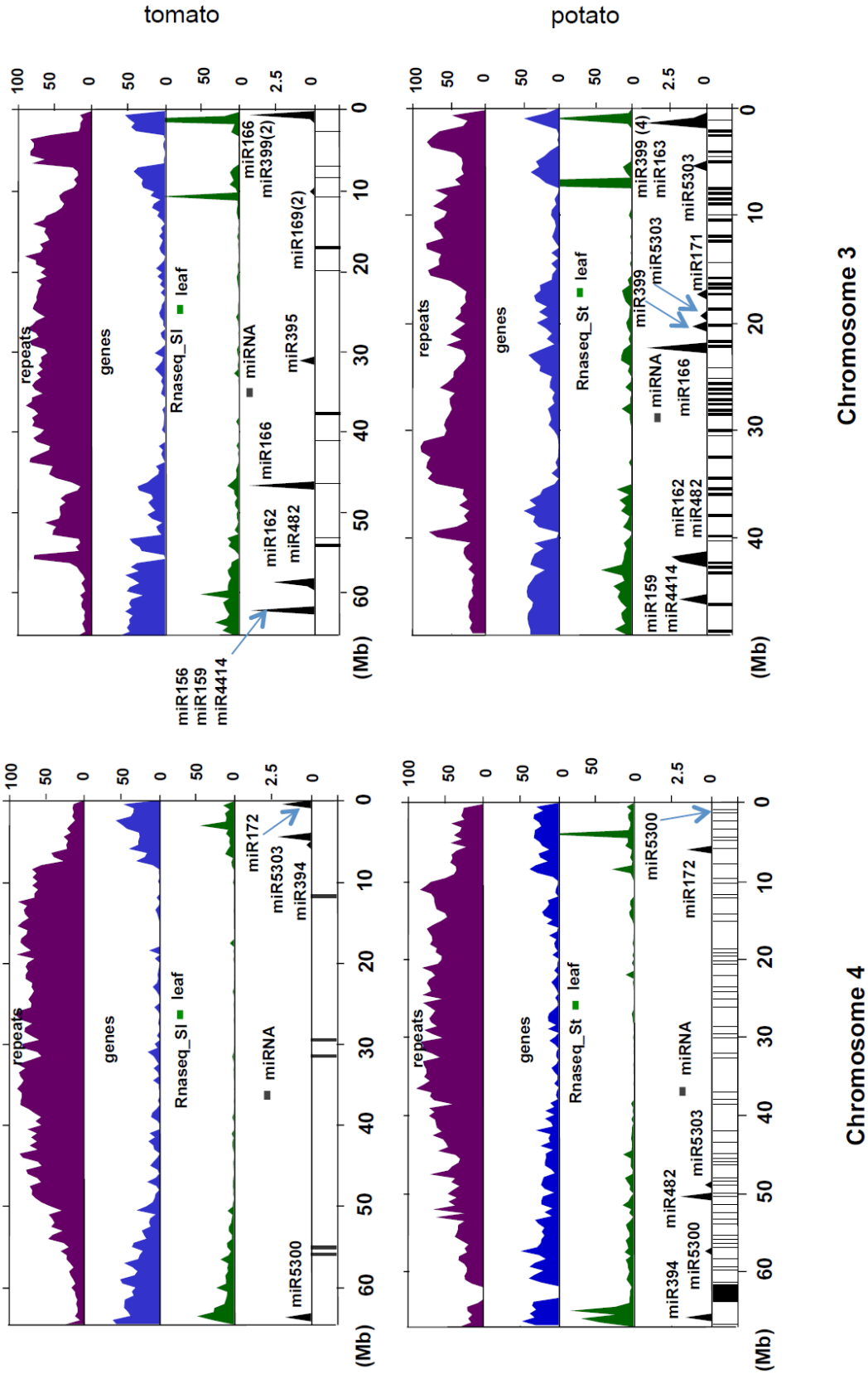


Supplementary figure 2. (Continued on next page)

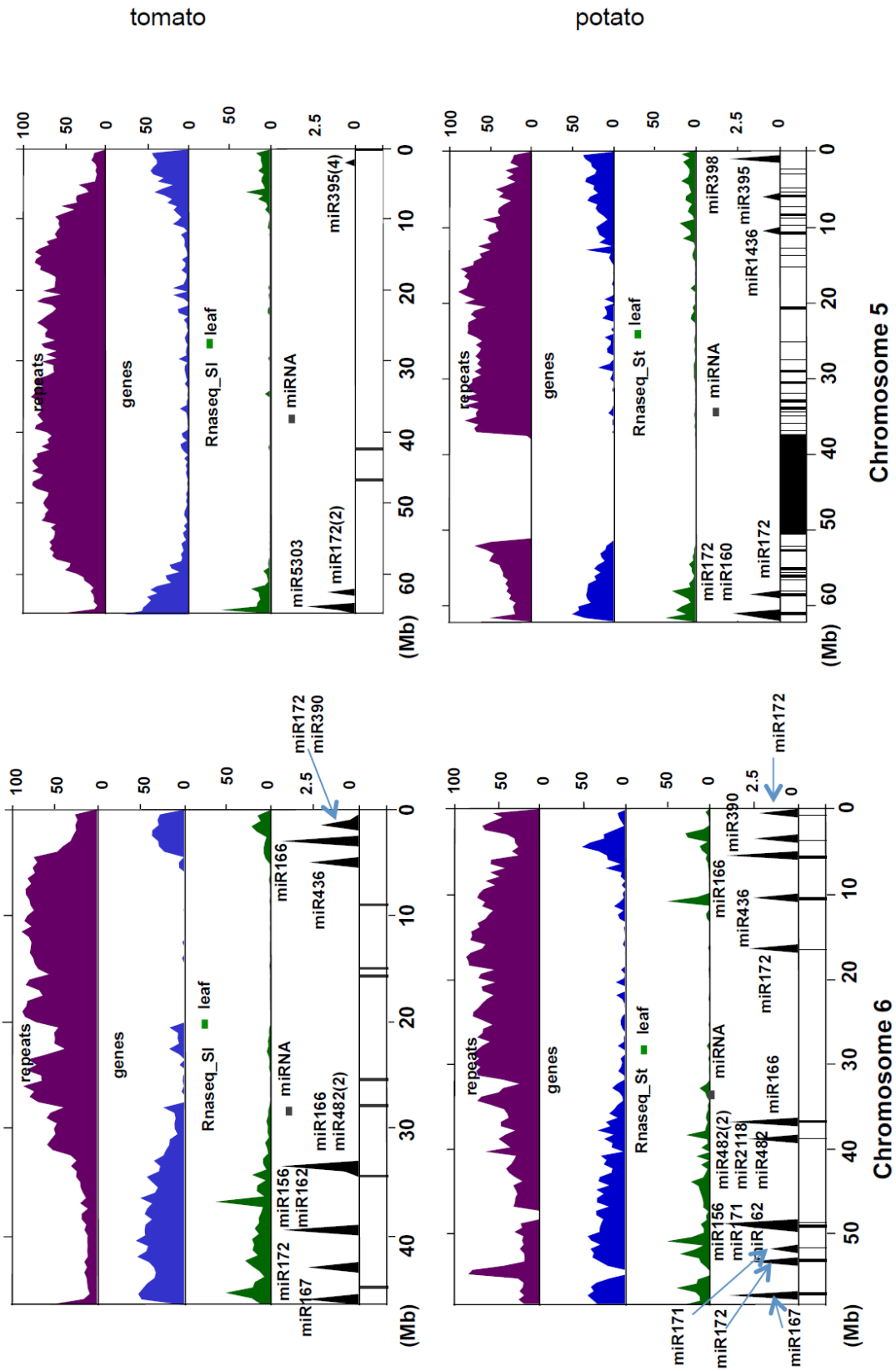
Supplementary Figure 2. Normalized distributions of small RNAs along the chromosomes of Arabidopsis, tomato, potato and soybean.
A. Arabidopsis small RNAs are highly abundant in the pericentromeric region of the chromosomes. Approximate centromeric location indicated by green shading. Small RNAs are from wild type flower, normalized abundances divided by the number of matching genomic locations ("hits-normalized abundances" or HNA) are indicated by the vertical bars above each chromosome ideogram. The scale in megabases is indicated above. Data are from Genbank record GSM280228.
B. Tomato small RNA levels as measured by HNA (as in panel A) are reduced in the pericentromeric region and most regions of the genome have either high or low levels of small RNAs with sharp transitions between regions with these abundances. The small RNA data are shown for the breaker-stage fruit library (SLY3).
C. Potato small RNAs show a similar bifurcated distribution to tomato; although pericentromeric regions are not marked, the HNA small RNA abundances are highest in the gene-rich, more telomeric regions of the chromosomes. Blue blocks indicate gaps inserted into the potato genome for alignment purposes.
D. Soybean small RNAs show a more continuous distribution than tomato, but also show a higher level of small RNAs in the gene-rich more telomeric regions of the chromosomes compared to the pericentromeric regions (not marked). Data are from soybean flowers, Genbank GEO record GSM769284.



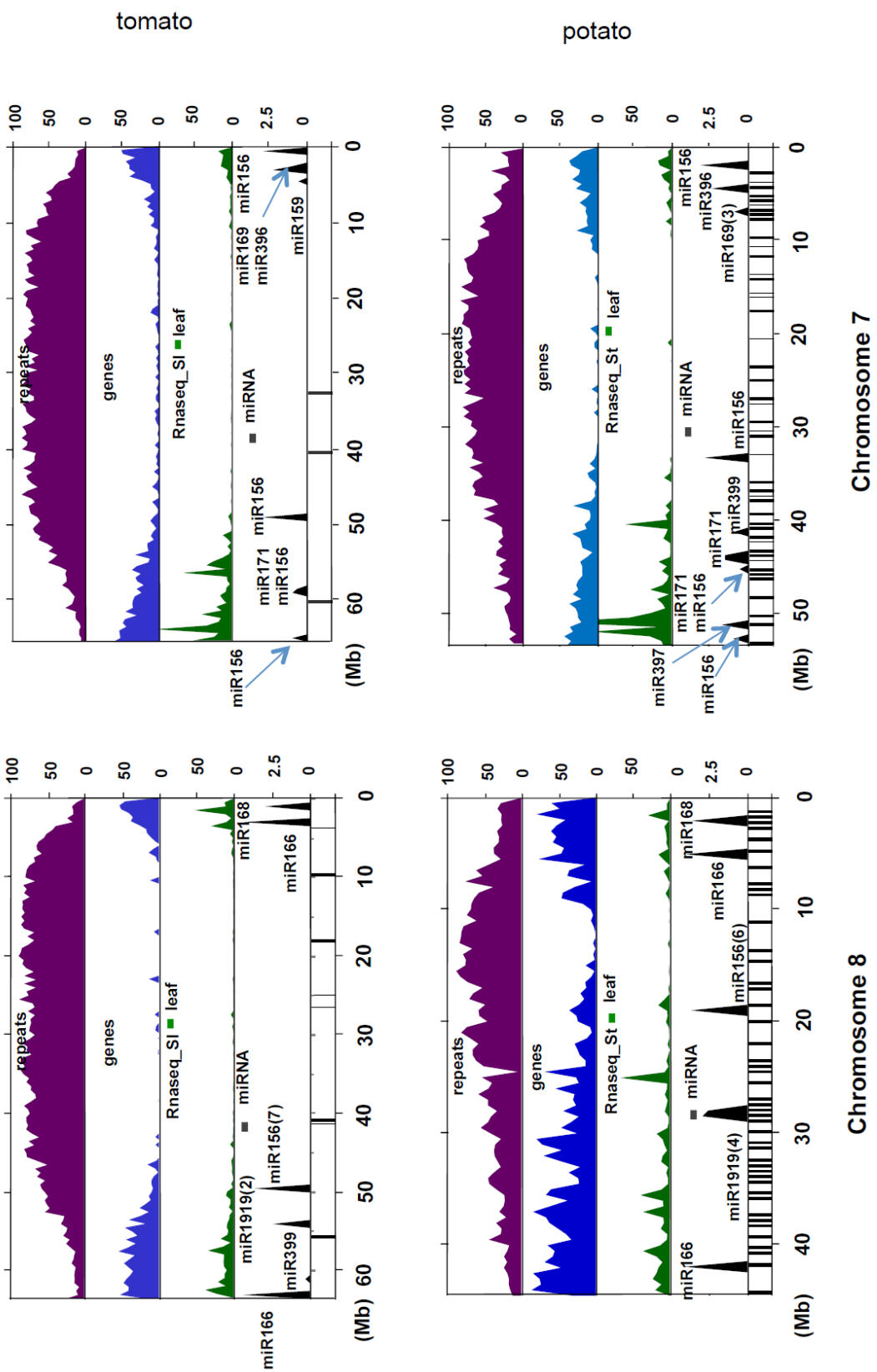
Supplementary figure 3. (Continued on next page)



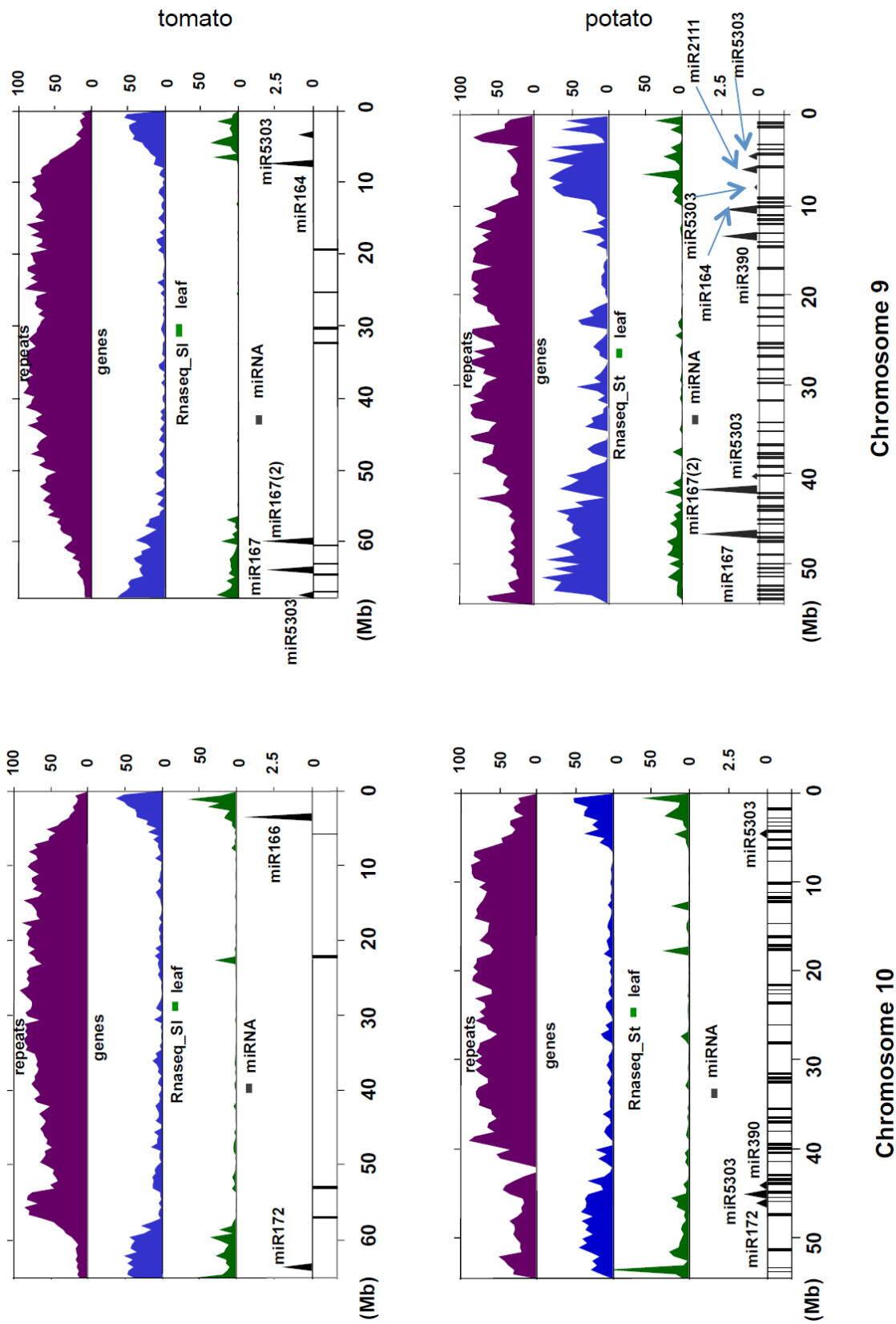
Supplementary figure 3. (Continued on next page)



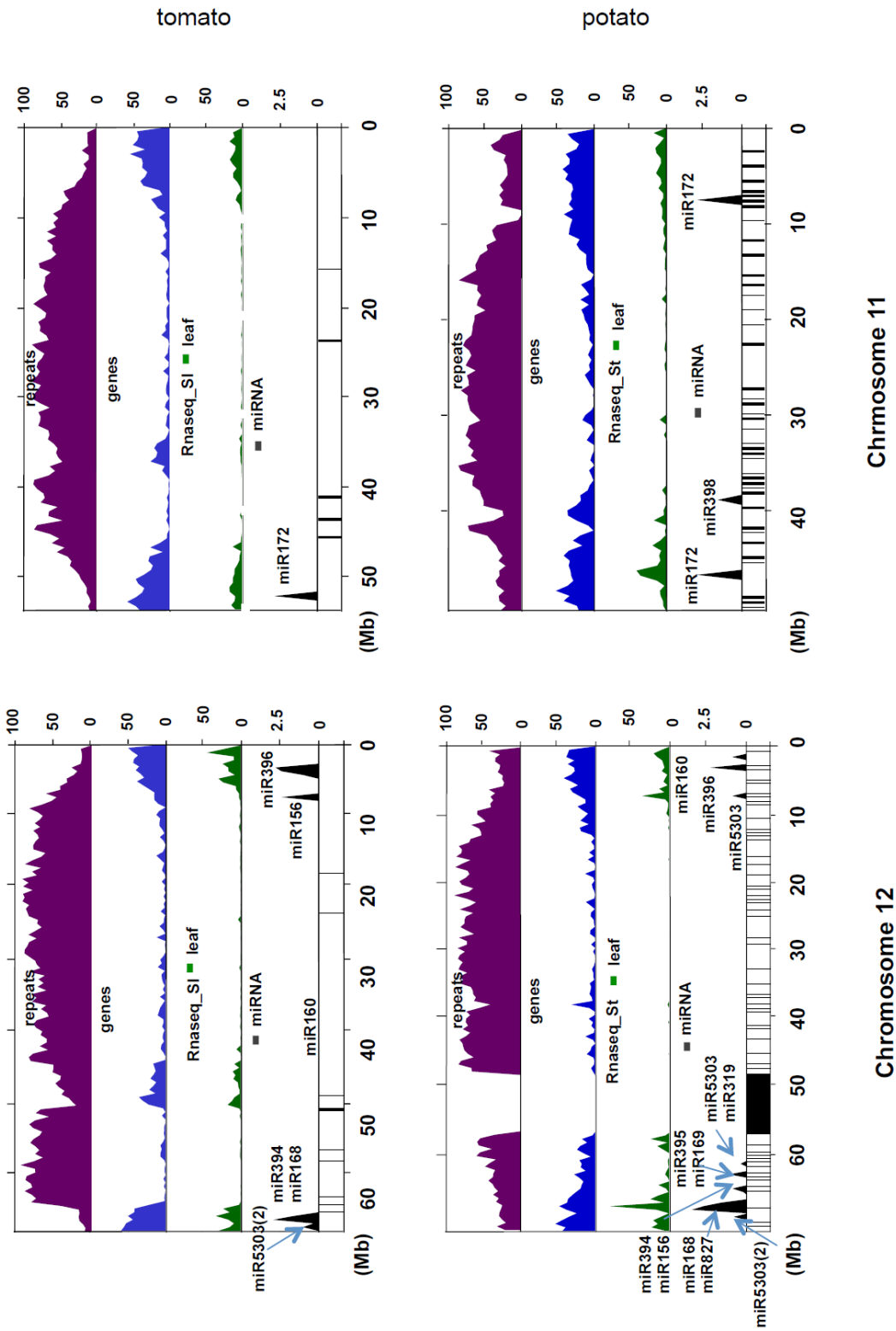
Supplementary figure 3. (Continued on next page)



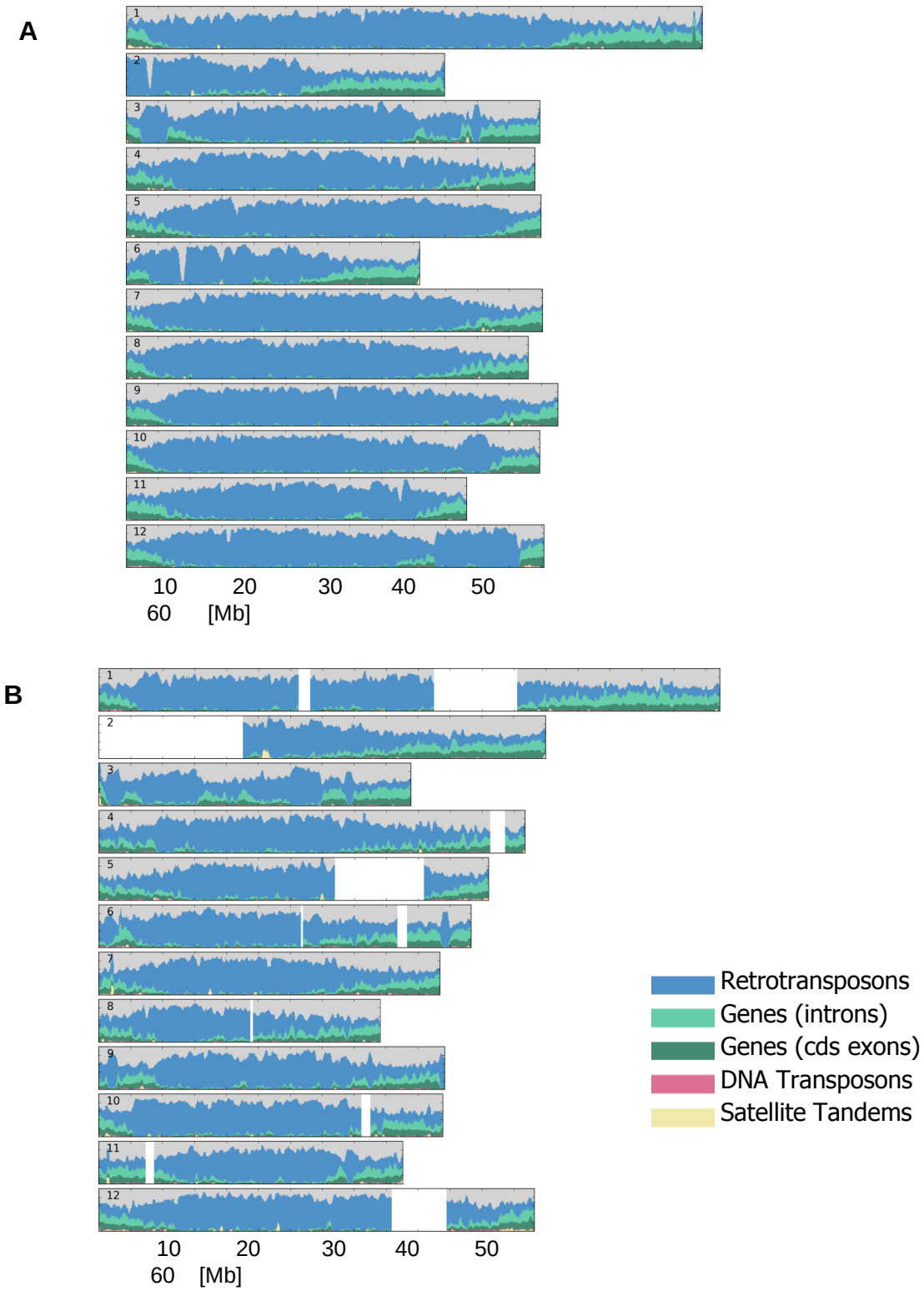
Supplementary figure 3. (Continued on next page)



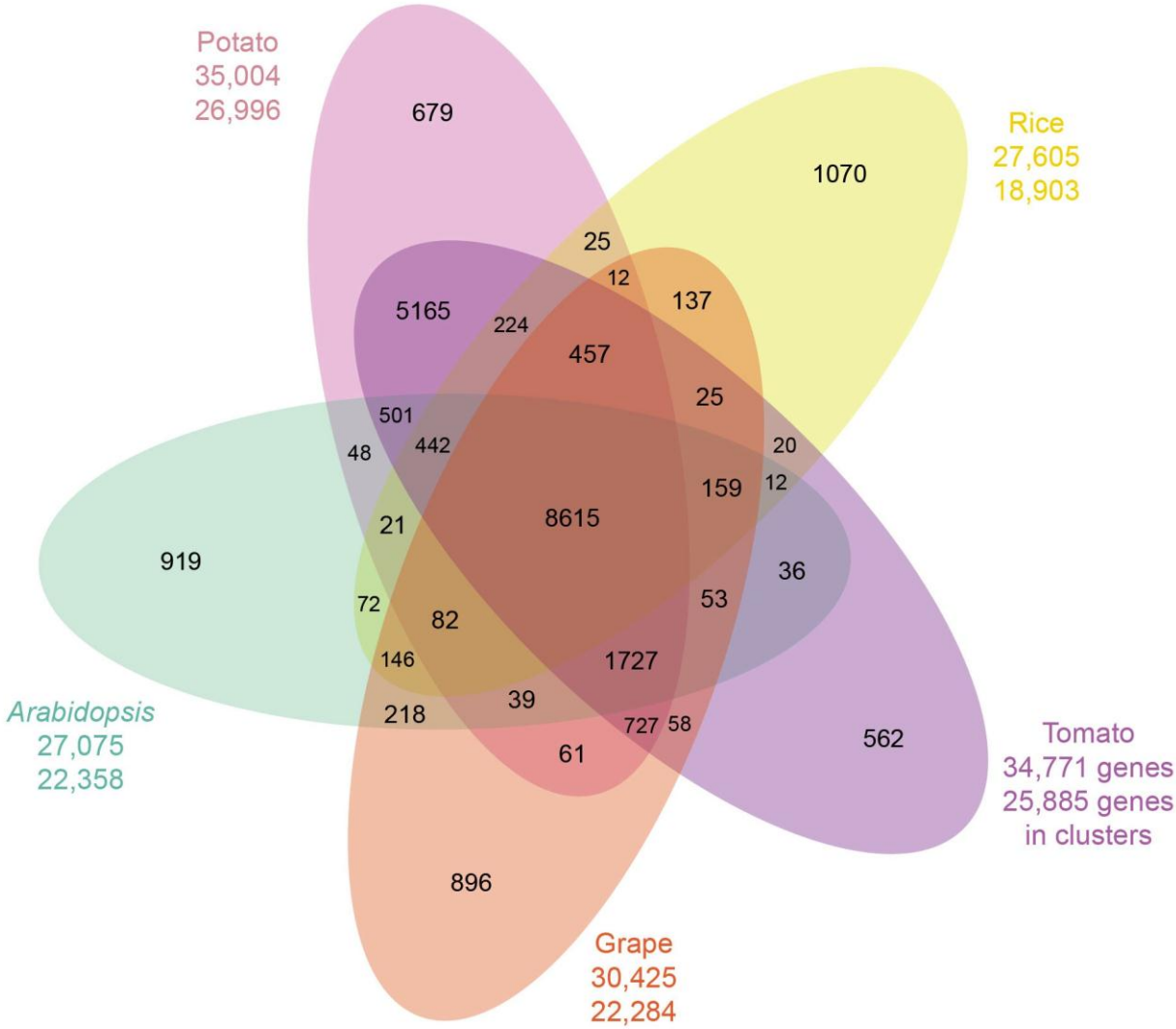
Supplementary figure 3. (Continued on next page)



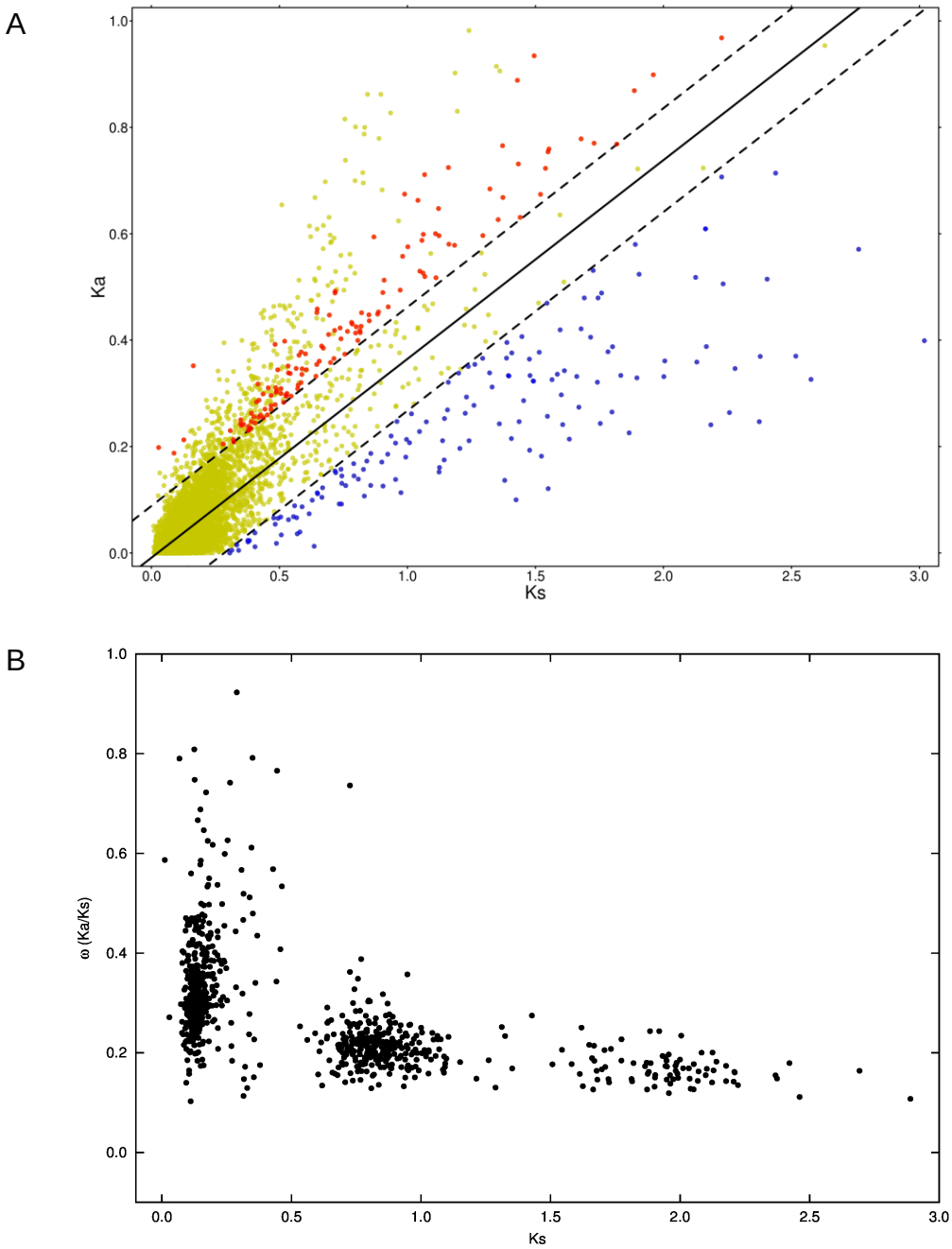
Supplementary Figure 3. Comparison of chromosome features between tomato and potato. Data are reported according to Panel “e” of **Figure 1**. miRNA names are indicated. Gaps between the scaffolds are indicated by black lines on the right of each panel. Large, gap-rich regions in the potato genome are shown as black boxes.



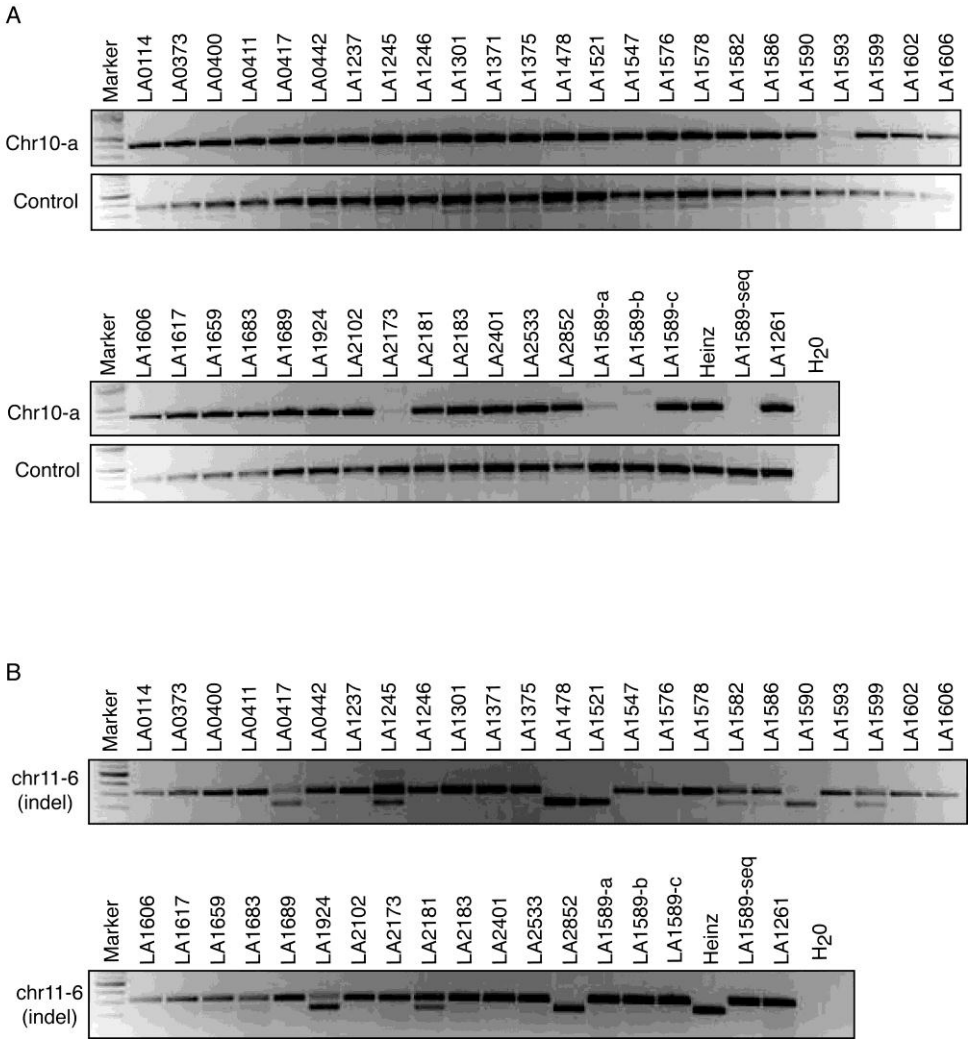
Supplementary Figure 4. Chromosomal Landscapes of Tomato (A) and Potato (B). The different colours show the relative proportion (in %) of each of the four main genomic components per 0.5 Mb window as stacked bar-chart. Including the unassigned (grey) regions all values add up to 100% per window.



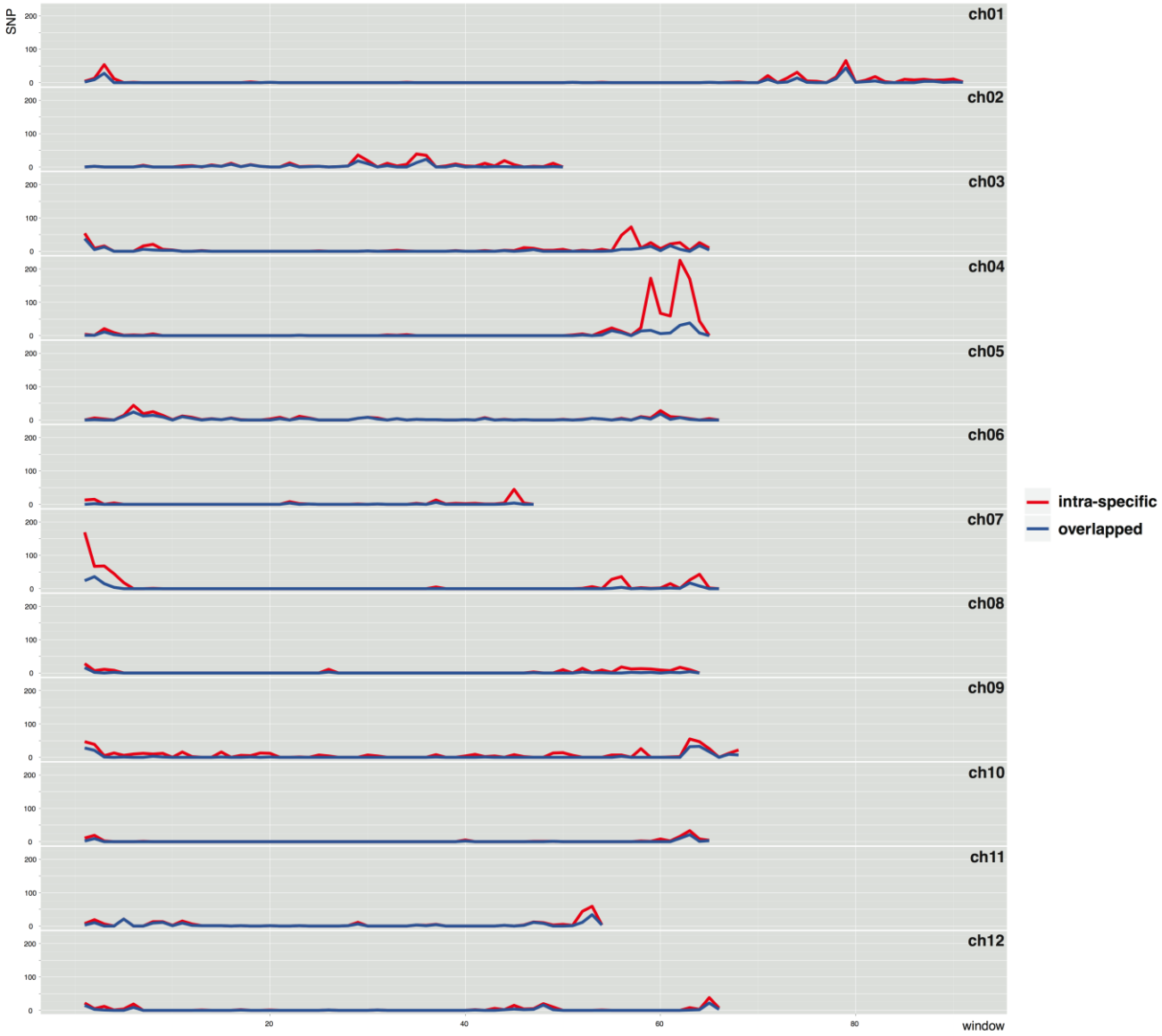
Supplementary Figure 5.
Distribution of orthologous gene families in tomato, potato, grapevine, Arabidopsis and rice, calculated with OrthoMCL[182]; BLASTP e-value cut-off 1e-05 and MCL inflation parameter of 1.5). The annotations used were from iTAG (tomato and potato, this paper), RAP2 (rice), TAIR9 (Arabidopsis) and Genoscope (grape).
A total of 154,880 sequences from the five different organisms was clustered into 23,208 gene-groups (plus singletons) using OrthoMCL. In each intersection of the Venn diagram the number of gene-groups ("families") are represented. Of the 34,771 protein-coding genes predicted for tomato, 25,885 were clustered in a total of 18,783 gene-groups (≥2 members); From the 18,783 gene-groups 8,615 are common to all five genomes while 1,727 gene-groups are confined to eudicots (tomato-potato-grape-Arabidopsis). Within the eudicots, 727 are restricted to plants with fleshy fruits (tomato-potato-grape); In total 5,165 gene-groups contain genes from the Solanaceae only whereas 562 groups contain genes from tomato only and 679 groups from potato only.



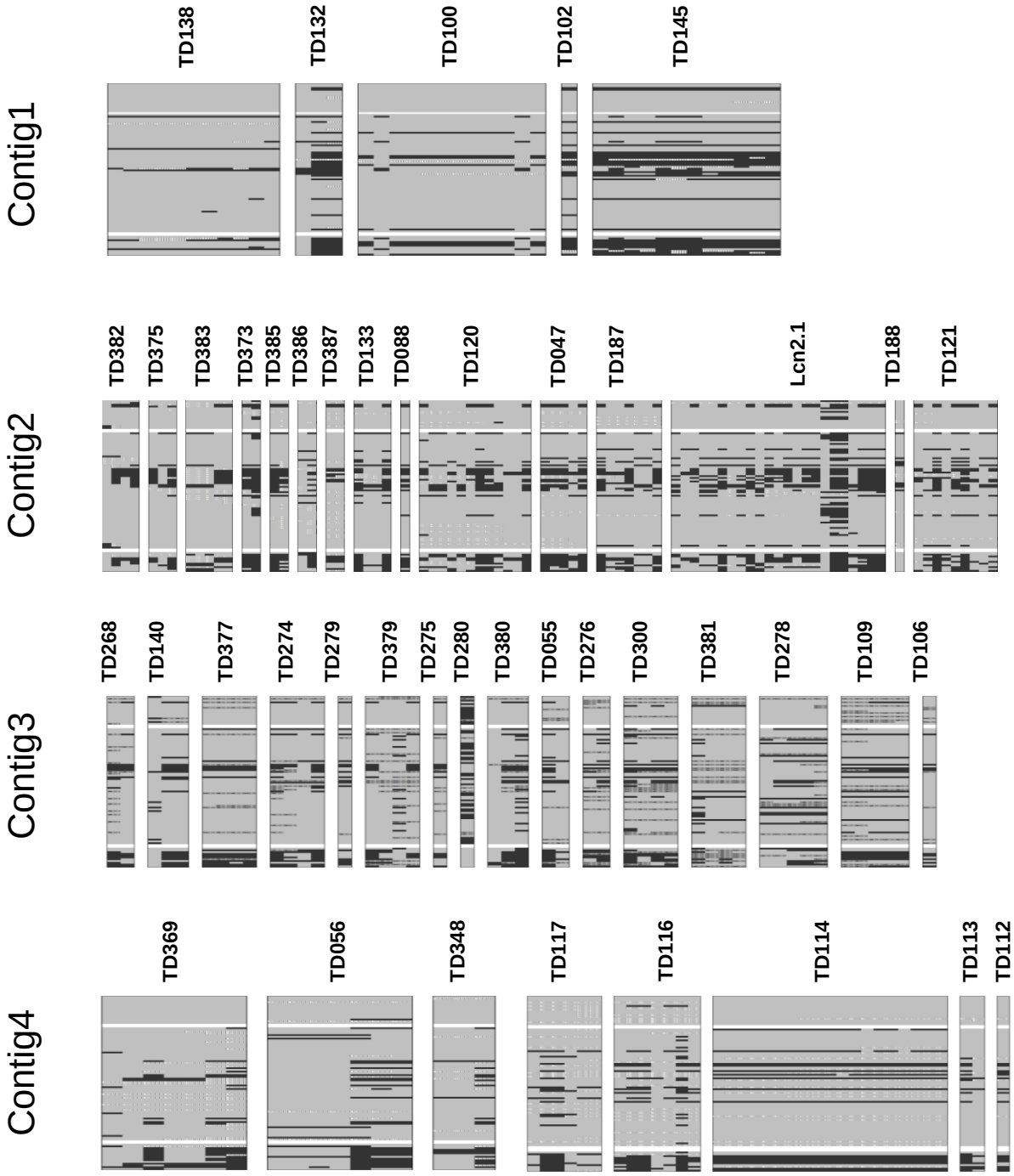
Supplementary Figure 6. Solanaceae comparative genomics.
A. Scatter Plot of Ks vs. Ka of orthologs between tomato and potato. The dashed line represents the 95% prediction interval about the linear regression. Red and blue dots represent high and low *ω* (*Ka/Ks*) gene pairs, respectively. Yellow dots above the 95% prediction interval failed to pass a two-tailed Fisher's exact test, indicating that the number of base substitutions was too small or the numbers of synonymous and non-synonymous substitutions were too similar to be deemed high *ω.*
B. Collinear blocks can be divided into 3 groups by *Ks* values, with those > 1.5 attributed to pan-eudicot triplication. The mean potato-tomato *ω* of all gene pairs in these blocks is 0.176±0.091, suggesting that the remaining genes are now largely under purifying selection.



Supplementary Figure 7. Newly discovered genomic variation in *S. pimpinellifolium* germplasm.
A. Multiple accessions of *S. pimpinellifolium* and independently derived single seed descent stocks of LA1589 were tested for the presence or absence of an approximately 5 Mb deletion on chromosome 10. Only two of 37 accessions appear homozygous for the deletion (LA1593 and LA2173); however, the presence of this chromosomal segment in independent stocks of LA1589 suggests that the deletion is likely polymorphic both within and between accessions. This might explain why homozygosity for the deletion was only found in LA1593 and LA2173, if the pool of DNA used for PCR from other accessions included individuals that are polymorphic for the deletion. Chr10-a represents one of 20 PCR markers used to verify the presence of the Chr. 10 deletion in LA1589.
B. A randomly selected newly identified insertion-deletion (indel) marker from chromosome 11 distinguishing the genomes of *S. lycopersicum* cv. 'Heinz 1706' and *S. pimpinellifolium* LA1589 was used to test whether intra-specific polymorphism in *S. pimpinellifolium* extends beyond the Chr. 10 deletion. Indel marker chr11-6 reveals a pattern of polymorphism distinct from that of the Chr. 10 deletion, and further shows the existence of heterozygosity in the *S. pimpinellifolium* germplasm. DNA was prepared from pools of four plants.

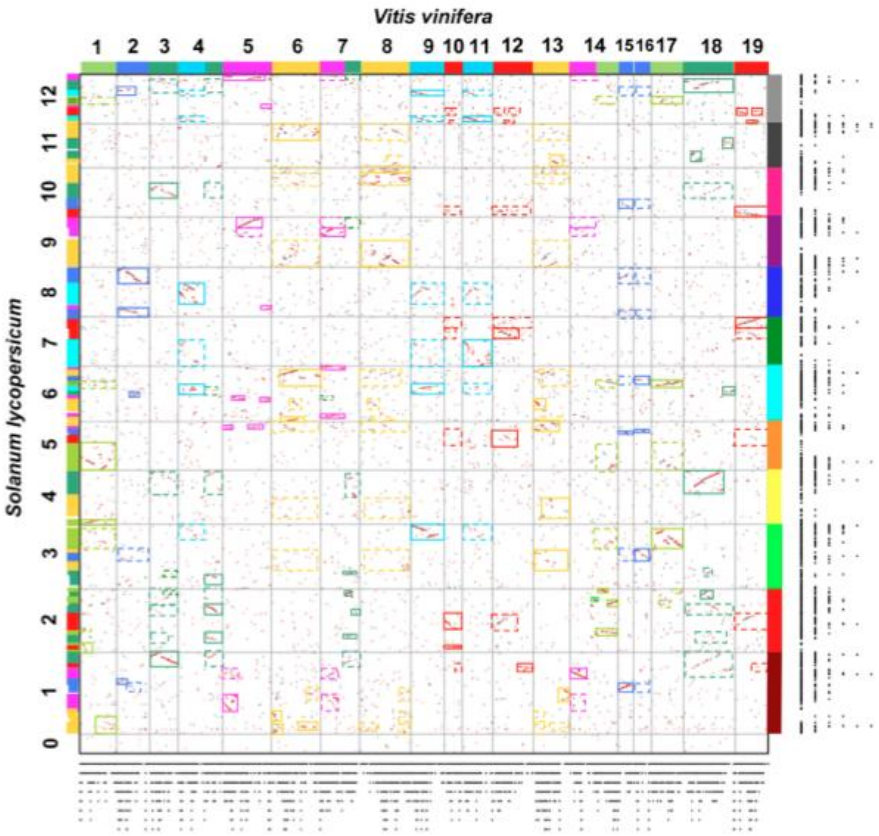


Supplementary Figure 8. Chromosomal distributions of *S. lycopersicum* intra-specific SNPs and their overlap with *S. pimpinellifolium* SNPs (1 Mbp window). Intra-specific polymorphisms (SNPs) were overlaid with *S. pimpinellifolium* SNPs to investigate the overlap of SNP position and base change. The same variant base in *S. pimpinellifolium* and intra-specific accessions of *S. lycopersicum* (alterative allele from the reference 'Heinz 1706' genome) at the same positions were designated as overlapping SNPs, which were then plotted simultaneously with intra-specific SNPs in 1 Mb window along each chromosome.

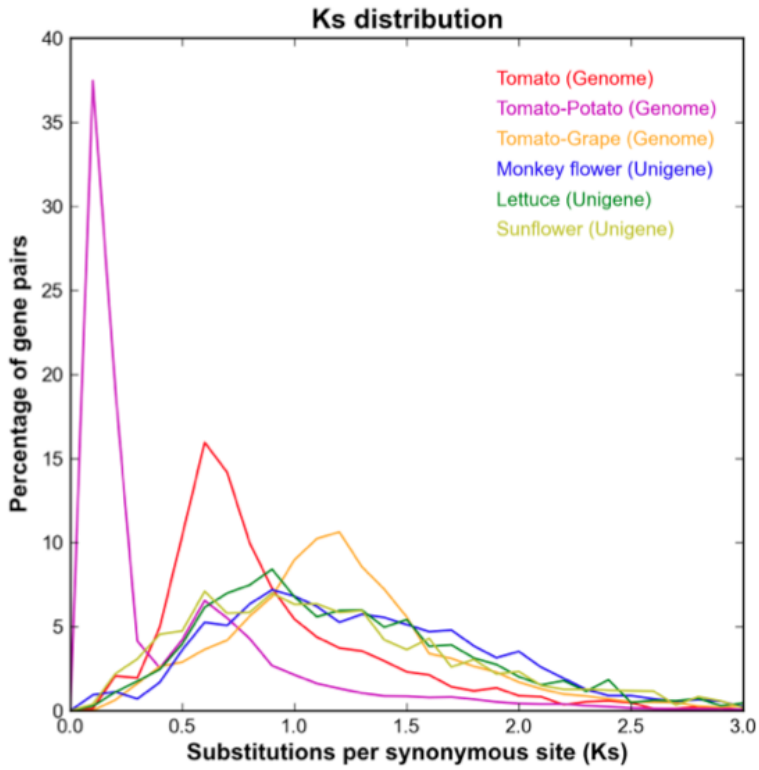


Supplementary Figure 9. Graphical haplotypes of 90 accessions for markers located on four physical contigs on chromosome 2. Rows represent accessions and columns represent polymorphic sites. The 44 DNA sequenced fragments are separated by white rows. For each polymorphic site, the most frequent allele is represented in light gray and the alternative allele is represented in black. Missing data are represented in white. The three groups of accessions comprising tomato (17 top lines), cherry tomato (63 middle lines) and *S. pimpinellifolium* (10 bottom lines) are separated by continuous white lines (adapted from Ranc *et al*, unpublished data).

A

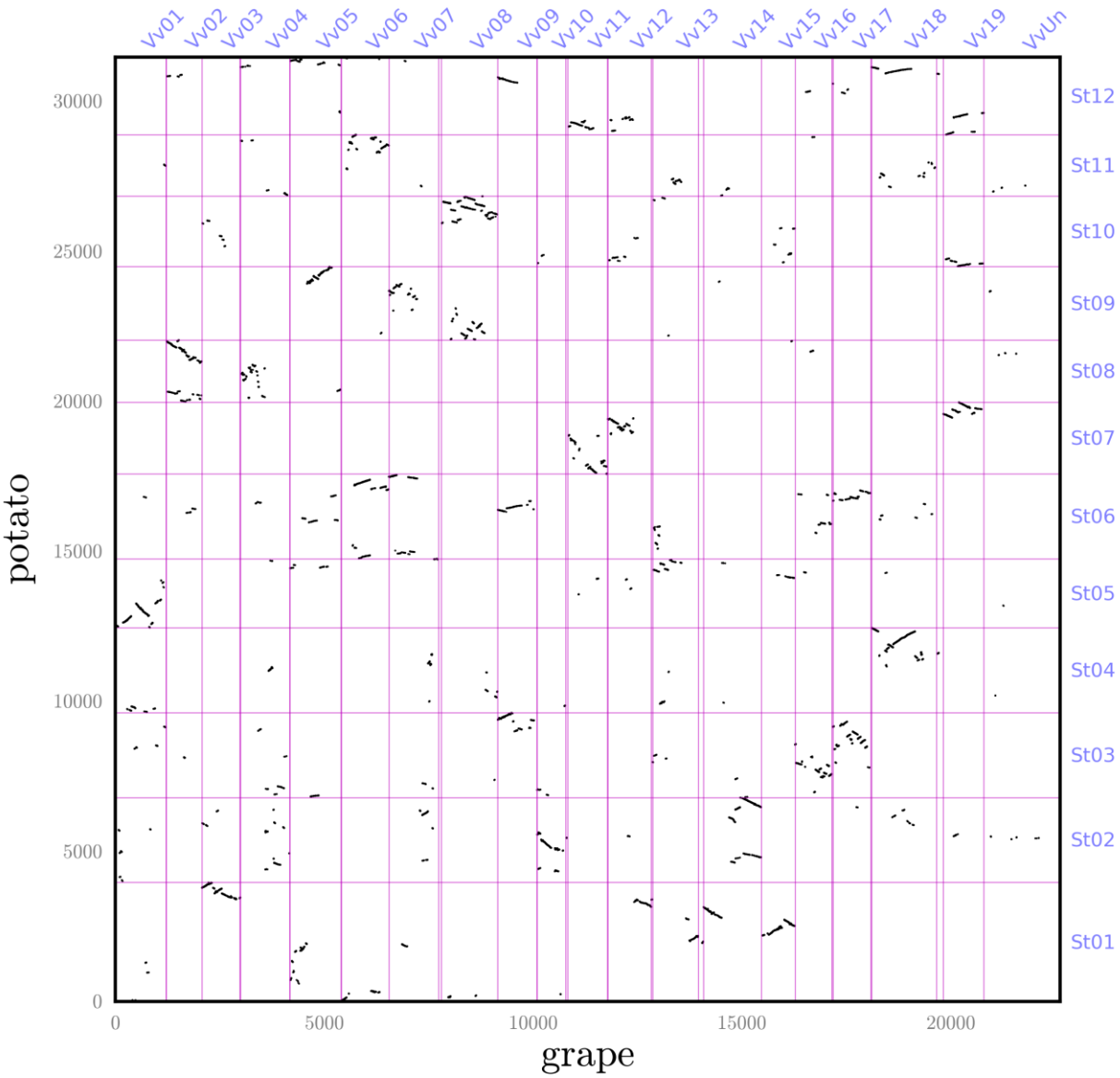


B

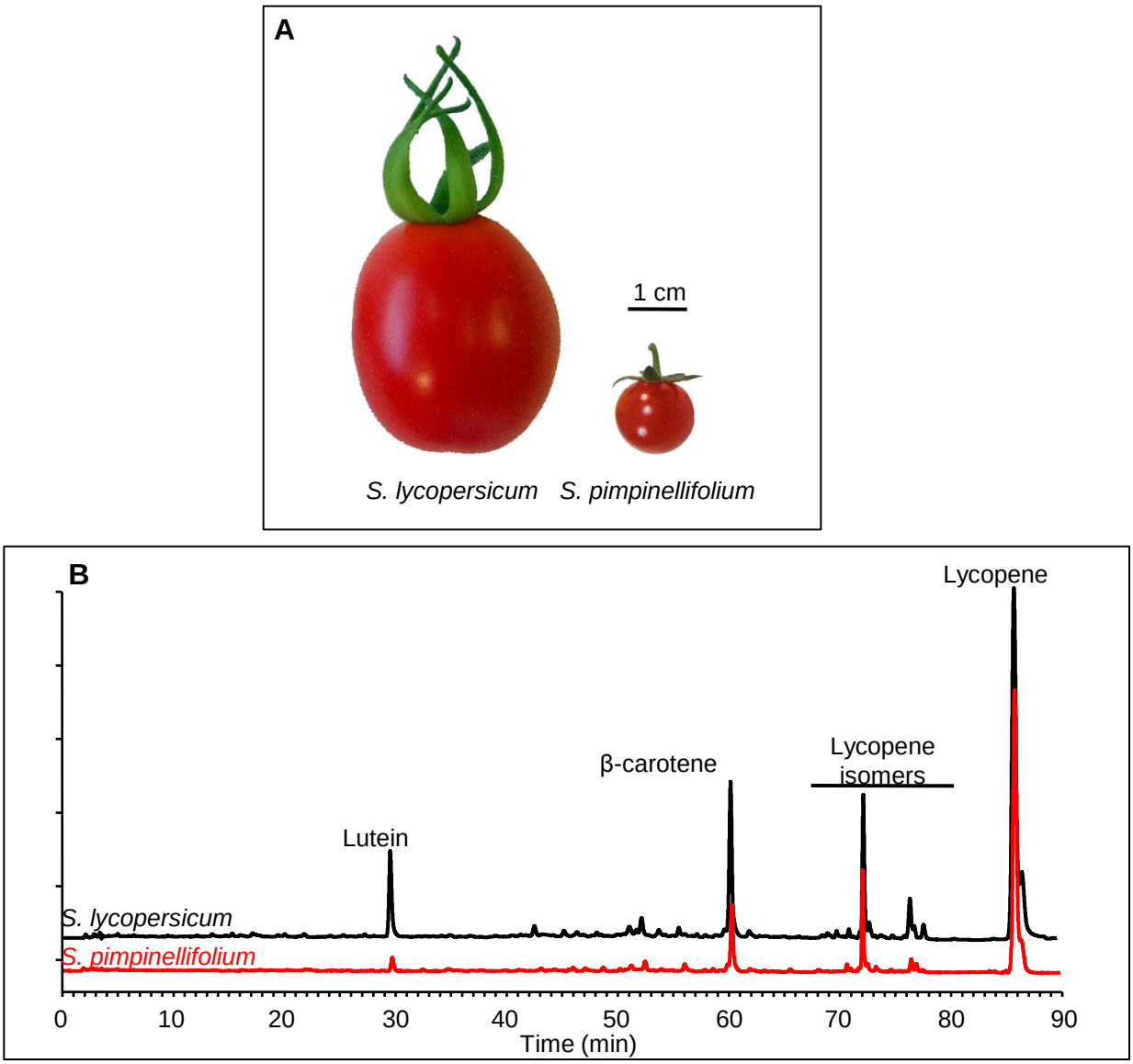


Supplementary Figure 10. (Continued on next page)

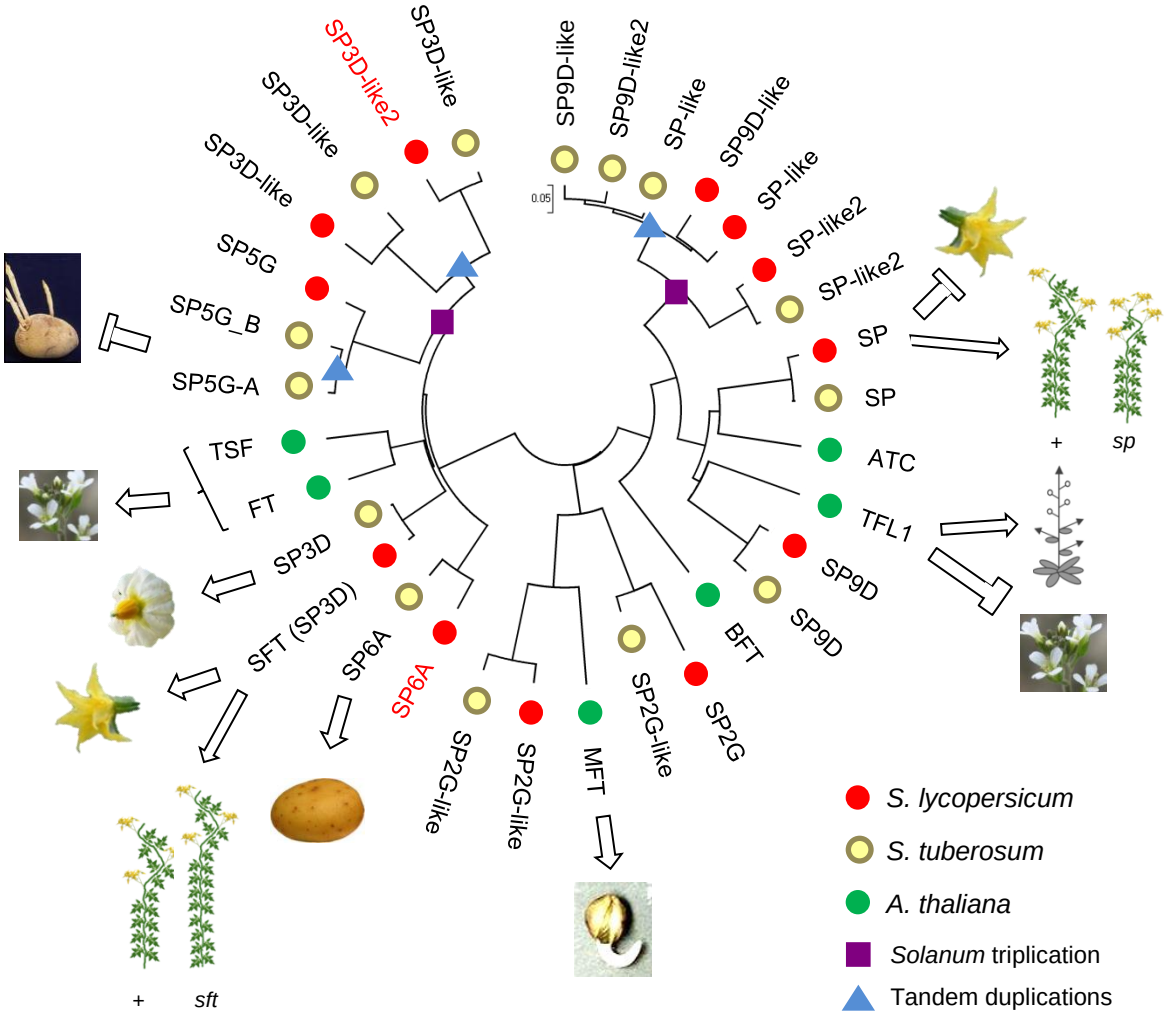
Supplementary Figure 10. The *Solanum* genome triplication is superimposed on the pan-eudicot triplication.
A. Two sequential triplications affected the tomato genome. Grape triplet chromosomes have 1-3 (average 1.99) best-matching tomato chromosomal segments (shown by dots in solid lines), produced by whole-genome triplication in the Solanum lineage, and several other secondary hits (dashed boxes). The color schemes follow the 7 putative eudicot ancestral chromosomes (above and to the left), and the tomato chromosomes (right). The dashed lines below and to the right show DNA duplication depths in grape and tomato, respectively.
B. Distributions of synonymous nucleotide substitution (Ks) rates. MCScan was used to identify syntenic blocks in the tomato genome assembly, calculating Ks distributions between syntenic gene pairs in tomato (red) and tomato-grape (magenta). Lacking genome sequences, for lettuce, sunflower and monkey flower (green, yellow, and blue) unigenes were clustered and assembled from ESTs by TGICL, and Ks values were calculated for pairs of unigenes that were each best matches for a pair of syntenic tomato genes.



Supplementary Figure 11. Dot plot comparing potato and grape genomes shows that syntenic regions with redundancy level of 1 in grape (therefore not derived from the gamma triplication) have redundancy levels of 3 in the potato genome.

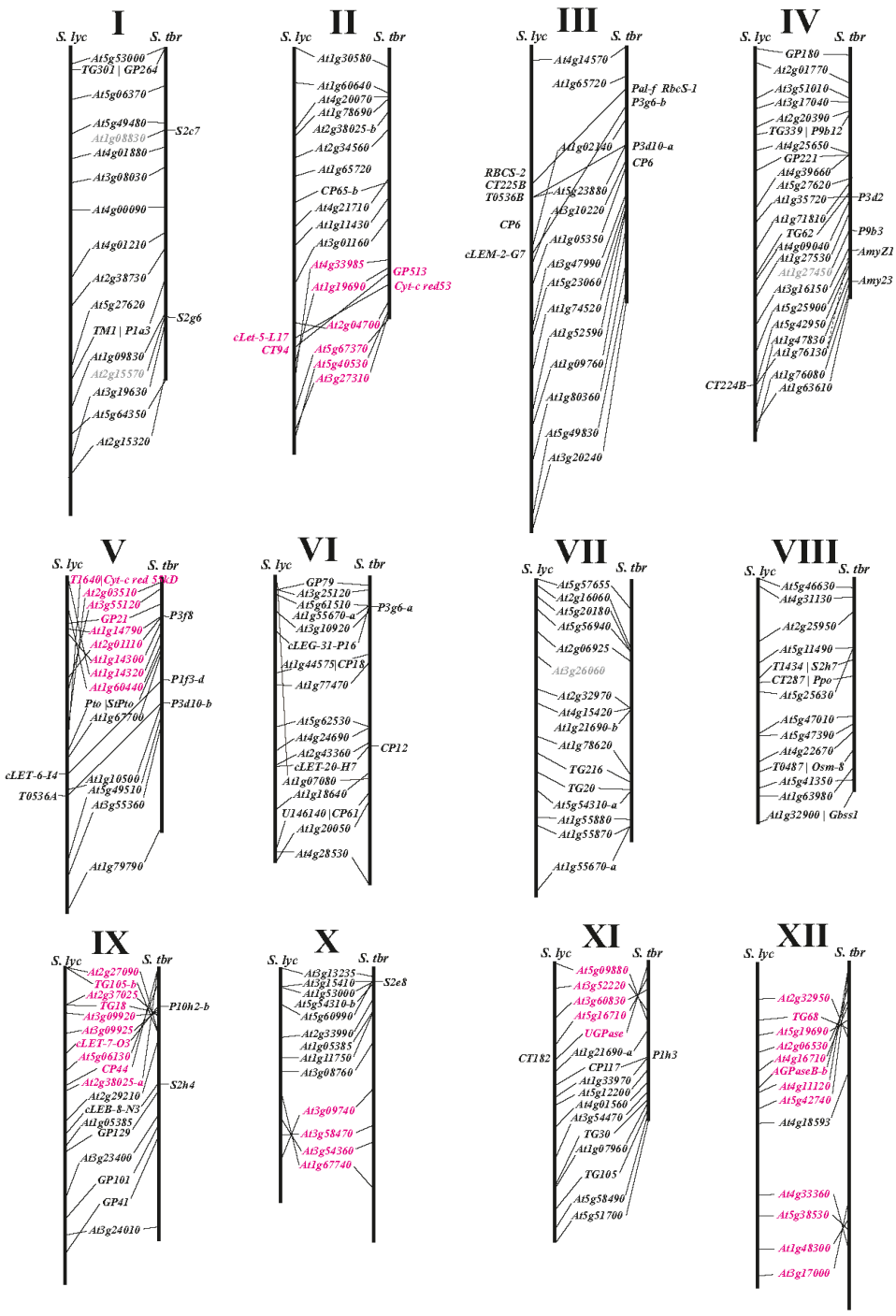


Supplementary Figure 12.
A. Ripe (58 days post anthesis) fruits of *S. lycopersicum* and *S. pimpinellifolium*
B. HPLC chromatograms of carotenoids from fruits shown in A

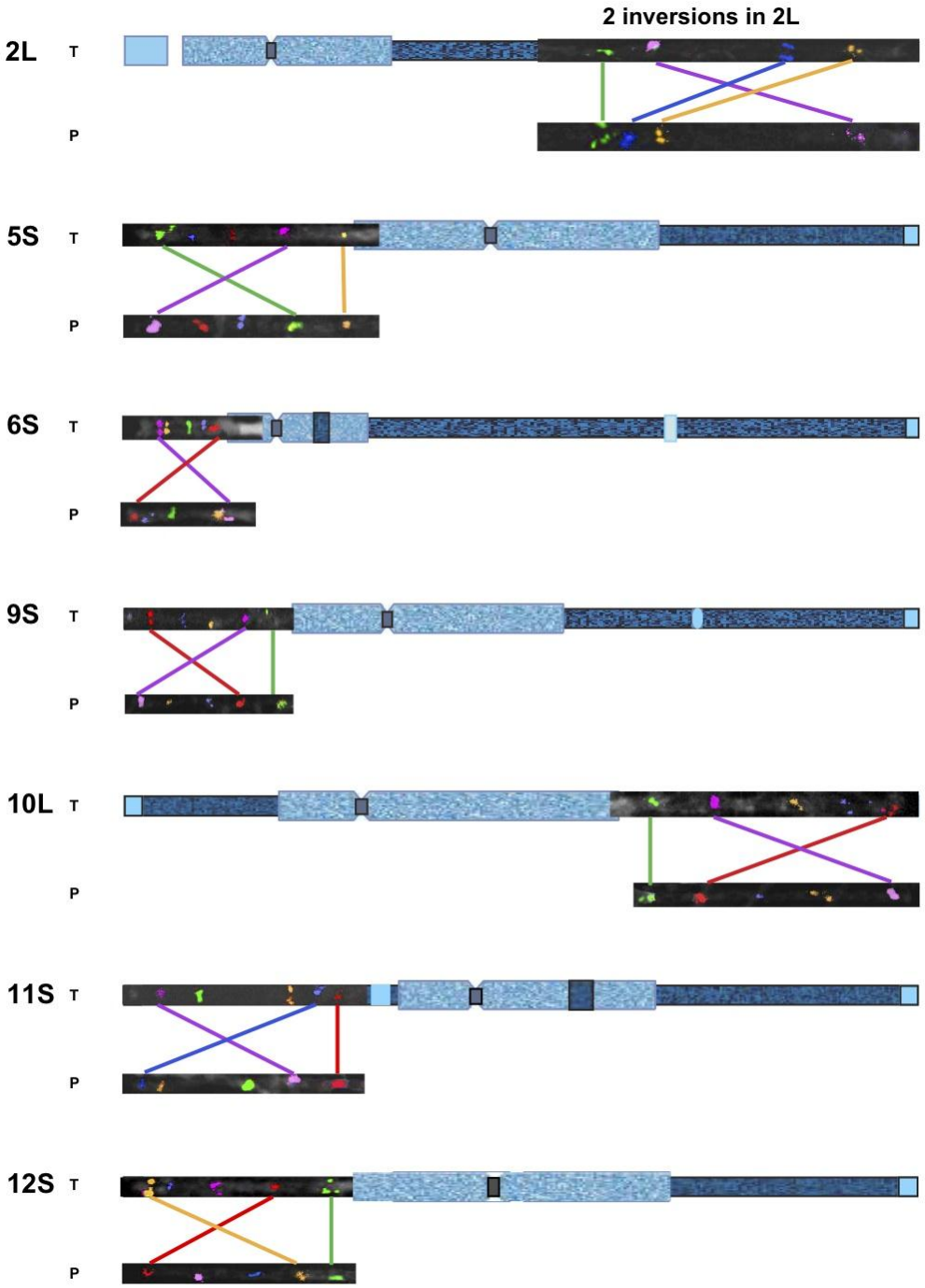


Supplementary Figure 13. Unrooted dendrogram of *Arabidopsis FT* family orthologs in tomato and potato. Some of them have been have been assigned specialized functions. In general, they are involved in flowering-related processes and plant architecture. *SP6A* is a pseudogene in *S. lycopersicum.* Functions have been depicted according to[93,225,226,227,228,229,230,231].

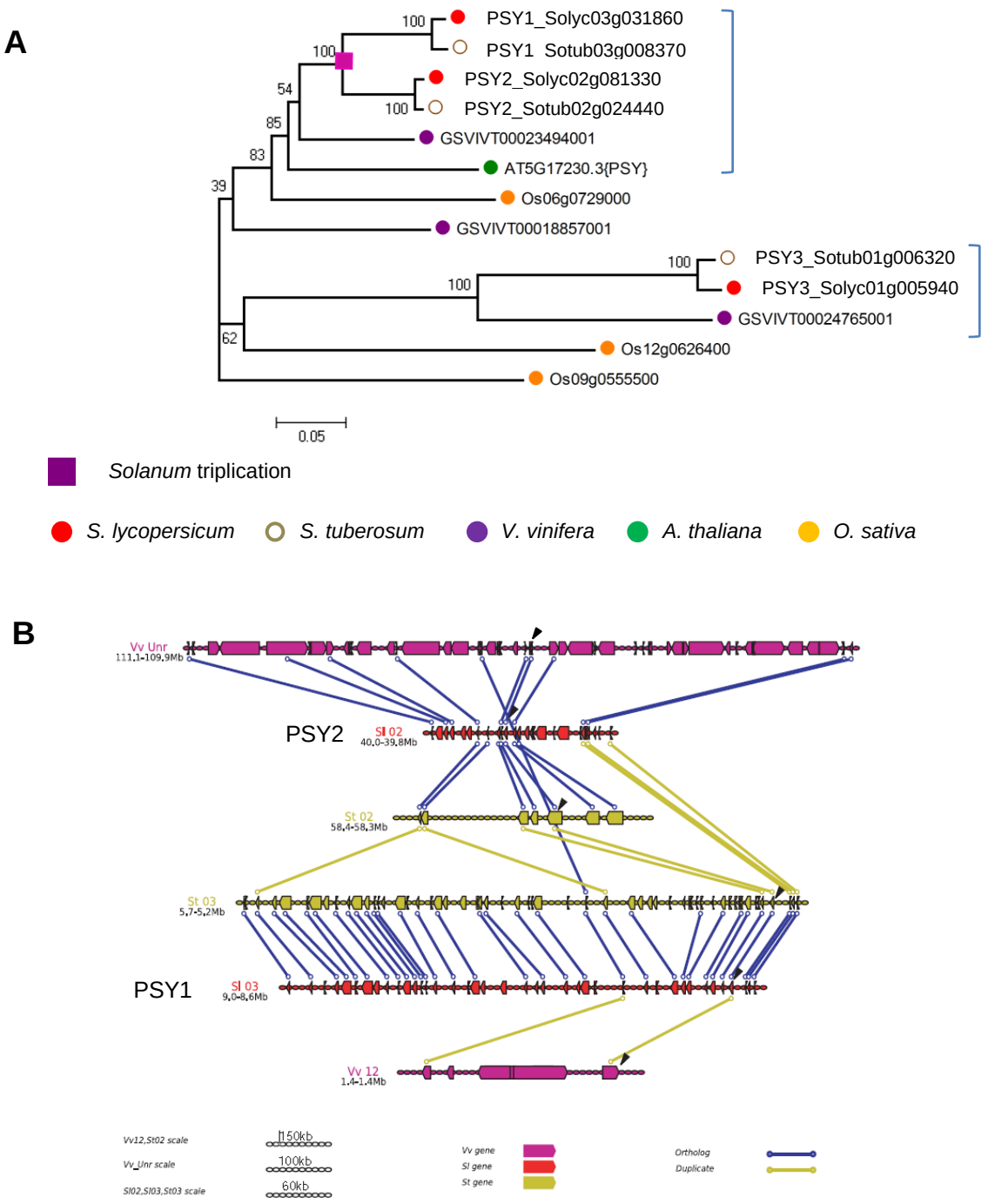
FT/TFL1 is a phosphatidylethanolamine-binding protein family involved in the control of the morphological switch between shoot growth and flower structures with the exception of MFT that regulates seed germination[228]. In *Arabidopsis* it comprises 6 members, but in Solanaceae the family underwent an expansion caused in part by the *Solanum* triplication. Tomato SFT has a function similar to FT, that is, flowering induction[226,230], and its antagonist SP has a function similar to its Arabidopsis ortholog TFL[226,231]. In *sft* mutants, sympodial branching is suppressed[230], whereas *sp* mutants show a determinate phenotype[231]. Solanaceae show expansions in these 2 subclusters, that further acquired species-specific functions: in potato, SP6A promotes tuberisation[225], whereas SP3D appears to have retained its function as flowering inducer[225]. In tomato, but not *S. pimpinellifolium*, *SP6A* is a pseudogene. Other family members have been identified that derive from the *Solanum* triplication (SP3D-like, SP/SP9D-like), but their function is still unknown.



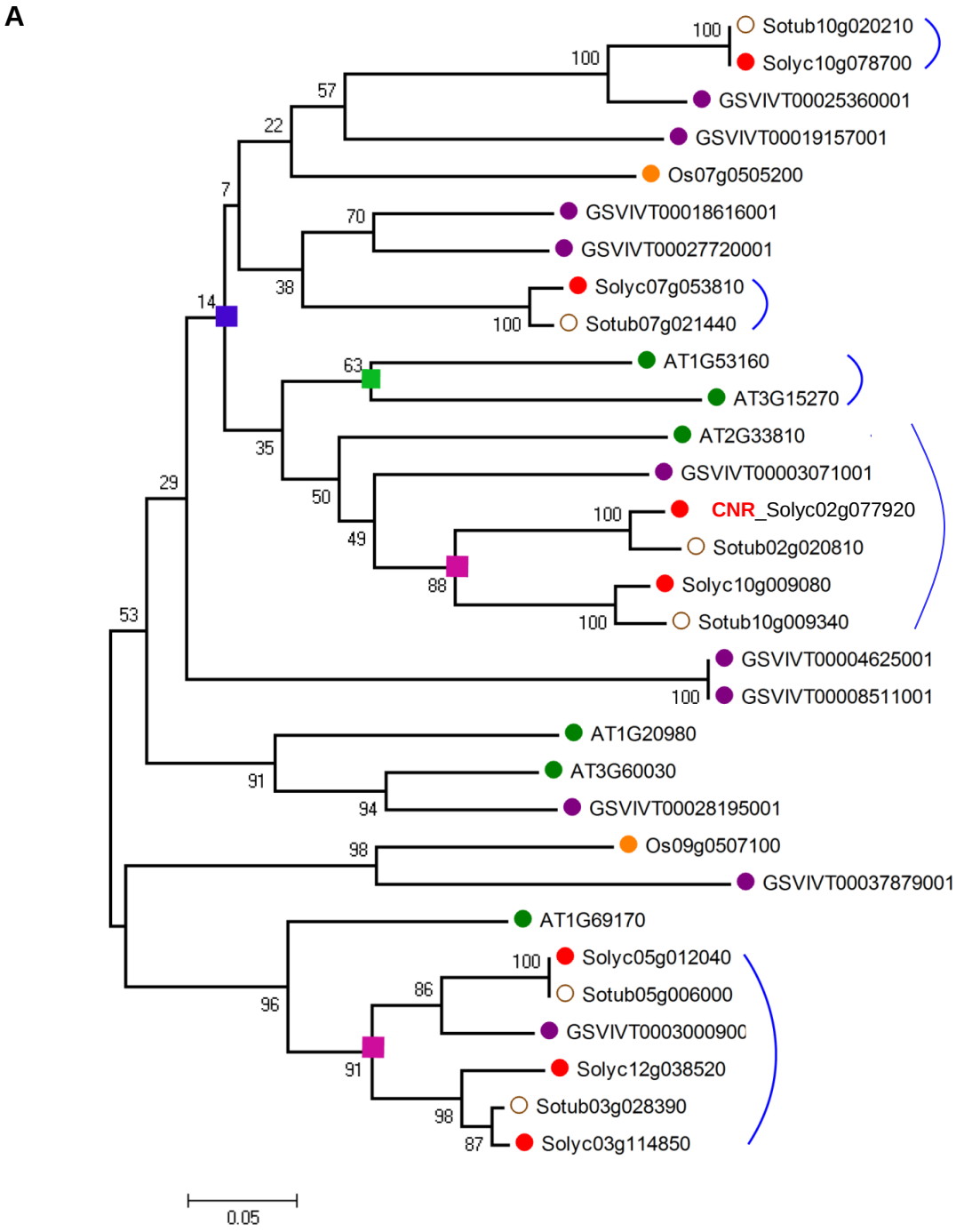
Supplementary Figure 14. Comparative genetic map of potato and tomato based on 141 CosII markers. The diploid potato population F1840 was used for mapping, where detailed linkage maps have been constructed for the 12 potato chromosomes based on RFLP (Restriction Fragment Length Polymorphism) markers and compared with the *Arabidopsis thaliana* genome. DNA polymorphisms were detected by single strand conformation analysis and mapped using the software package MAPRF (E. Ritter, NEIKER, Ap.do 46, E-01080 Vitoria, Spain). Tomato linkage groups (EXPEN2000 map) are shown on the left and potato linkage groups are shown on the right. The CosII markers (At*g*****) are named with the corresponding gene identifier of *Arabidopsis thaliana*. Markers defining inverted chromosome segments are highlighted red.



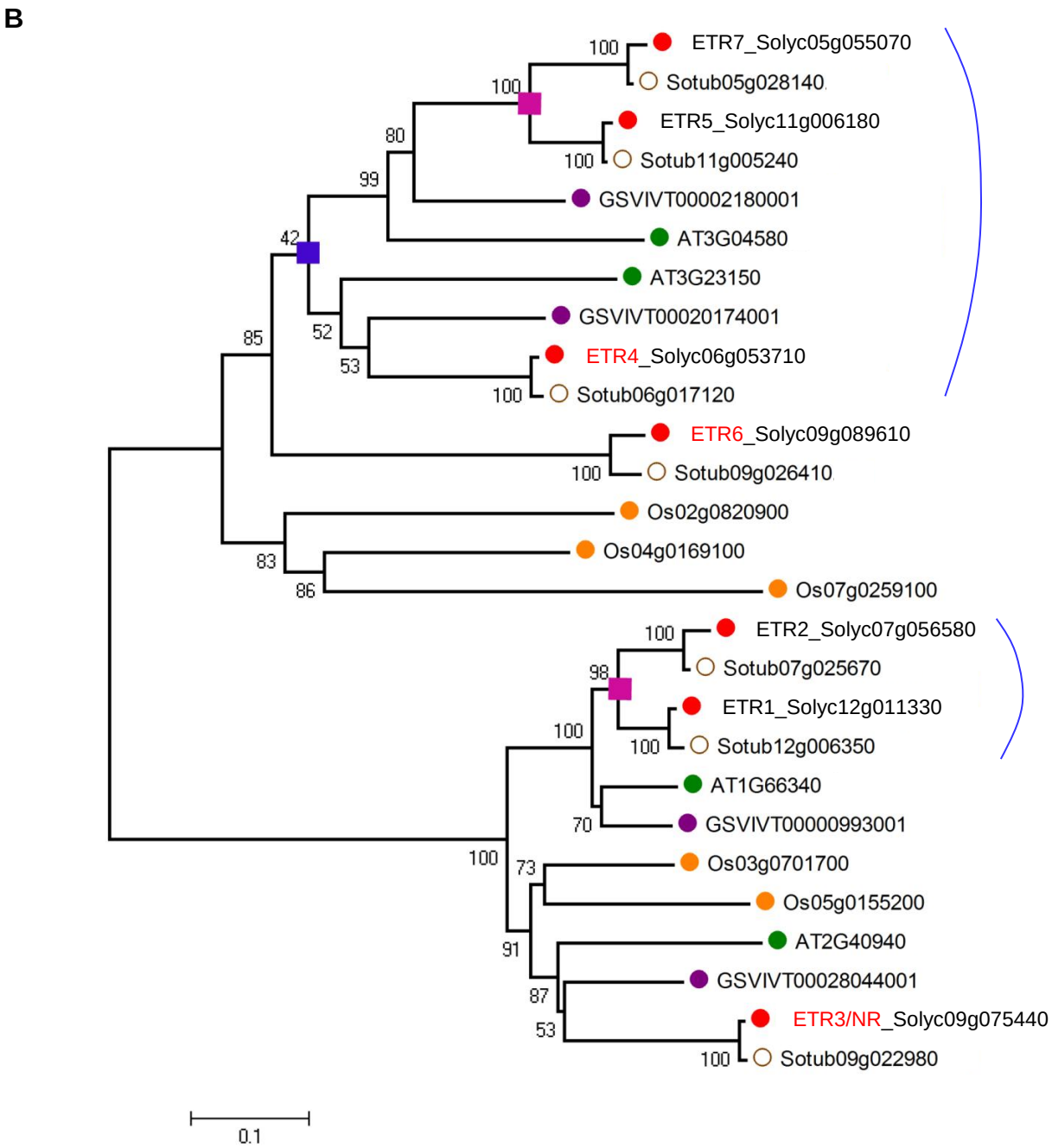
Supplementary Figure 15. FISH images of the paracentric inversions between tomato (T) and potato (P). Only chromosomes are drawn with observed inversions between the two species. Four or five BACs of tomato and/or potato was selected per each chromosome arms to show the inversions between the two species. The BACs were labeled with five different fluorophores (see in the Material and Methods). 2L=chromosome 2, long arm; 5S=chromosome 5, short arm; 6S=chromosome 6, short arm; 9S=chromosome 9, short arm; 10L=chromosome 10, long arm; 11S=chromosome 11, short arm; 12S=chromosome 12, short arm. Coloured lines show the rearrangements between the inverted regions.



Supplementary Figure 16. The *Solanum* triplication contributed to neofunctionalisation of the *PSY* gene family. A. Tomato (red), potato (white), grape (purple), *Arabidopsis* (green) and rice (orange) gene phylogeny is shown in the tree with corresponding genome ID nomenclature. The purple square indicates the *Solanum* triplication, deduced by combined phylogenetic/synteny analyses. Brackets at the right sides of trees show subgroups supported by gene collinearity.
B. Synteny analysis for *PSY* genes. Genomic fragments of tomato, potato and *Vitis* were aligned to show microsynteny for related genes. Black triangles are used to show the considered genes.

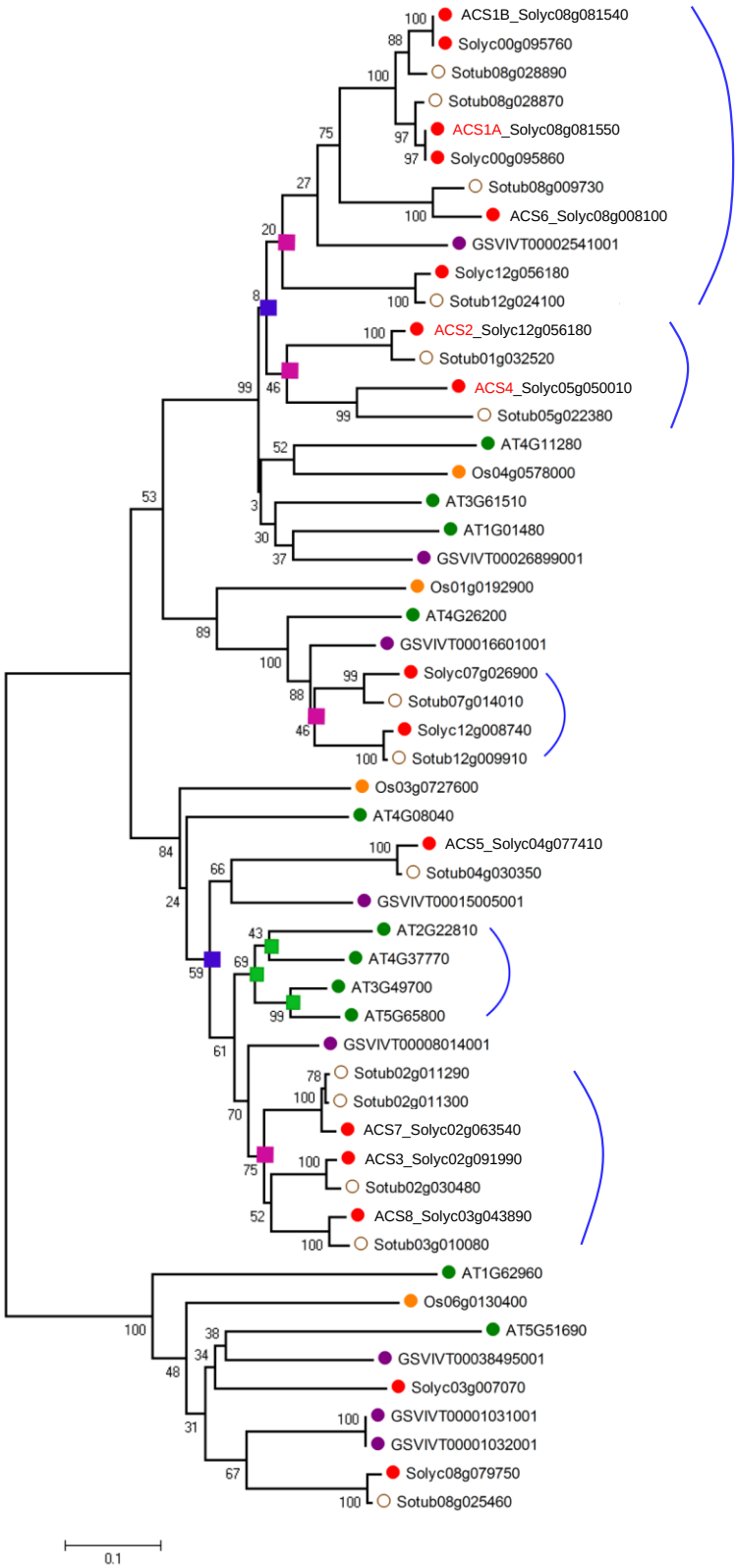


Supplementary Figure 17. (Continued on next page)

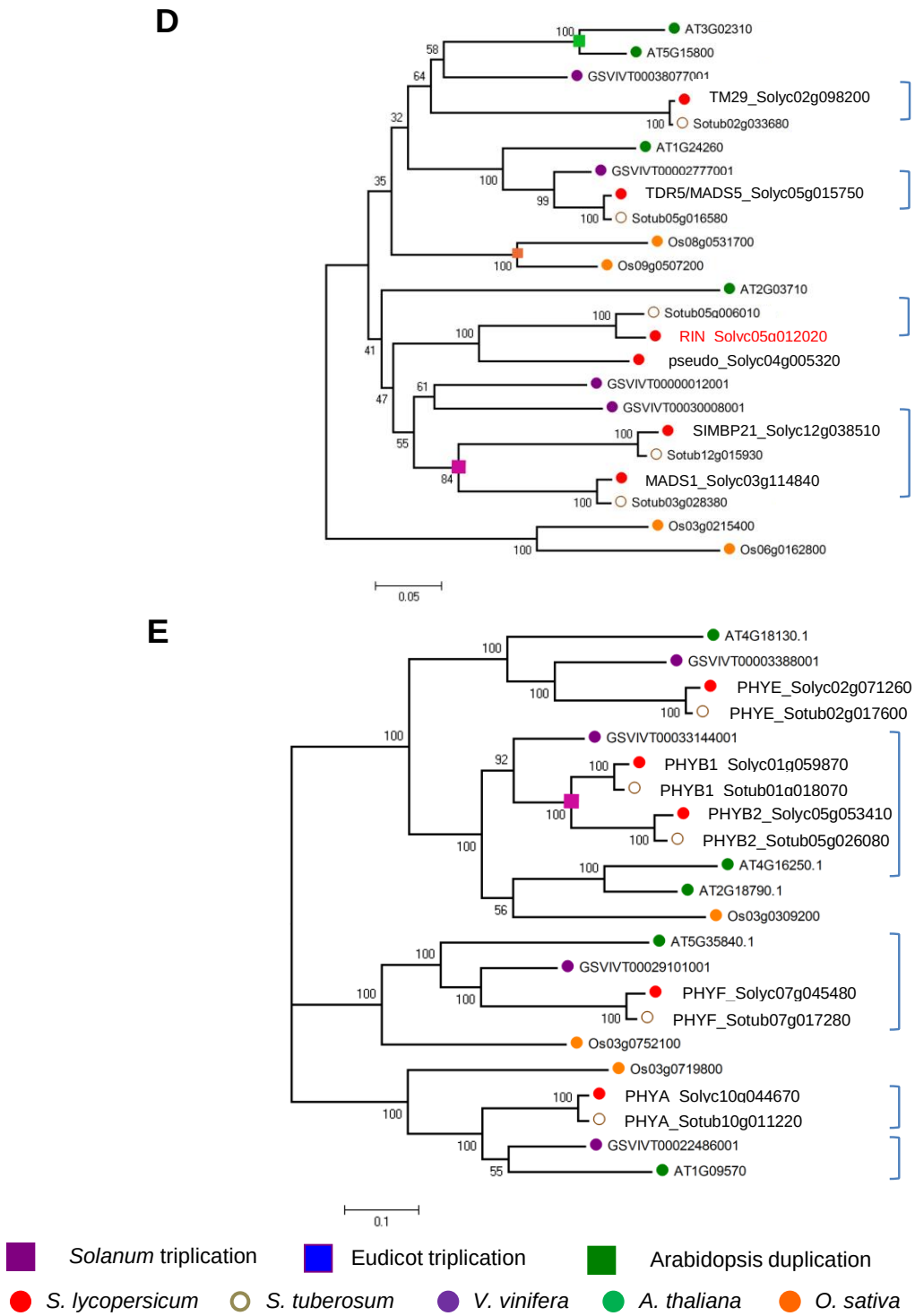


Supplementary Figure 17. (Continued on next page)

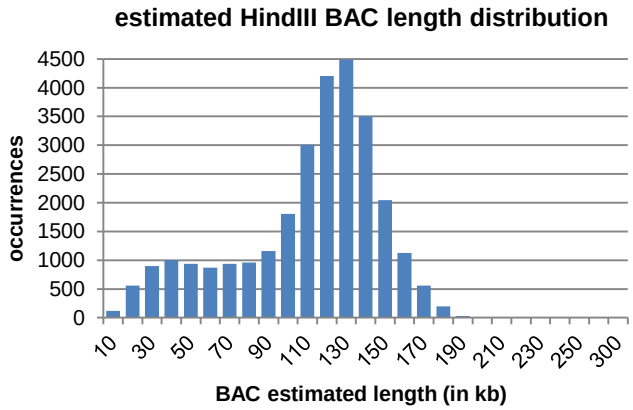
C

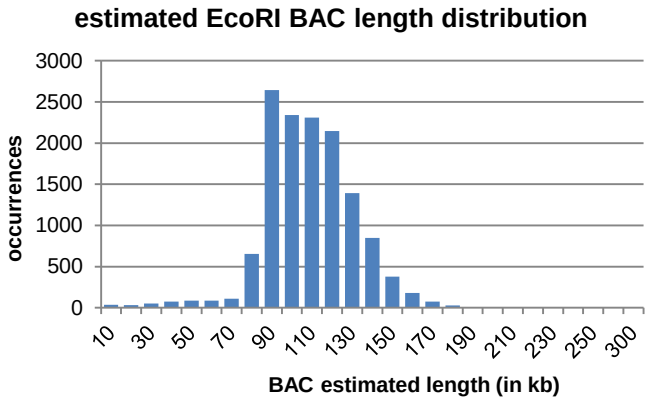


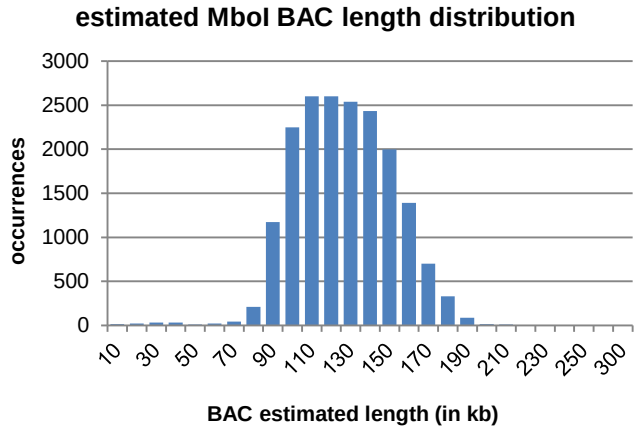
Supplementary Figure 17. (Continued on next page)

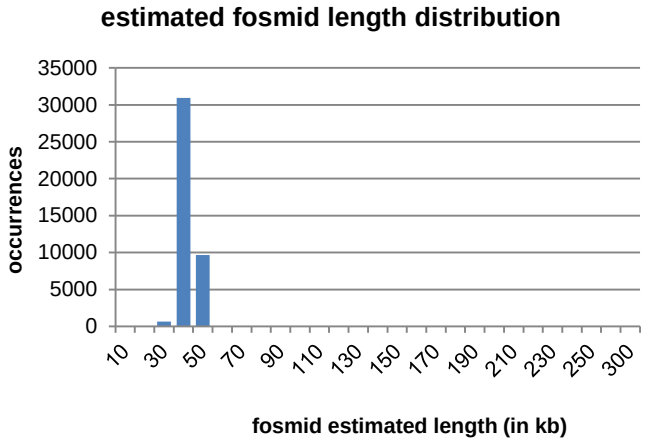


Supplementary Figure 17. Genome polyploidisation events contributed to expansion of the *CNR* (A) *ETR* (B) *ACS* (C) *MADS-RIN* (D) and *PHY* (E) gene familes. Tomato (red), potato (white), grape (purple), *Arabidopsis* (green) and rice (orange) gene phylogeny is shown in the tree with corresponding genome ID nomenclature. Fruit-specific gene names are indicated in red. Purple squares indicate the *Solanum* triplication and blue boxes the eudicot triplication, deduced by combined phylogenetic/synteny analyses. Brackets at the right sides of trees show subgroups supported by gene collinearity.

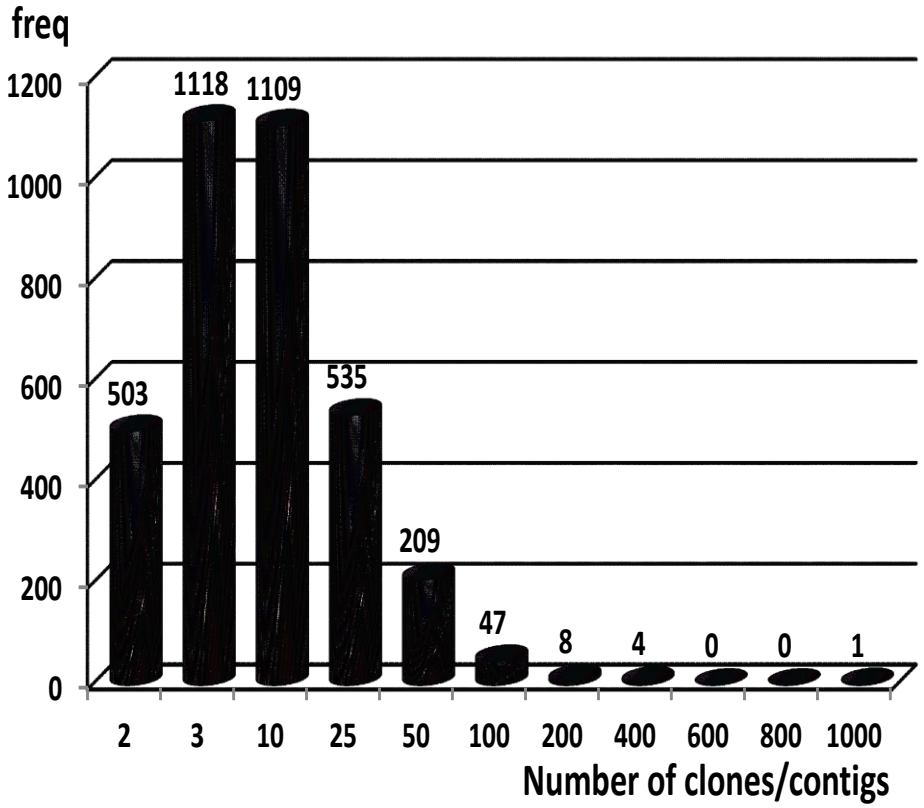




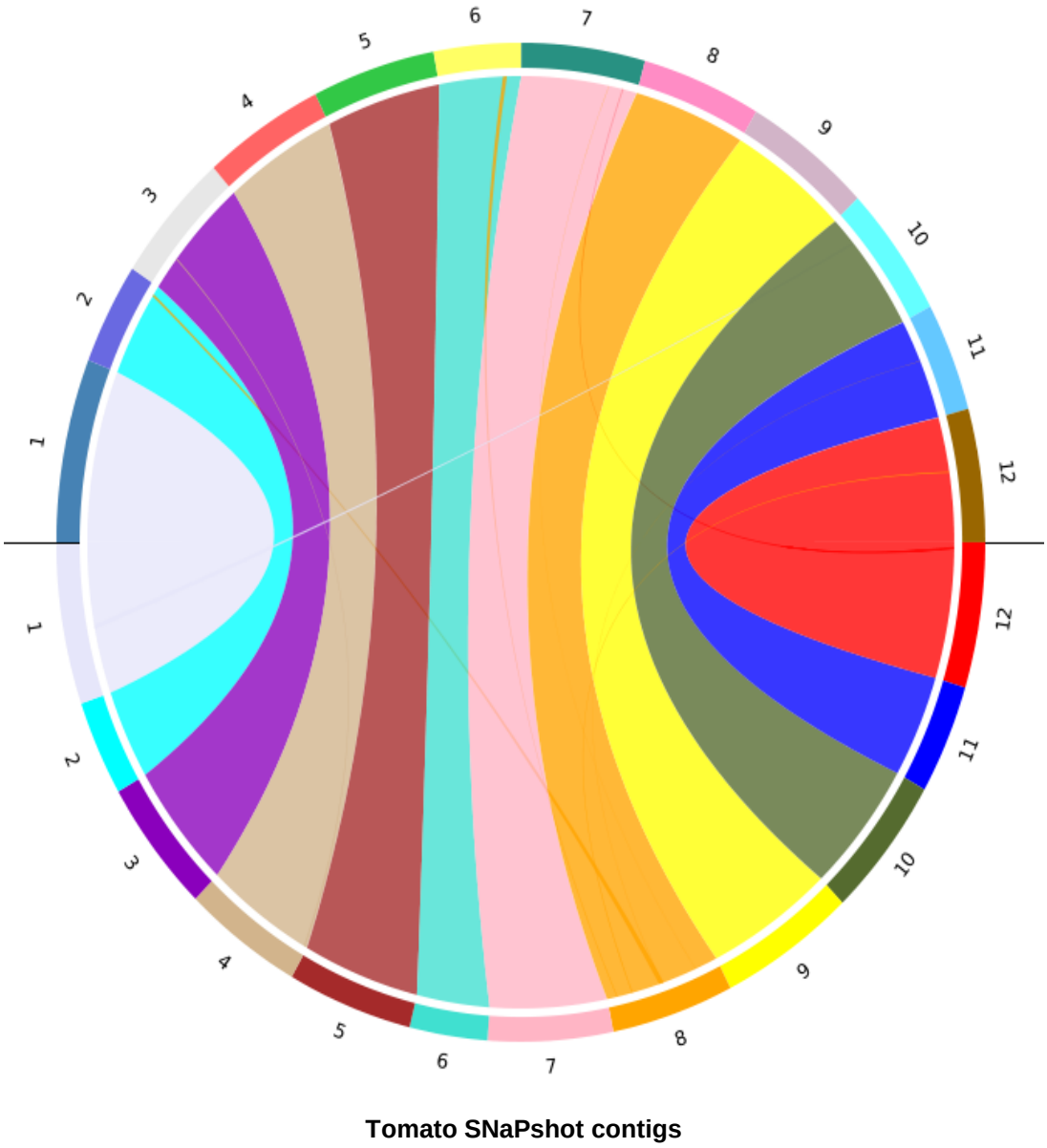




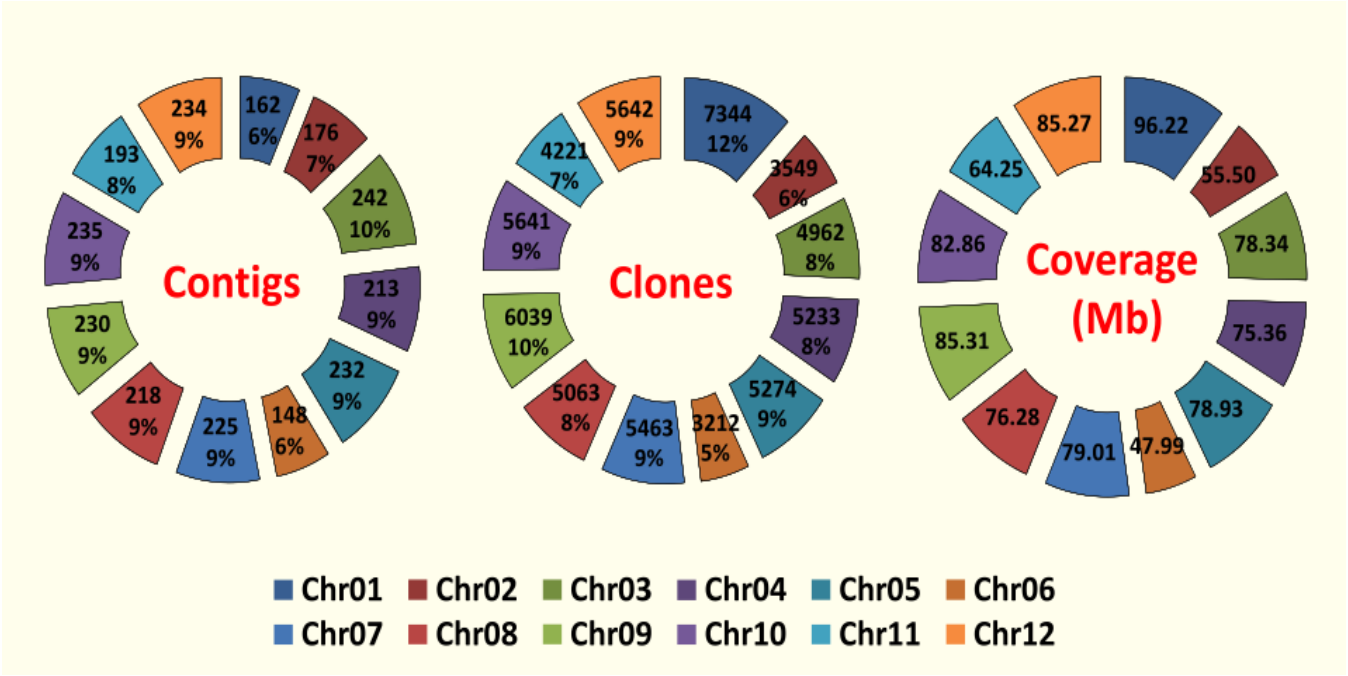
Supplementary Figure 18. Estimated BAC and fosmid length distribution based on mapping of mate-pairs to the *S. lycopersicum* assembly.



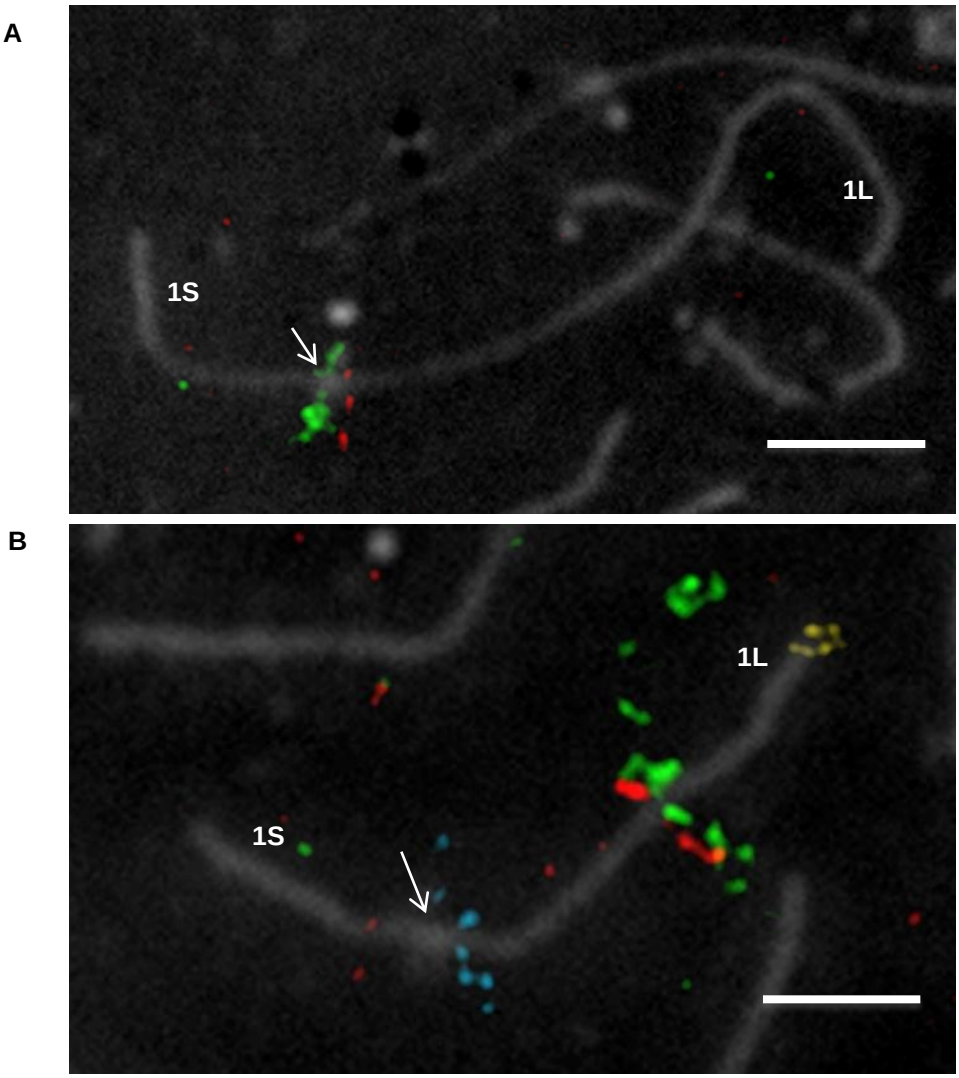
Supplementary Figure 19. Distribution of SNAPshot physical map contigs (ordinates) by clone count (abscissae). Clone count range: 2 clones, and 3-9, 10-24, 25-49, 50-99, 100-199, 200-399, 400-599, 600-799, 800-999 and 1000 or more clones respectively.



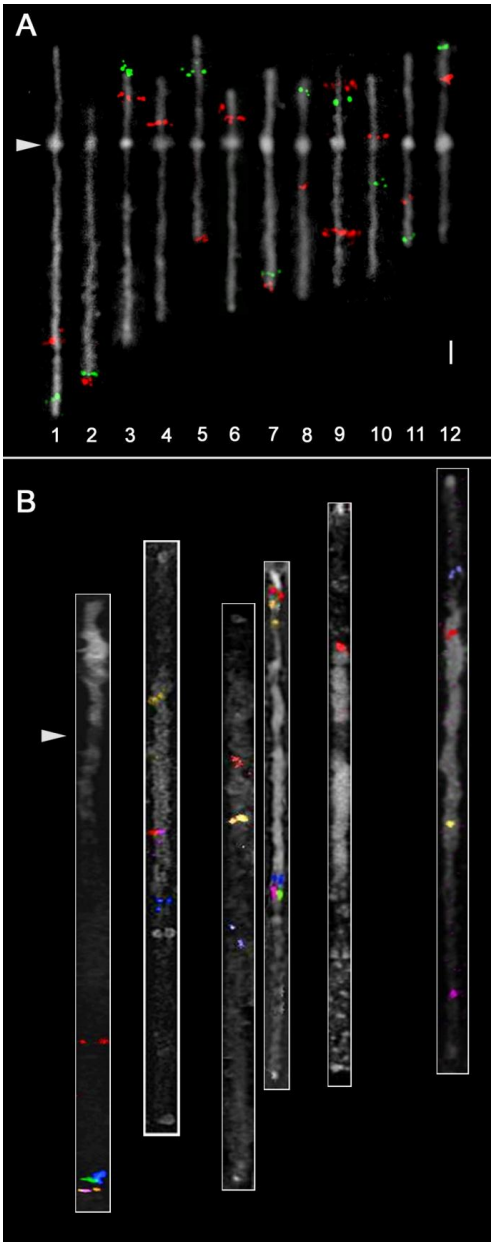
Supplementary Figure 20. Alignment of the anchored SNAPshot physical map FPC contigs to their respective pseudomolecules *via* end sequences. Coloured connectors between top and bottom semi-circle depict BAC end sequence (BES) alignments.



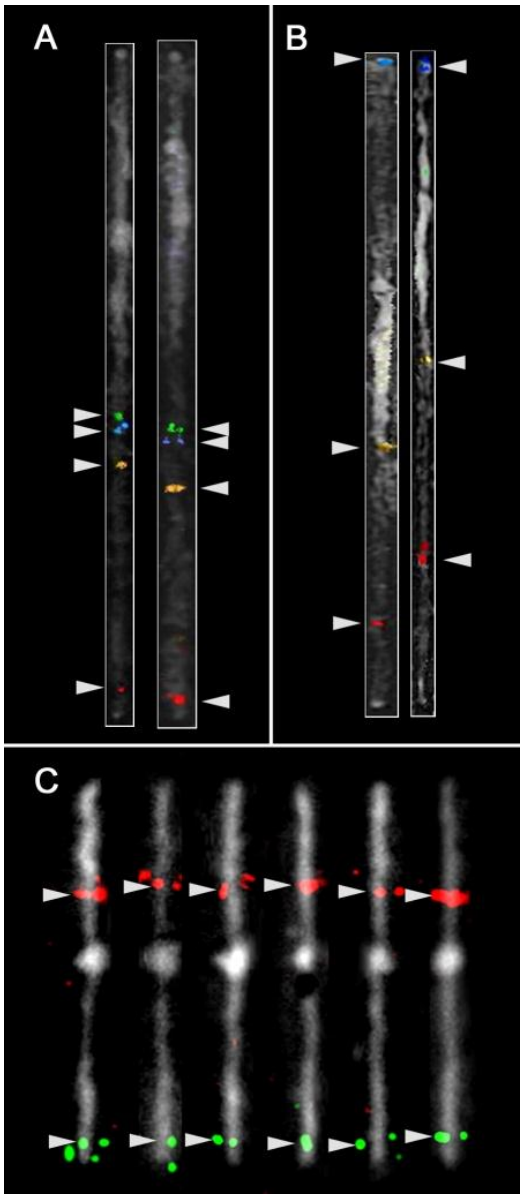
Supplementary Figure 21. Statistics of the alignment of the SNAPshot physical map to the genome sequence pseudomolecules. Contigs (left circle): Represents the total number of FPC contigs aligned to a particular chromosome (absolute number and percentage of the total contigs). Clones (middle circle): Represents the total number of FPC clones aligned to a particular chromosome (absolute number and percentage of the total clones). Coverage (right circle): Represents the estimated coverage (Mb) of the FPC contigs aligned to a particular chromosome (absolute number). Value estimated under the assumption that 1 CB (FPC consensus band) = 1.2 kb.



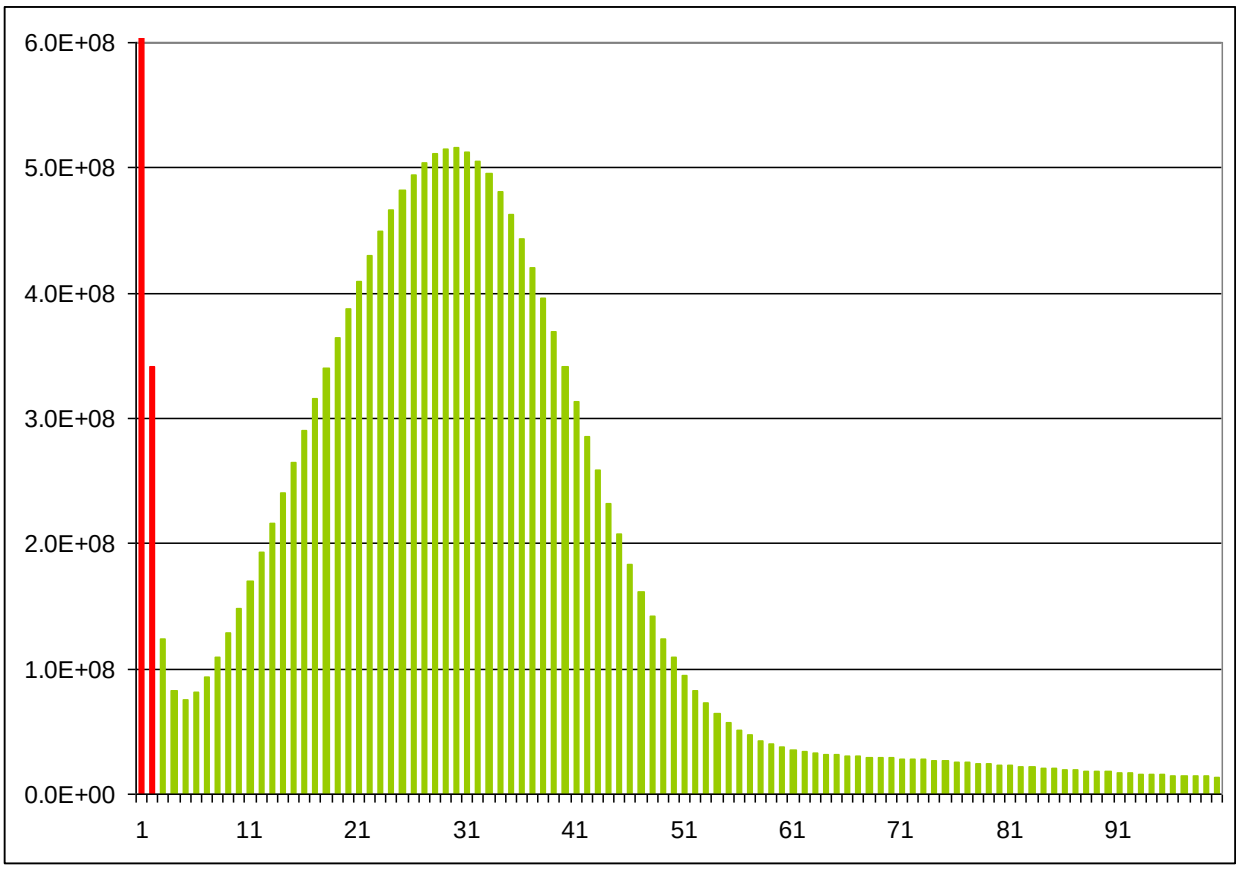
Supplementary Figure 22. Determining scaffold order and orientation and estimating gap sizes in DNA pseudomolecules by using BAC-FISH on spreads of tomato synaptonemal complexes (SCs = pachytene chromosomes). A. A reversed phase contrast image of tomato SC *1* with the short (1S) and long (1L) arms indicated. Chromatin loops extend above and below the SCs. BAC s083L21 (green) is at the bottom border of scaffold 3, and BAC E016I11 (red) is at the top border of scaffold 2. Although the scaffold numbering based on sequence assembly is out of order, the gap defined by these two BACs (gap 3-2 in **Supplementary Table 13**) consists largely (possibly completely) of the centromere as defined by the borders of the kinetochore (arrow - just visible as a brighter white spot between the two fluorescence signals). Gap 3-2 measures 0.7 µm (S.D. 0.2 µm, *n* = 15) and is estimated to be approximately 2.4 Mb in length (based on 3.6 Mb/µm in the centromere of chromosome 10, scaffold 4). B. A reversed phase contrast image of tomato SC *10* with the short (1S) and long (1L) arms indicated. An arrow points to the kinetochore/centromere. Chromatin loops extend above and below the SCs. BAC s083L21 (blue) is at the top border and BACs 121P17 (red) is at the bottom border of scaffold 3 that spans both eu- and heterochromatin and includes 31.1 Mb in 4.7 µm of SC. BAC E036N16 (green) is at the top border and BAC E008A07 (yellow) is at the bottom border of scaffold 5 that is in euchromatin and includes 8.0 Mb in 4.5 µm of SC. Although the scaffold numbering based on sequence assembly is out of order, unsequenced gap 3-5 is in euchromatin with its borders defined by BACs121P17 (red) and E036N16 (green). This gap is 0.2 µm (S.D. 0.1 µm, *n* = 14, **Supplementary Table 13**), and assuming 1.54 Mb/µm in euchromatin, the gap consists of approximately 0.3 Mb. Bar equals 5 µm.



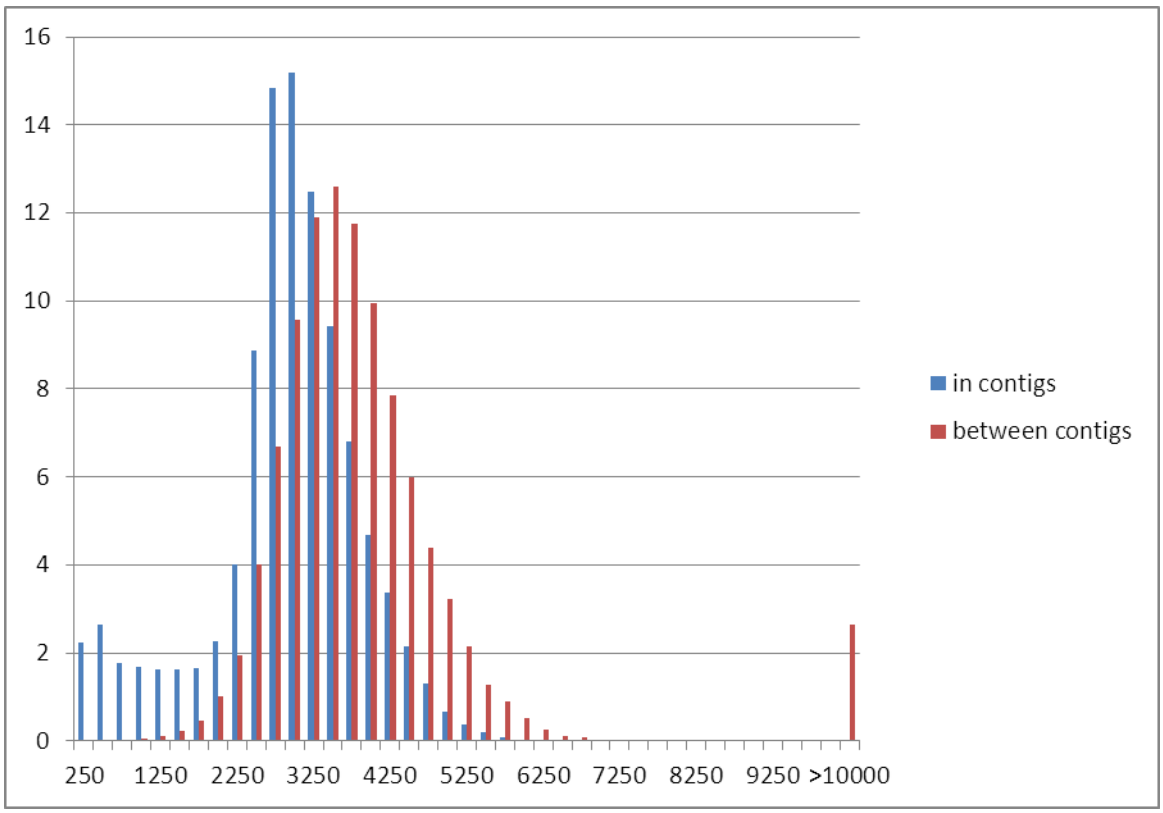
Supplementary Figure 23. Examples of BAC-FISH on (**A**) spread SCs and (**B**) stirred pachytene chromosomes from tomato. Each SC and chromosome was labeled with different BAC probes (**Supplementary Table 78**), straightened, and assembled into this montage. (**A)** The twelve tomato SCs labeled by FISH. The SCs (from inverted, DAPI-stained phase images) are all aligned by their kinetochores (arrowhead). SC in pericentric heterochromatin is more heavily stained than SC in distal euchromatin. The length of each SC has been adjusted to fit the standard tomato SC karyotype. (**B**) Examples of stirred pachytene chromosomes *2, 4, 6, 7, 9* and *12* (under corresponding SCs in A above) labeled with different probes and aligned by their centromeres (arrowhead). Proximal (pericentric) heterochromatin is more heavily stained than distal euchromatin (the opposite staining pattern from **A** above). The scale bar = 2 μm for **A** and 1 μm for **B**.



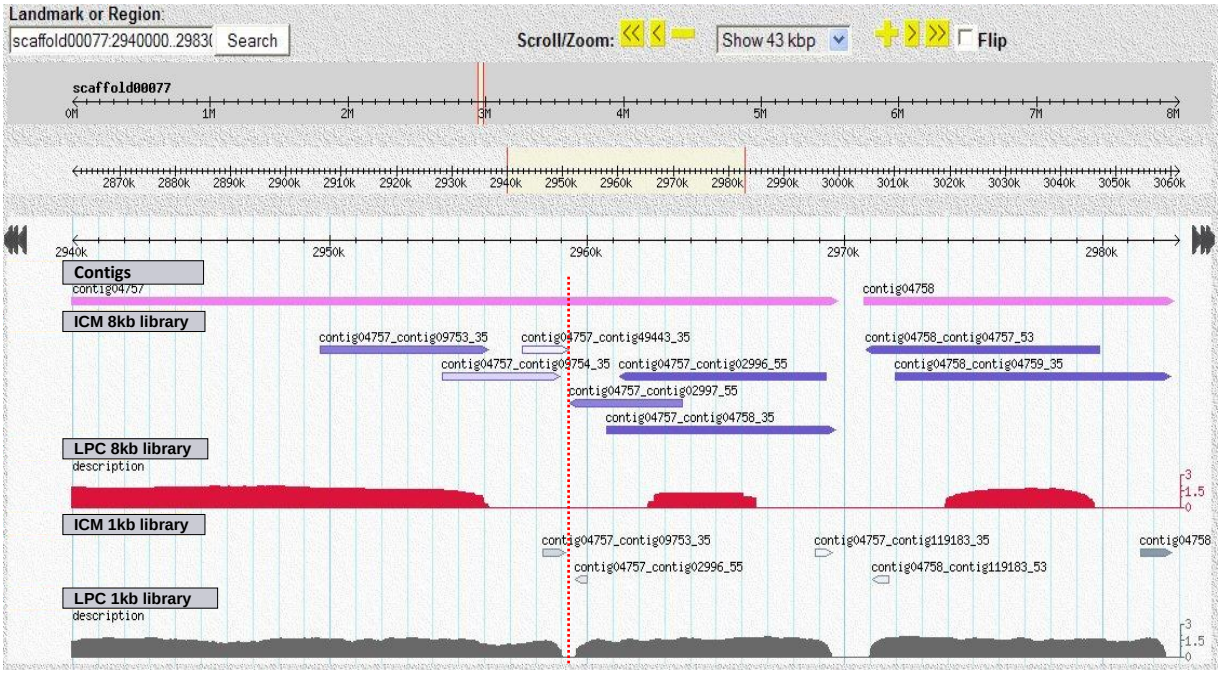
Supplementary Figure 24. Reproducibility of FISH localisations on stirred pachytene chromosomes and on SC spreads. (**A**) Chromosome 6 from two different chromosome sets that had been fixed with 1:3 acetic ethanol and hybridized with BAC probes 215M16 (green), 024F02 (blue), 177K13 (yellow), 034C13 (red). The two chromosomes have been adjusted to be the same length using Adobe Photoshop. (**B**) Pachytene chromosome 7 from two different chromosome sets that had been fixed with 1:3 acetic ethanol and hybridized with the BAC probes E110K10 (blue), 044H04 (yellow), 059A10 (red). The two chromosomes have been adjusted to be the same length using Adobe Photoshop. In both (**A**) and (**B**), the pericentric and telomeric heterochromatin segments are stained more intensely with DAPI (brighter white in these images) than the distal euchromatin. The order of BACs on both examples of chromosome 6 and chromosome 7 is the same. However, differential stretching of the chromosomes may lead to substantial differences in the location of the BACs on the chromosomes (especially in **B**). (**C**) SC 11 from six different sets of SCs that were hybridized with the same BAC probes (291F09 – red, 323E19 - green). The six SCs have been aligned by their dark kinetochores and adjusted to be the same total length using Adobe Photoshop. Some BAC signals extend to either side of the SCs because the spreading process disperses chromatin loops laterally. The position of each BAC as measured from the kinetochore is highly reproducible on different SC sets.



Supplementary Figure 25. 31-mer distribution in the Illumina data of the *S. lycopersicum* genome.
The 31-mer frequency (y-axis) is plotted against the volume of 31-mers in the read data with that frequency (y-axis). The 31-mers with a frequency below three (in red) are assumed to represent the majority of erroneous 31-mers (i.e., those derived from sequencing errors), whereas the 31-mers with a frequency of three and higher (in green) are assumed to contain few erroneous 31-mers. The peak of the Poisson-shaped distribution (which represents the volume of unique 31-mers in the tomato genome) is at a frequency of 30, meaning that the 31-mers cover of the genome roughly 30-fold. The x-axis and y-axis in the figure have been limited to 100 and $6.0 * 10^8$, respectively.

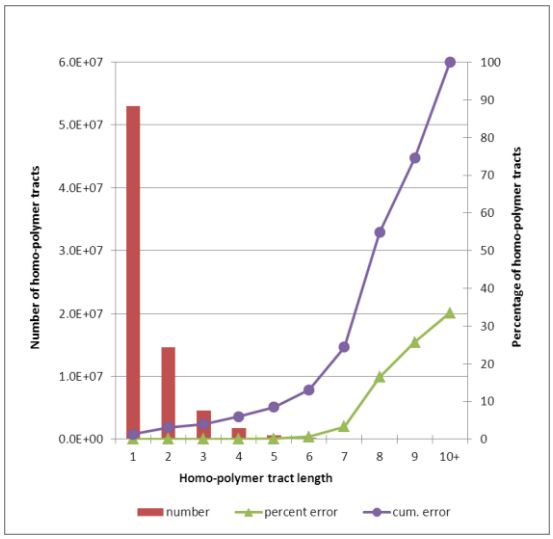


Supplementary Figure 26. Distance between SBM matepair sequences in contigs (blue) and between contigs (red) of the Newbler *de novo* assembly of the *S. lycopersicum* genome.
The x-axis represents the distance between the two sequences of a mate pair, measuring from the outermost end of the sequences, in bins of 250 bp; the y-axis shows the fraction of matepairs with the corresponding distance.

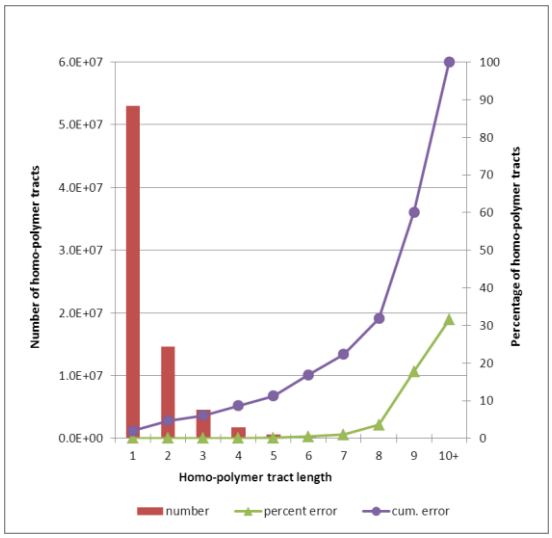


Supplementary Figure 27. GBrowse visualisation of a chimeric contig (contig04757), as revealed by LPC and ICM evidence.
The dotted line highlights the misassembled region where the local physical coverage (shown in log10 scale) of both SOLiD mate-pair libraries drops to zero. The arrows shown in the ICM sections allow to identify which contigs are connected to the left and right part of the misassembled region. The darkness of the arrows increases with the number of mate-pairs linking the two contigs. The two digits reported at the end of each arrow name indicate the ends of the contigs (3' or 5') that should be merged: the first digit referring to the misassembled contig and the second digit to the linked contig.

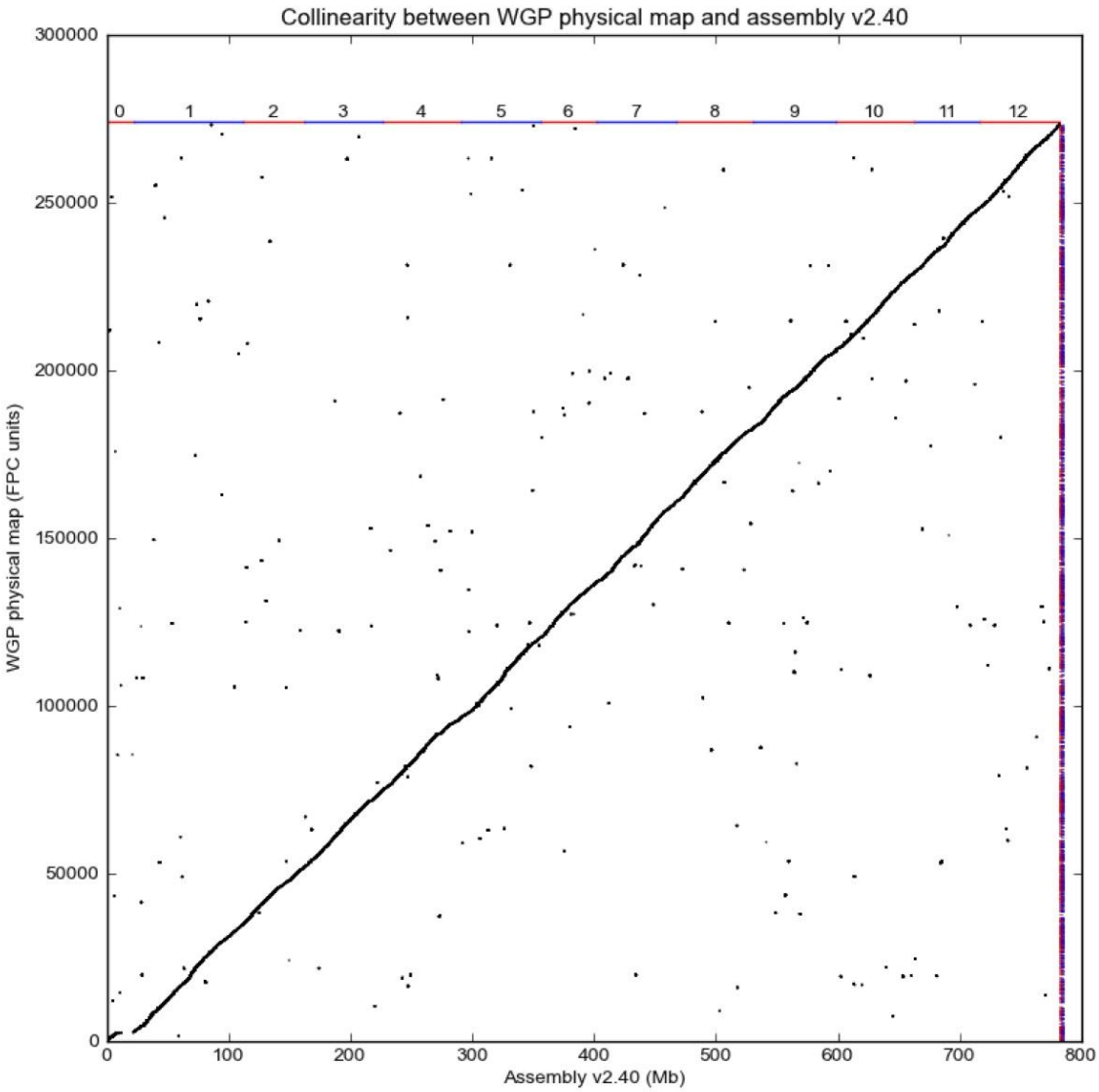
A



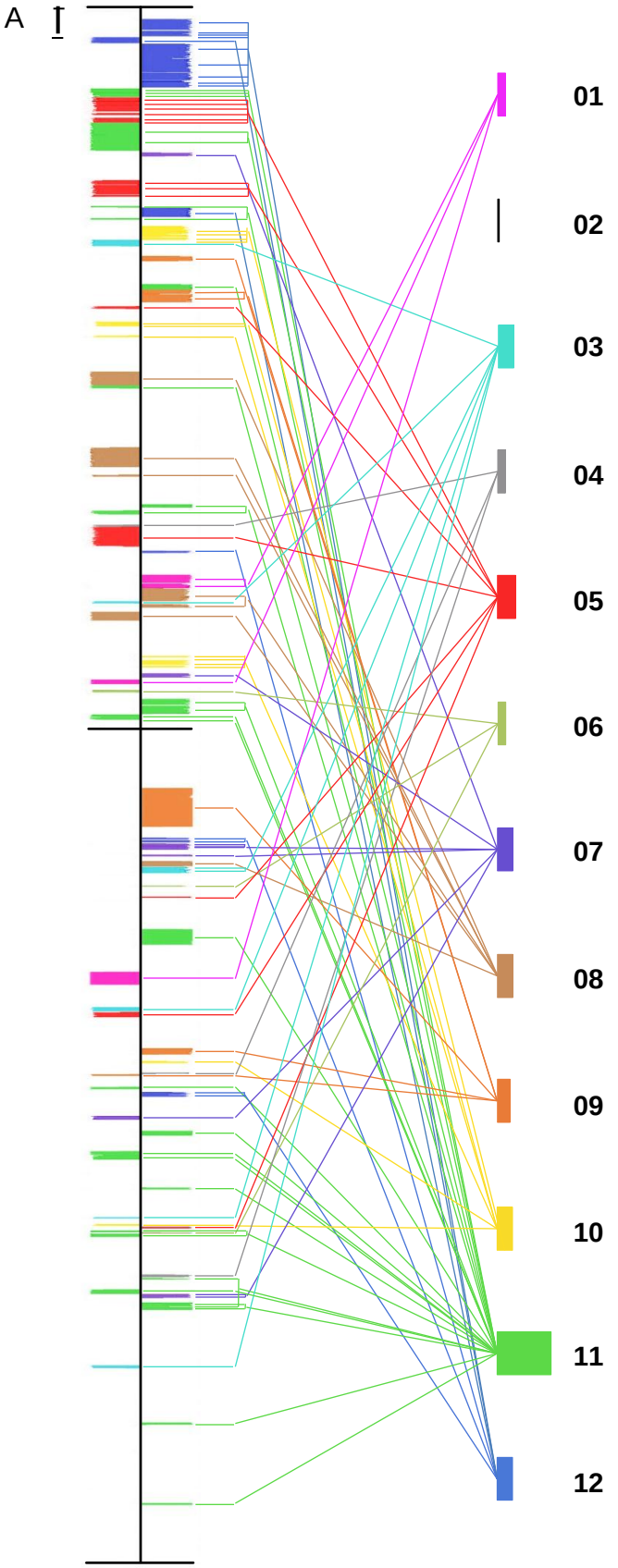
B

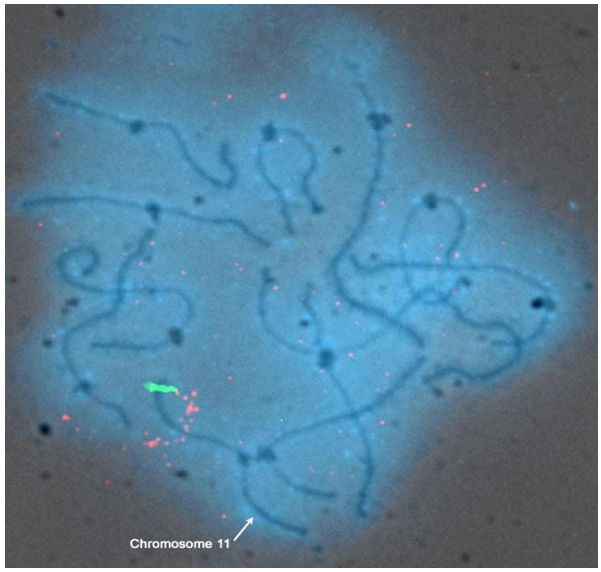


Supplementary Figure 28. Distribution of errors in homopolymer tracts (a) before and (b) after integration of the Illumina and SOLiD data.
Red bars present the number of homopolymer tracts in 108.3 Mb of the genome assembly covered by BAC sequences. The green lines indicate the percentage of homopolymer tracts in the assembly that deviated in length compared to the BAC reference sequence. The purple lines show the cumulative percent error in the homopolymer tracts. Homopolymer tracts of 6 nt and shorter represent the vast majority of the sequence (99.9%), have a low error rate (below 0.6% before correction, and below 0.5% after correction) and have a limited contribution to the error rate (13.1% of all homopolymer errors before correction, and 16.9% after correction). Through error correction, the error rate of homopolymers with length 7 and 8 dropped from 3.3% and 16.4% to 1.0% and 3.5%, respectively.

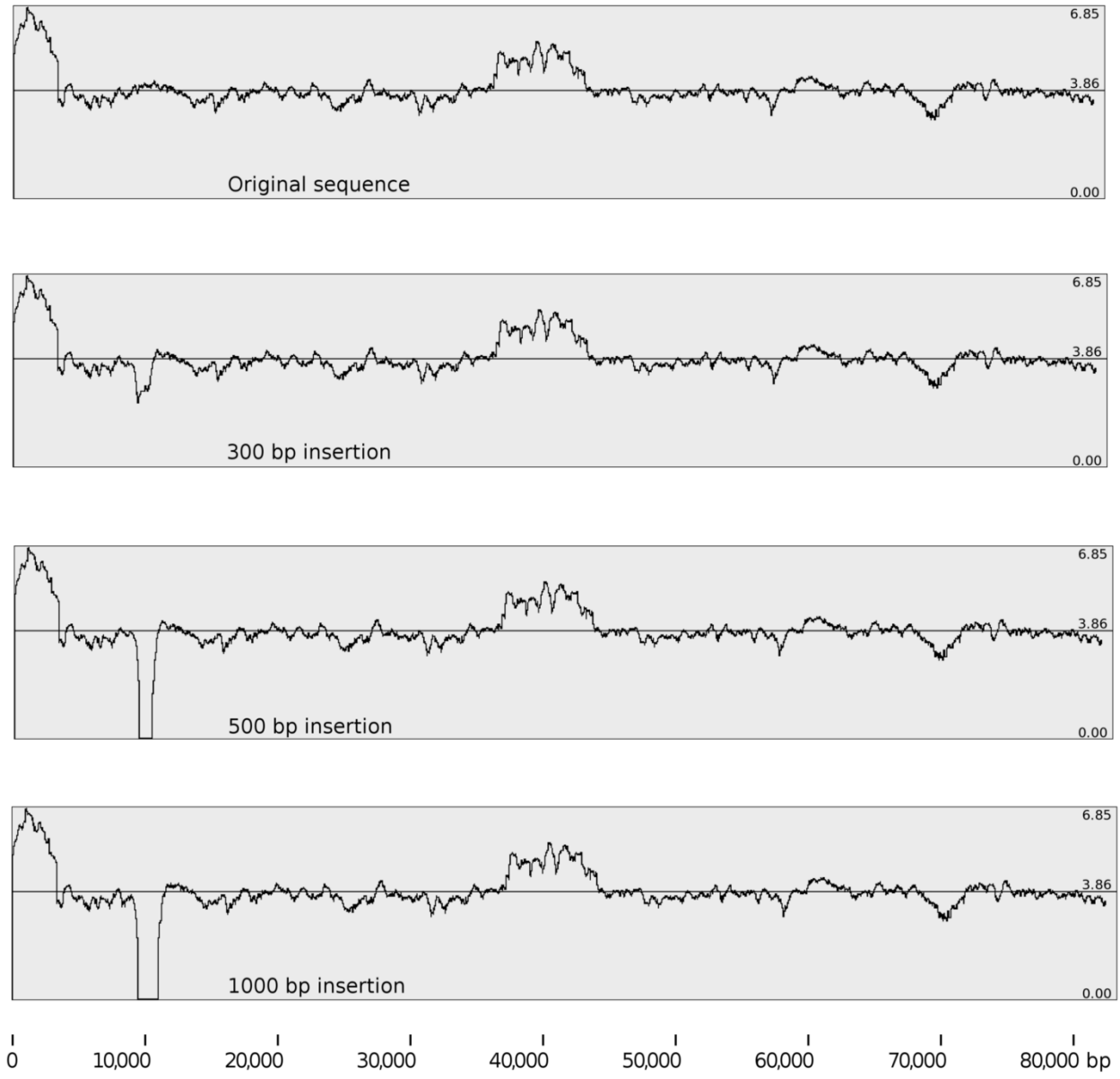


Supplementary Figure 29. Alignment of the anchored WGP physical map FPC contigs to tomato assembly v2.40 via WGP sequence tags. On the x-axis are the assembled pseudomolecules, including chromosome 0, with a cumulative length of 781 Mb (chromosome layout on top in red and blue). On the y-axis are 2,521 WGP FPC contigs, ordered by their strongest anchor point (on the right-hand side is a line at the starting point of each WGP contig in red and blue). Each dot is a BAC with its position on the WGP map and the assembly. There are 52,617 BACs organized on the WGP contigs, of which 97% (51,224) could be firmly anchored on the assembly and thus displayed in this figure. The Pearson correlation coefficient (r) is 0.90 (p-value 0.0).

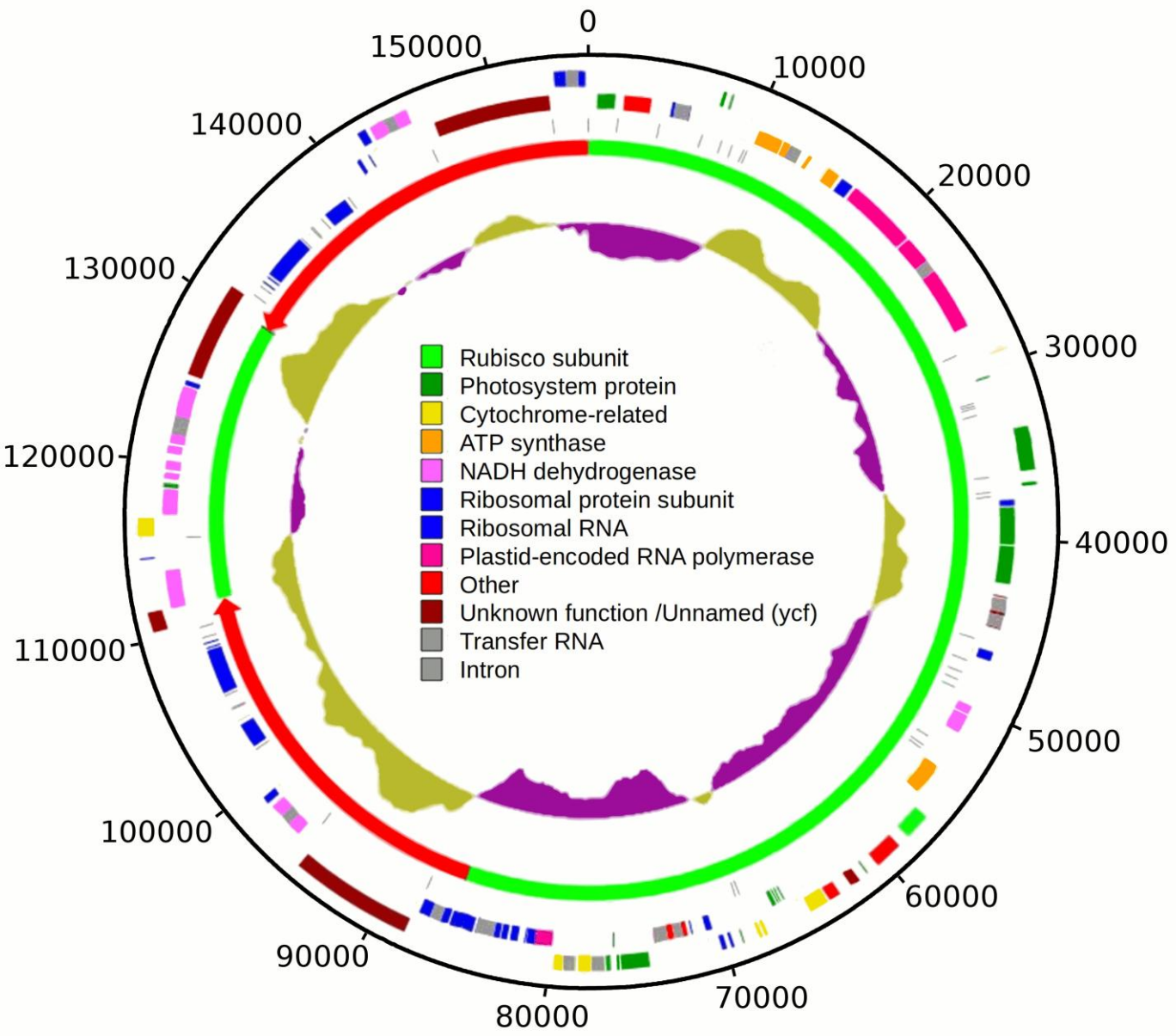




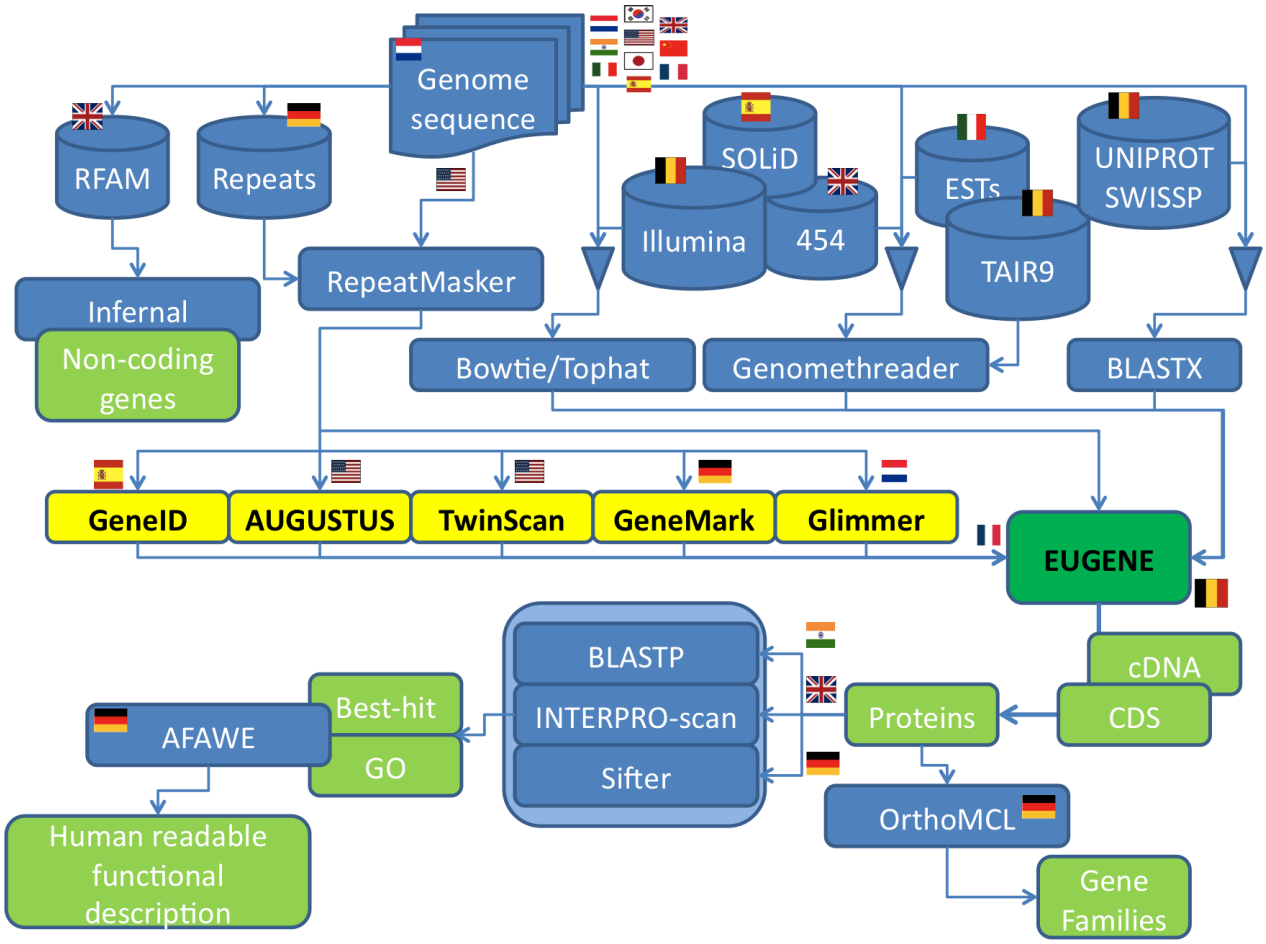
Supplementary Figure 30. Nuclear sequences of mitochondrial origin (*numts*) in tomato. **A.** Schematic representation of a linear map of the tomato mitochondrial genome. The vertical black line represents the current tomato mitochondrial genome assembly (v1.5) and the scale bar (upper right) represents 10 Kb. Black horizontal lines separate the genome between the identified scaffolds (top segment) and contigs (bottom segment). Coloured blocks extending through the genome represent conserved sequences found by applying the Progressive Mauve Alignment software using the nuclear chromosomes as a reference both in sense (right) and antisense (left) orientations. Blocks height and width represent conservedness degree and *numts* length, respectively. Lines connecting each mitochondrial block with nuclear chromosomes (vertical dashes on the right) represent mitochondrial insertion events. Thickness of nuclear chromosomes are proportional to the number of *numts*. **B.** Mitochondrial DNA-derived plasmids hybridized to LA1706 pachytene chromosomes. An equimolar mix of 40 clones with insert sizes ≥ 2.5 kb was labeled with digoxigenin by nick translation and used as a FISH probe (red foci). Simultaneously the BAC clone Le_HBa323E19 was labeled with biotin and used as FISH probe (green foci) to identify the long arm of chromosome *11*. Note the red foci extending to either side of the SC in the long arm of chromosome *11*, which mark long loops of chromatin that are rich in mtDNA. Scattered red foci may mark other sites of mtDNA insertions on other chromosomes. This pattern of hybridisation is in agreement with numerous insertions of mtDNA on chromosome *11* shown in panel A.



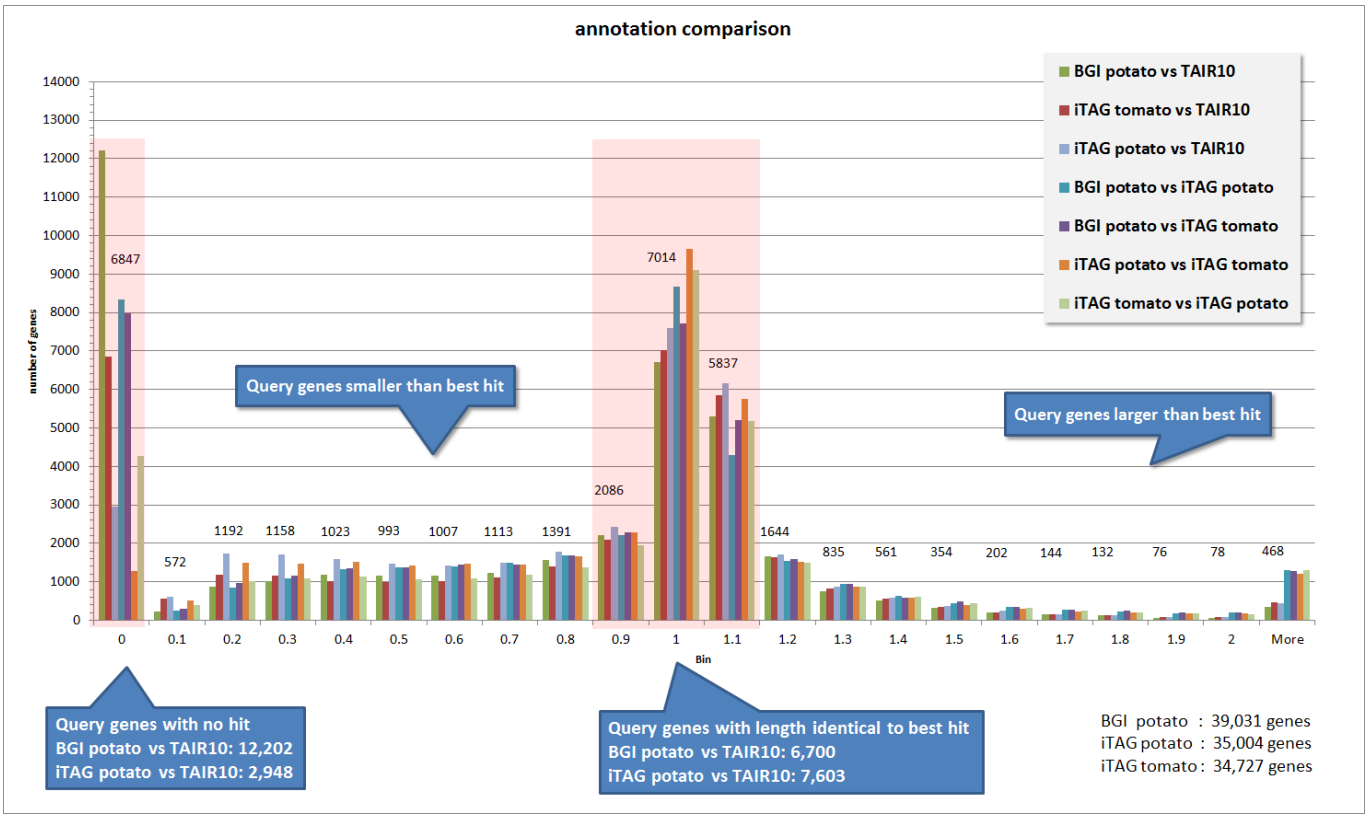
Supplementary Figure 31. Drop of physical coverage due to inappropriate insertion of DNA fragments. A stretch of about 80,000 bp of the tomato genome was considered. The top panel refers to the original sequence while the other panels represent the same sequence in which fragments of plastid DNA, respectively 300, 500 and 1,000 bp long, had been inserted near position 10,000. To build the graphs, the physical coverage given by the 1 kb SOLiD mate-pair library was considered. For each base of the genome the number of mate-pair inserts covering that position was computed, with the condition that the two mate-pairs should align at least 300 bases respectively before and after the position of the base. Values are expressed in log base 10. As seen above, a 300 base pair insertion shows a decrease of the coverage, while insertions of 500 and above are clearly detected, showing zero coverage.



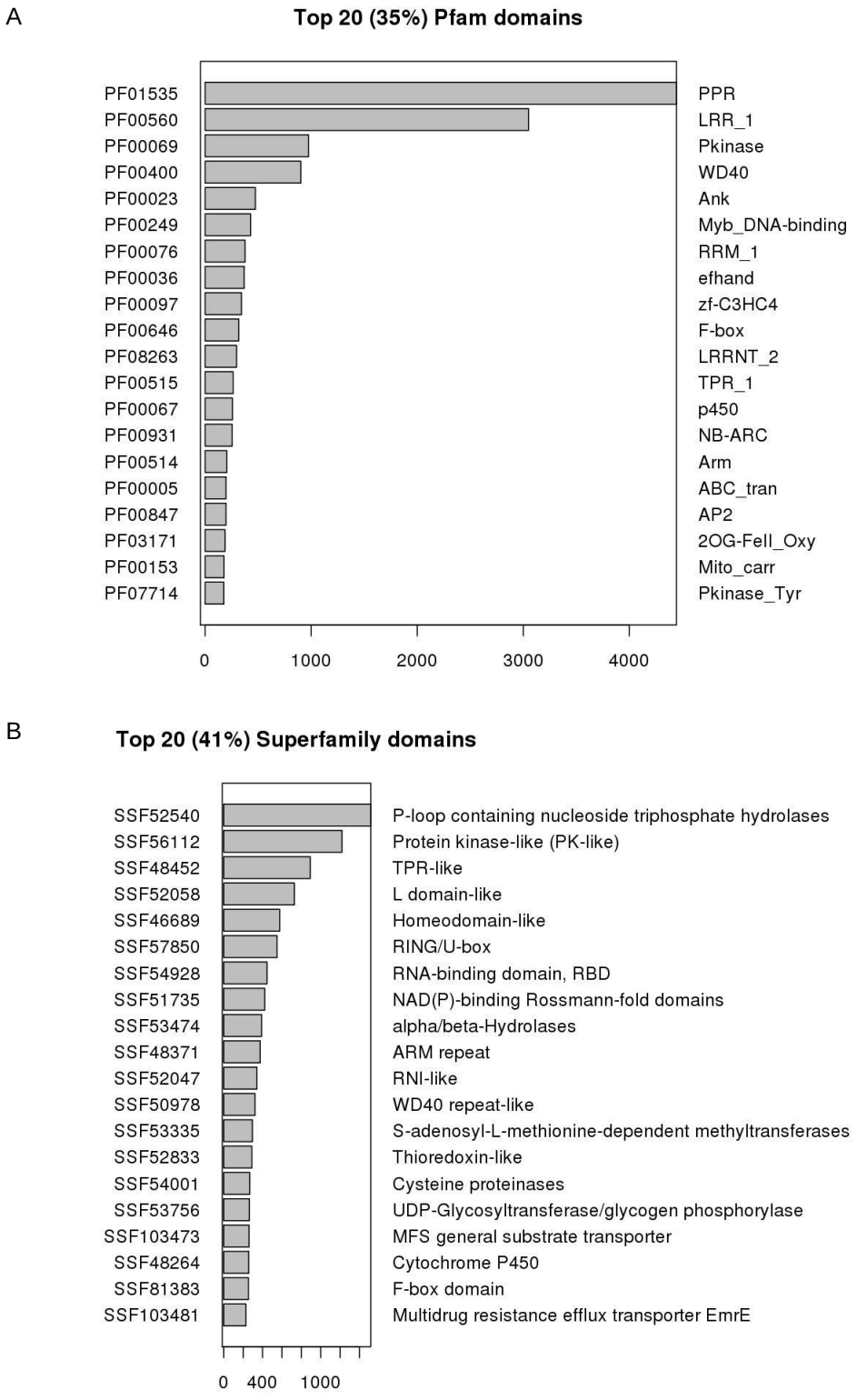
Supplementary Figure 32. Circular map of the chloroplast genome of *S. lycopersicum*. From the outside inwards the circles show: 1) genomic coordinates; 2,3) coloured coding sequences and introns on the plus and minus strands; 4) rRNA and tRNA molecules; 5) the genomic organisation into LSC, SSC (both green) and IR (red); 6) the *per base* chloroplast insertion value, which measures the number of times a plastid base has been detected throughout the nuclear genome. This value is expressed in relation to the mean *per base* insertion value for the chloroplast: purple color means few insertion events, light brown more insertion events. The higher the graph, the less/more insertion events have been detected (i.e.: no insertion events have been observed from the region around 75kb, whereas up to 10 chloroplast fragments from the region around 129kb have been detected in the nuclear genome).



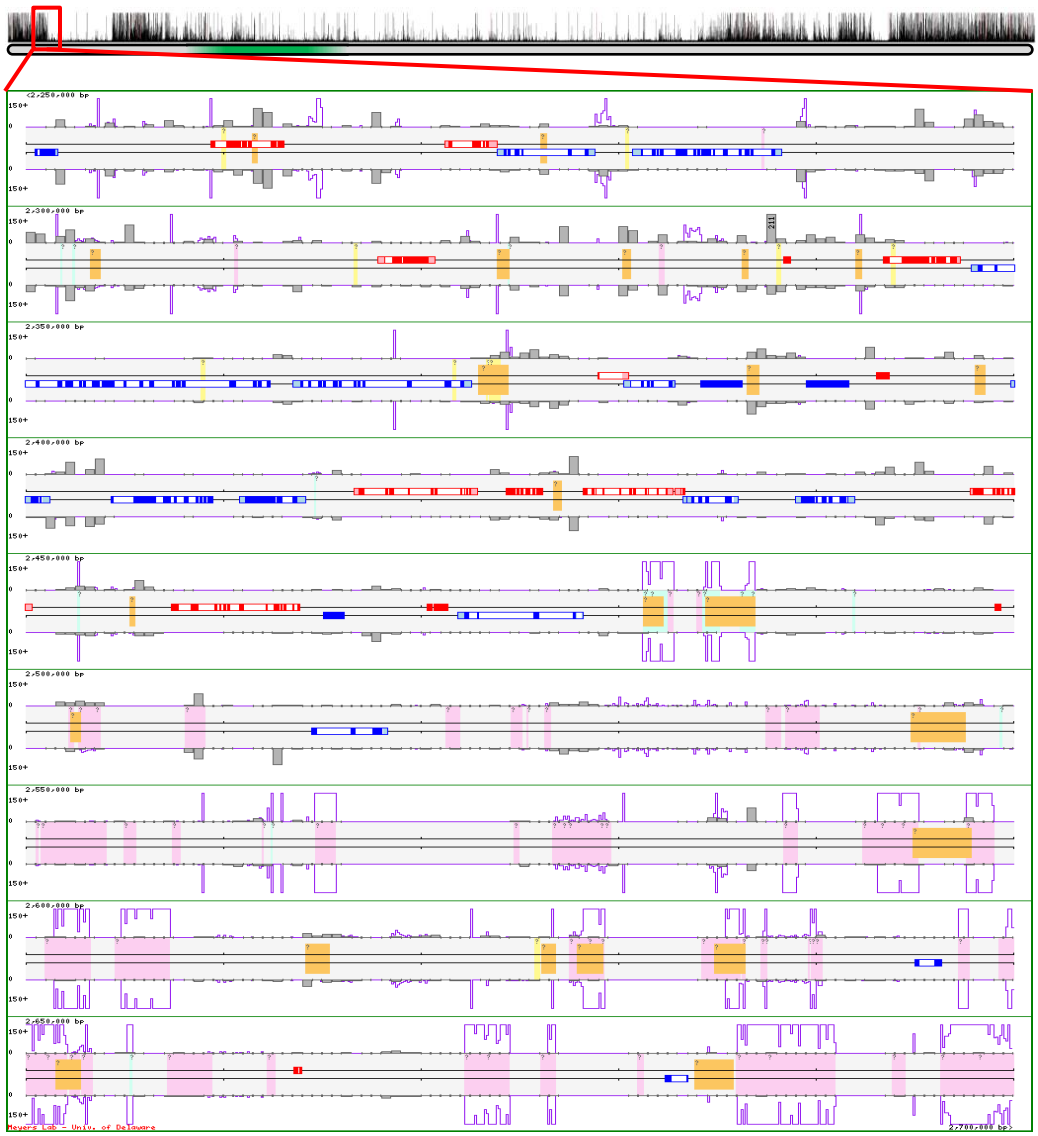
Supplementary Figure 33. The international Tomato Annotation Group (iTAG) annotation pipeline. The flowchart summarizes the different operations and timing of those operations during the whole process of the iTAG annotation.
Analyses were distributed and run in parallel where possible. Cylindrical shapes depict databases, blue squares programs and green squares (intermediate) results. The yellow squares list the other gene-callers used, whose results were integrated in Eugene (dark green box) together with all other extrinsic data. The flowchart before Eugene is structural annotation, past Eugene is the functional annotation and analyses with protein sequences. The flags next each analysis refer to the lab running the analyses (for some countries more than one lab participated).



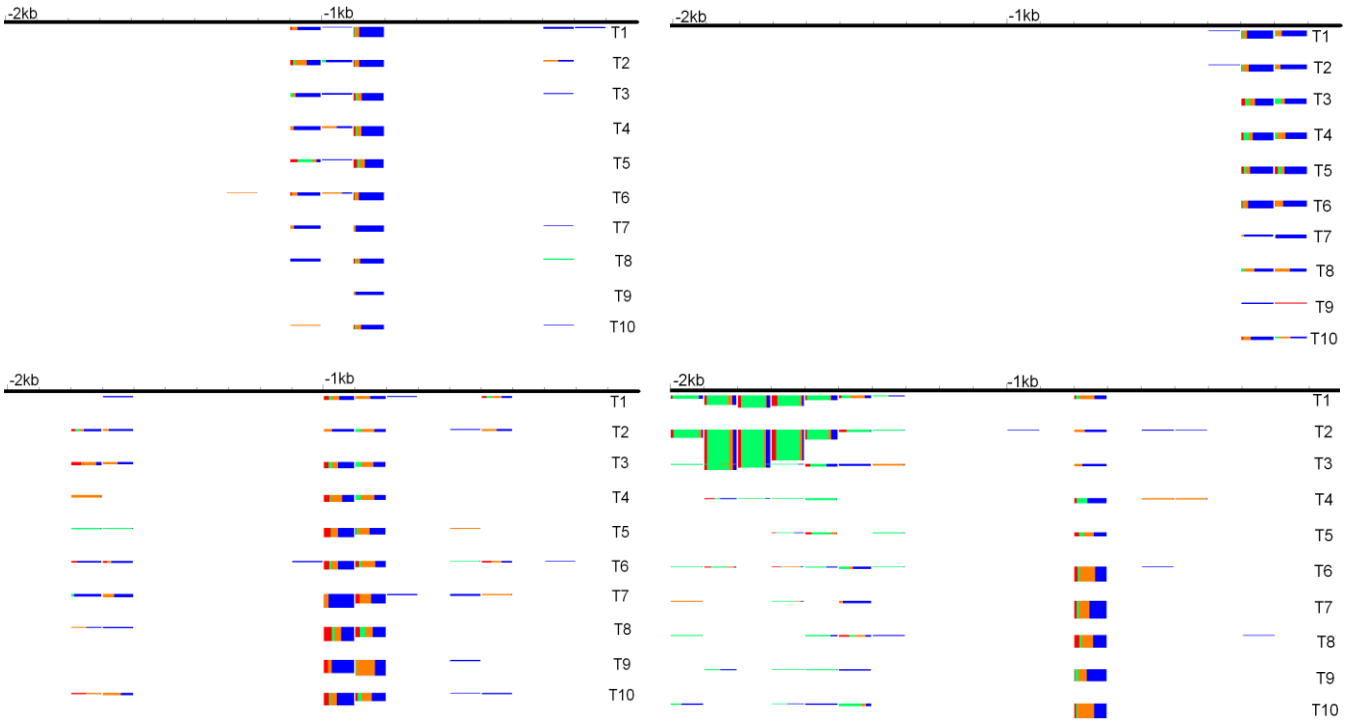
Supplementary Figure 34. Histogram of pairwise comparison of predicted protein length ratios with their best blast hit (BLASTP, e-value e^{-3}) relative to the reference TAIR10 proteome, tomato versus potato (iTAG) or BGI versus iTAG. Proteins with a ratio [0.9 - 1.1] (as indicated by the pinkish area in the middle of the plot; BGI-potato vs TAIR10: 14,199; iTAG-potato vs TAIR10: 16,203) can be considered as highly confident predictions (e.g. 43% of iTAG Tomato genes) as the protein sequences match at the sequence level as well as at the protein length level. Bins of ratios < 0.9 are proteins predicted to be shorter than their *Arabidopsis thaliana* TAIR10 reference, while bins of ratios > 1.1 indicate proteins predicted to be longer. The slightly larger amount to potato genes in the lower bins could be associated with the higher degree of fragmentation of the potato assembly (genes prematurely truncated due to genome sequence issues). The zero-bin, indicates the number of genes not reporting a hit to the reference given the threshold used. This bin will collect all real species specific genes and predicted small genes, but also prediction artifacts (false positives = over prediction) or gene models resulting from genome/assembly issues. The numbers given above the histogram-bars refer to the iTAG tomato prediction compared to TAIR10.



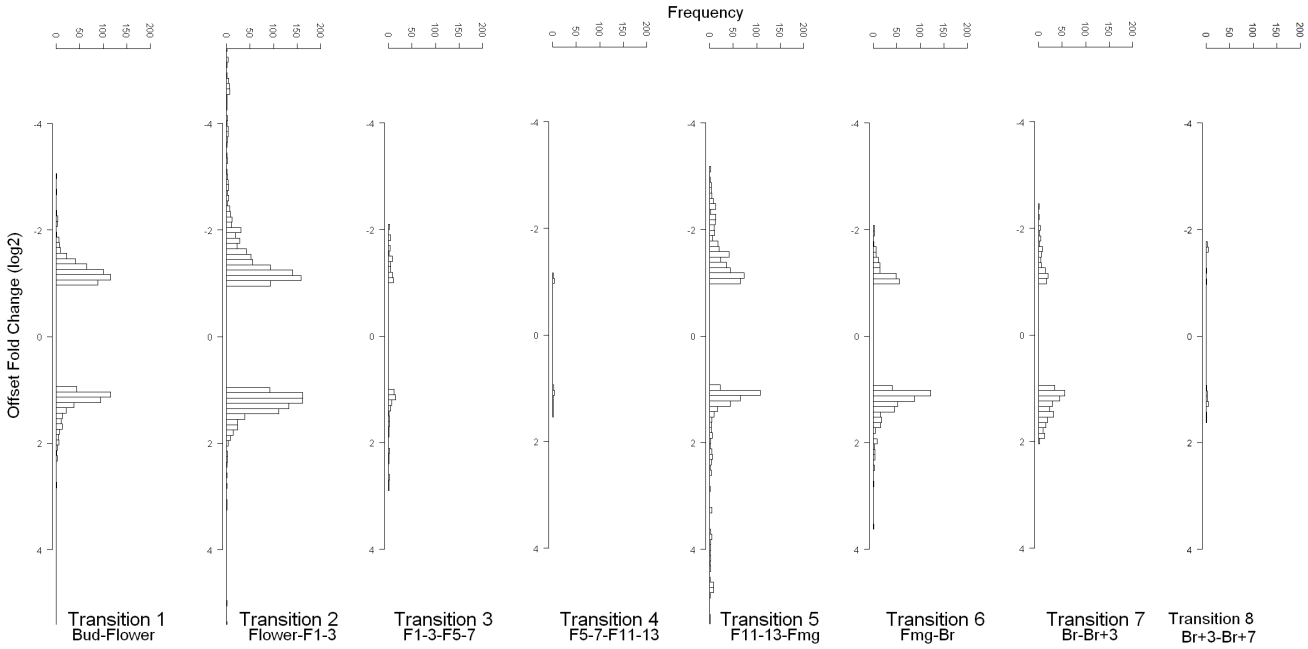
Supplementary Figure 35. INTERPRO domains in tomato annotation. The top 20 Pfam (A) and Superfamily (B) domains.



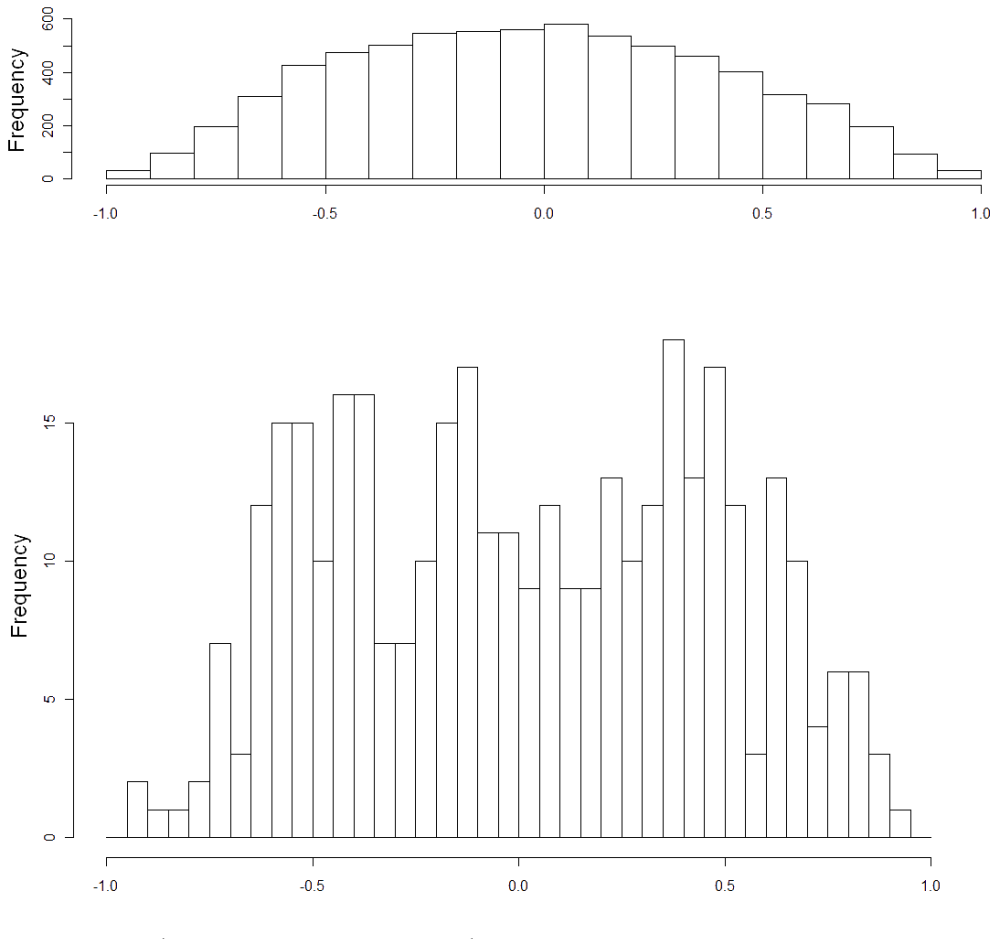
Supplementary Figure 36. Fine structure of a boundary between regions on tomato Chromosome *3* exhibiting high and low sRNA densities. The top image shows the full distribution of sRNA abundances on Chr. *3*, with the red block indicating the specific genomic location shown in the screenshot below. This 450 kb genomic region (below) sits on the boundary of a region of high to low small RNA density. Low small RNA abundances are associated with the high copy retroelements, predominantly Gypsy-class LTR retrotransposons, in the lower portion of the figure. High small RNA abundances are associated with the gene-dense, unannotated regions in the upper portion of the image. In the screenshot, gray histogram bars above or below the central axis (the genome) indicate the small RNA abundances, calculated as the sum of all small RNA "hits-normalized abundances" for three small RNA libraries, calculated for 500 nt bins. The maximum bar height is 150 "transcripts per 4 million" (TP4M). Blue or red boxes are annotated exons interrupted by introns (white boxes internal to the red/blue boxes). Purple lines indicate the k-mer frequency calculated for very 20 nt. Genomic repeats are indicated in yellow, pink, orange or turquoise shading, corresponding to DNA transposons, retrotransposons, inverted repeats, or low complexity repeats, as calculated by RepeatMasker or Einverted. Small question marks indicate that repeat information can be obtained by a mouse-over in the dynamic web browser used to generate the image.



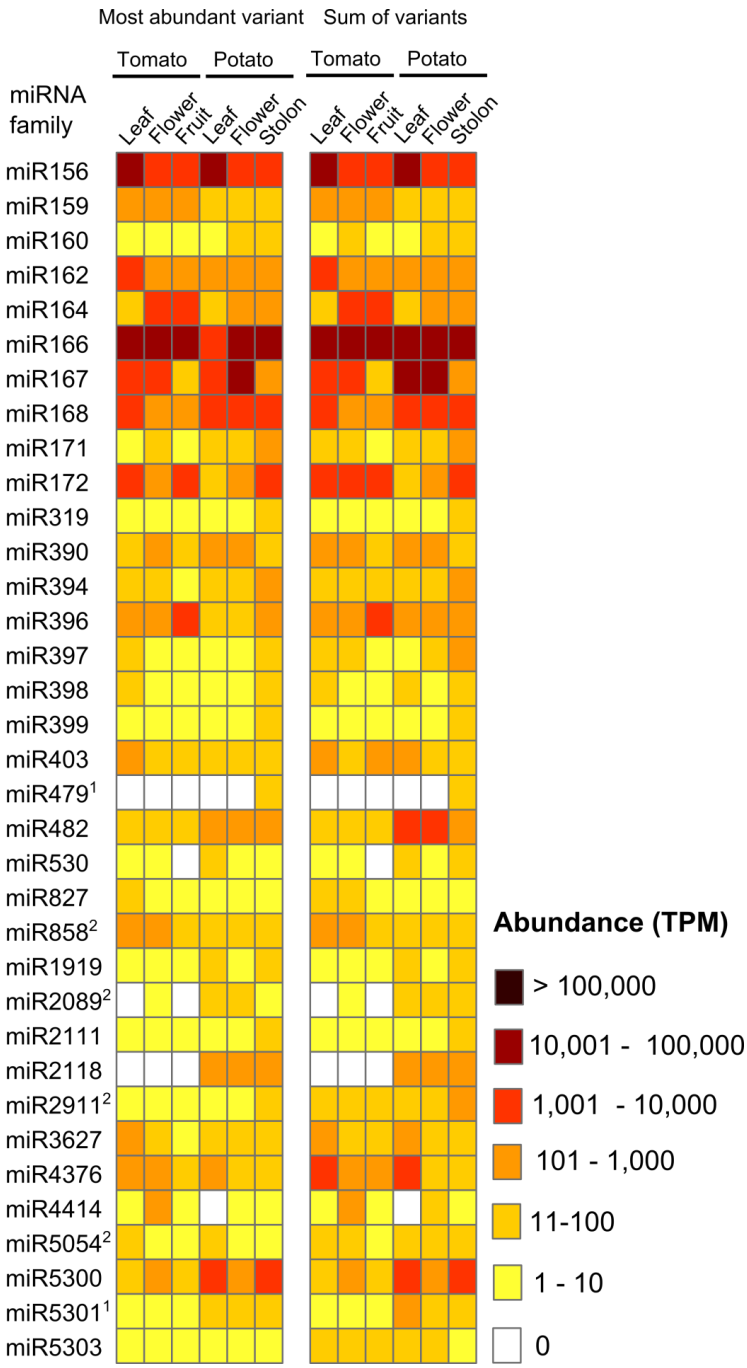
Supplementary Figure 37. Mapping of sRNA reads to promoters of protein coding genes.
Sequences obtained by Illumina sequencing of sRNA libraries at ten stages during fruit development (T1 – bud, T2 – flower, T3 – fruit 1-3mm, T4 – fruit 5-7mm, T5 – fruit 11-13mm, T6 – fruit mature green, T7 – breaker, T8 – breaker+3days, T9 – breaker + 5days, T10-breaker+7days)[127] were mapped to promoter regions of protein coding genes. The sRNAs are grouped in 100nt windows and for each window the size class distribution on redundant reads is shown. The colours corresponding to the 21-24 size classes are: 21 – red, 22 – green, 23 – orange, 24 – blue. The height of the following boxes is proportional with the log offset fold change (offset = 20, which means 20 was added to the read number of each sequence before calculating the fold difference) relative to the first time point. 2kb upstream region for each gene (Solyc08g0080940 top left, Solyc01g081310 top right, Solyc01g088010 bottom left and Solyc05g006350 bottom right) is shown.



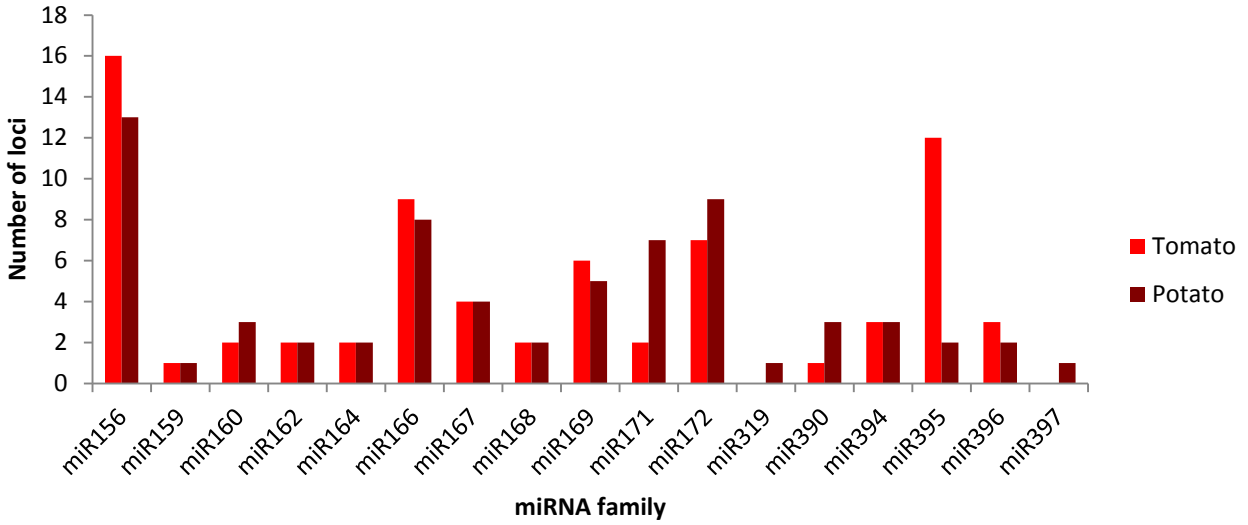
Supplementary Figure 38A. Distribution of offset fold change (OFC), with offset = 20, in log2 scale, computed for differentially expressed (at least 2-fold) promoter mapping sRNAs (2,687 promoters). The OFC was computed on consecutive time points for the 8 transitions corresponding to 9 time points (flower bud, open flower, fruit 1-3mm (F1-3), fruit 5-7mm (F5-7), fruit 11-13mm (F11-13), mature green (Fmg), breaker (Br), breaker+3 days (Br+3) and breaker + 7 days (Br+7)). The x axis shows the frequency of promoters with differentially expressed sRNAs ($|OFC|>2$). The y axis shows the OFC in log 2 scale (minus indicates up-regulation and plus refers to down-regulation). The histograms demonstrate that sRNAs are differentially expressed at many promoters during key developmental transitions (e.g. bud to flower, flower to fruit, mature green to ripening fruit) but their expression is very stable outside of these transitions (e.g. fruit growth or fruit ripening).

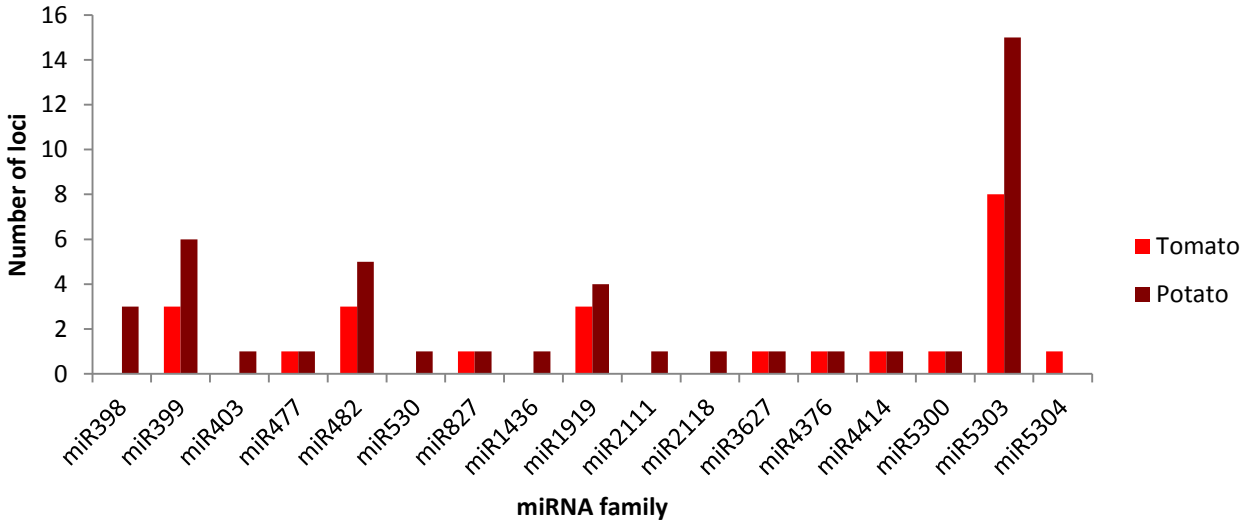


Supplementary Figure 38B. Distribution of correlations (Pearson Correlation Coefficients) between expression of mRNA and sRNAs mapping to the corresponding promoters for all 7,301 genes on the Affymetrix tomato genome array expressing sRNAs that could be mapped to the tomato genome (top panel) and for 358 genes present on the array and producing differentially expressed sRNAs from their promoters (bottom panel). The distribution presented in the top panel is not different from a distribution of correlation on randomly generated series with 9 points using a χ^2 test. The distribution presented in the bottom panel is significantly different from a distribution on randomly generated series ($p < 0.01$).



Supplementary Figure 39. Expression of conserved miRNAs in tomato and potato samples.
MiRNA indentification was performed using a genome-independent approach based on the comparison of plant miRNA sequences from miRBase (release 17, April 2011) and small RNA sequencing data from three tissues per species (See **Supplementary Section 2.9** for further details). Abundances of 35 miRNA families that were expressed beyond the background noise (10 transcripts per million [TPM]) in at least one sample are displayed in two heatmaps, showing the sum of abundances of all the sequence variants (right) or the abundance of the most frequent variant (left) for each miRNA family. Notes: [1] miRNAs identified as miR* of other families. [2] miRNAs that were not mapped on the genome with high confidence.

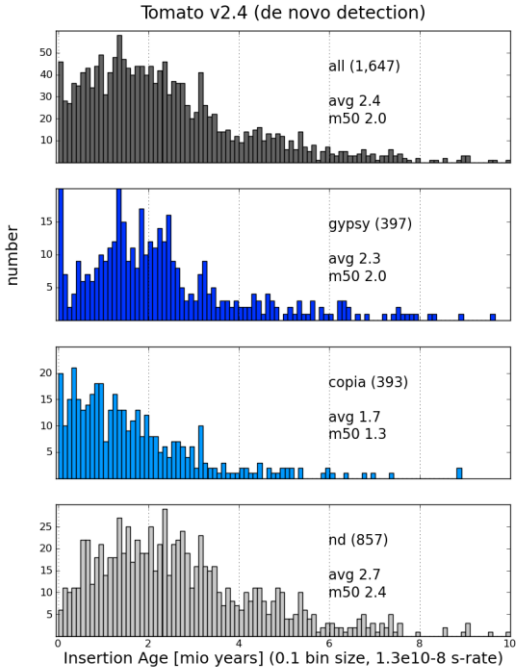


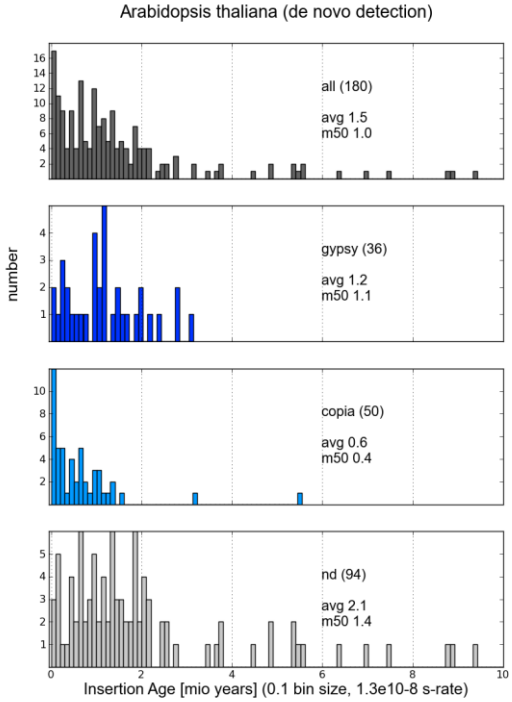


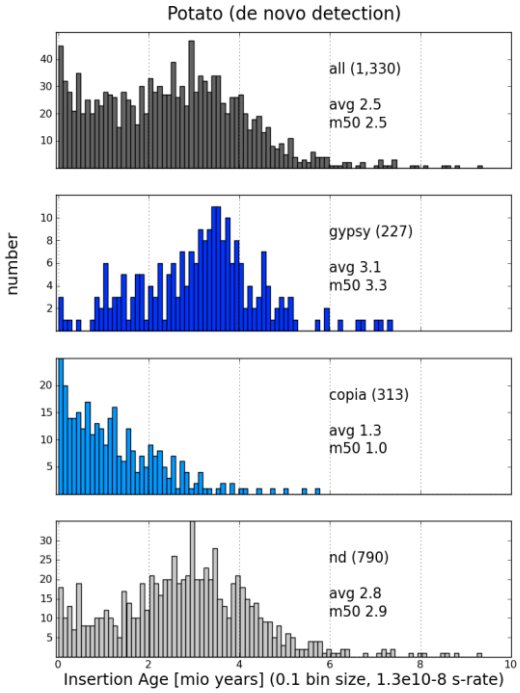
Total miRNA loci

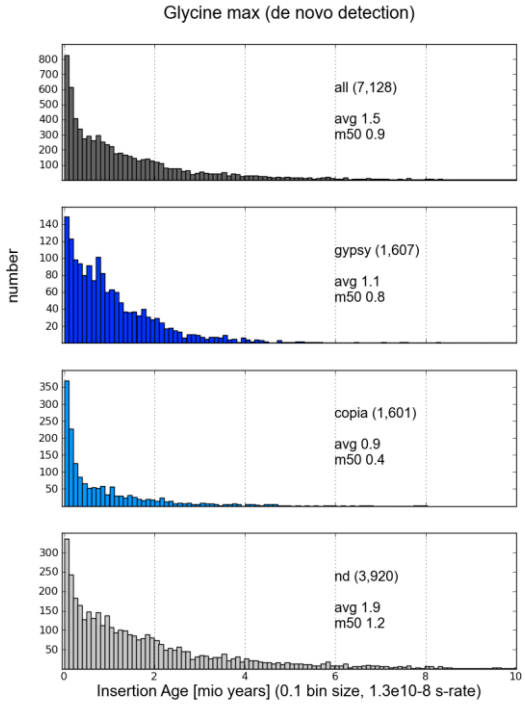
Tomato (*S. lycopersicum*)	96
Potato (*S. tuberosum*)	120

Supplementary Figure 40. Number of known miRNA loci mapped in the tomato and potato genomes. Genomic loci encoding conserved miRNAs were identified using data from miRBase (release 17.0, April 2011), small RNA expression data from three tissues per species and criteria described in Meyers et al, 2008 (See **Supplementary Section 2.9** for further details). The number of mapped loci is largely compatible between tomato and potato for a subset of miRNAs that are ubiquitous or widely conserved in plants (miR156, miR160, miR162, miR164, miR166, miR167, miR168, miR169, miR172). While the excess of mi5303 in potato is balanced by the abundance of miR395 loci in tomato, the higher number of loci identified in potato is accounted for by an increased number of loci for shared miRNA families as well as the presence of miRNAs that could not be identified in the tomato genome.

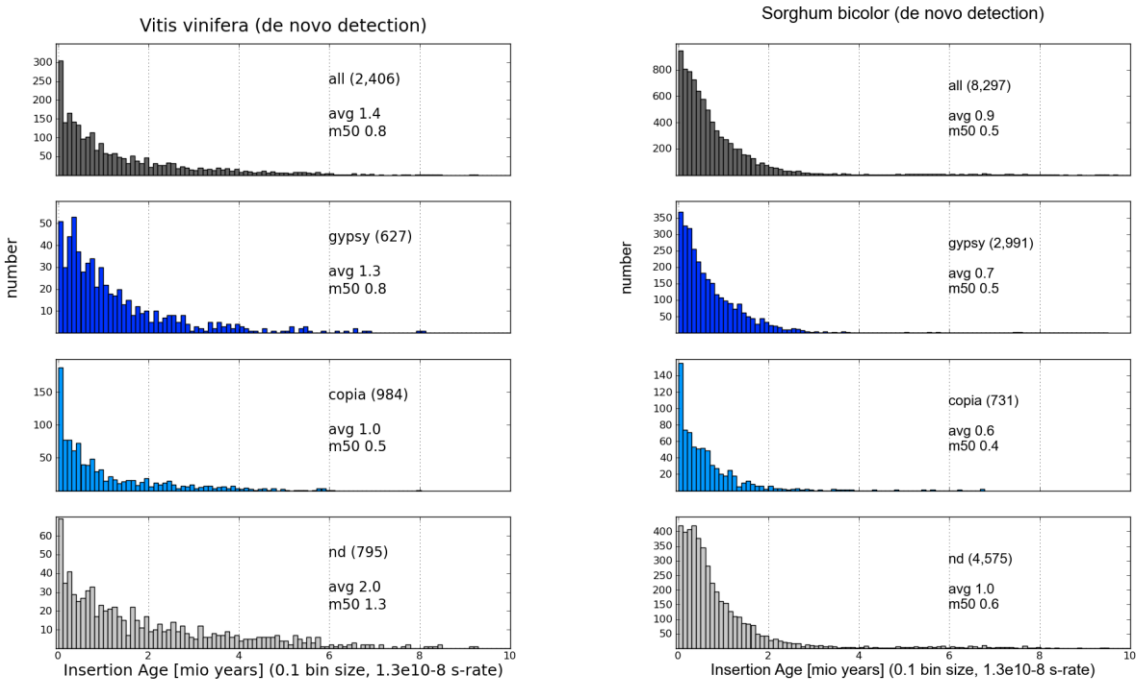




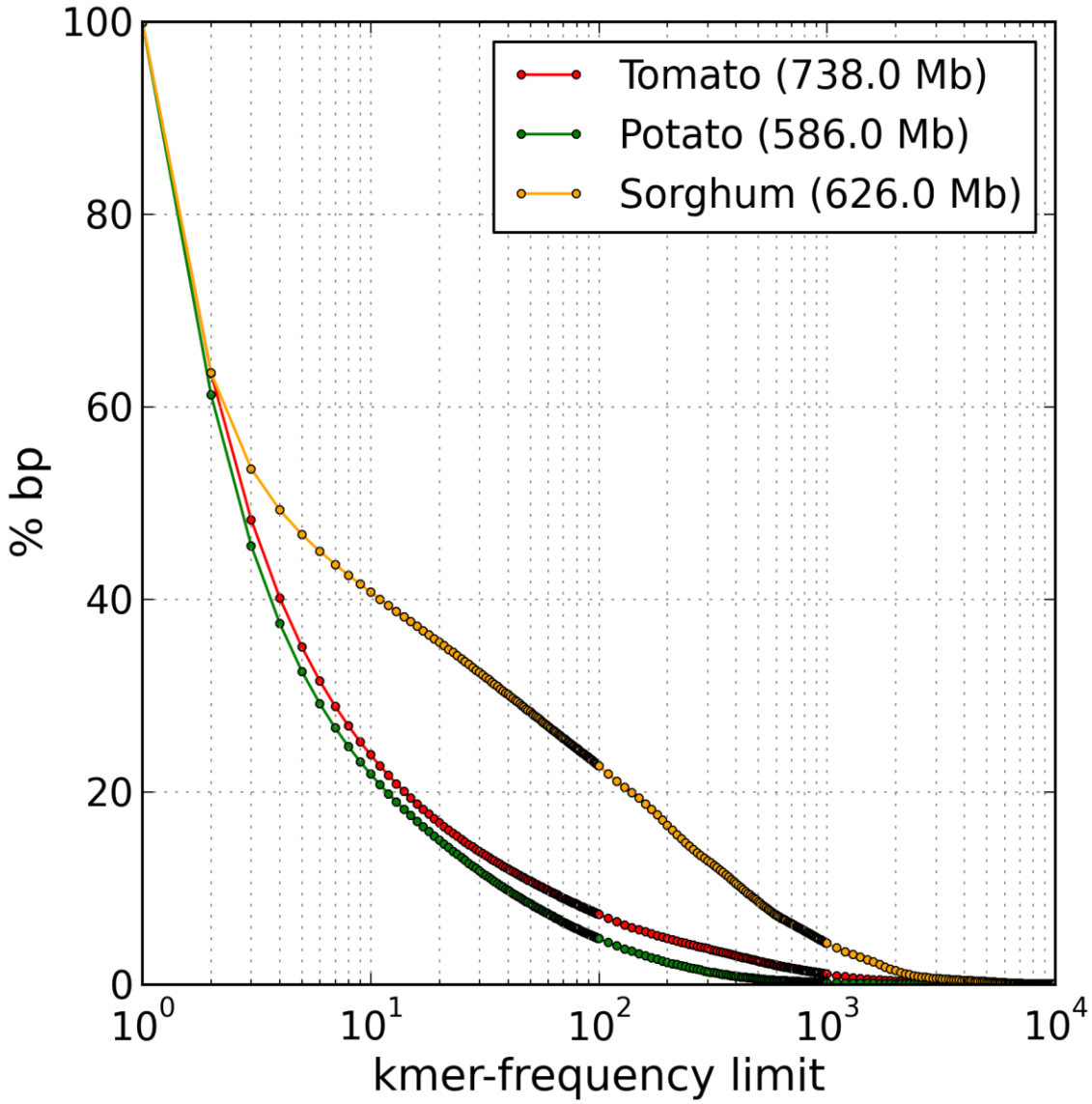




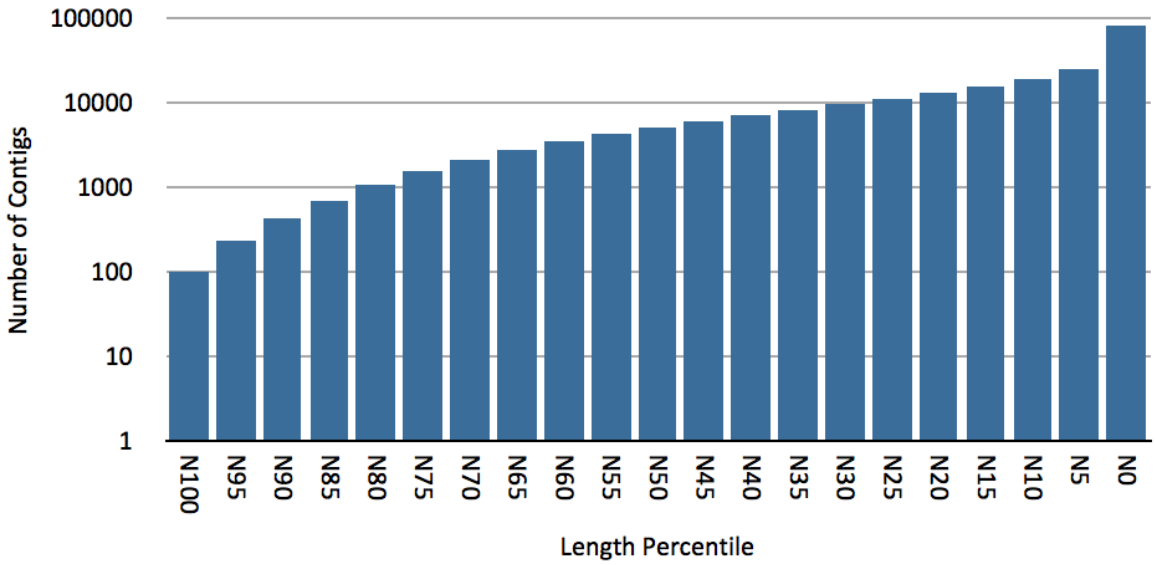
Supplementary figure 41. (Continued on next page)



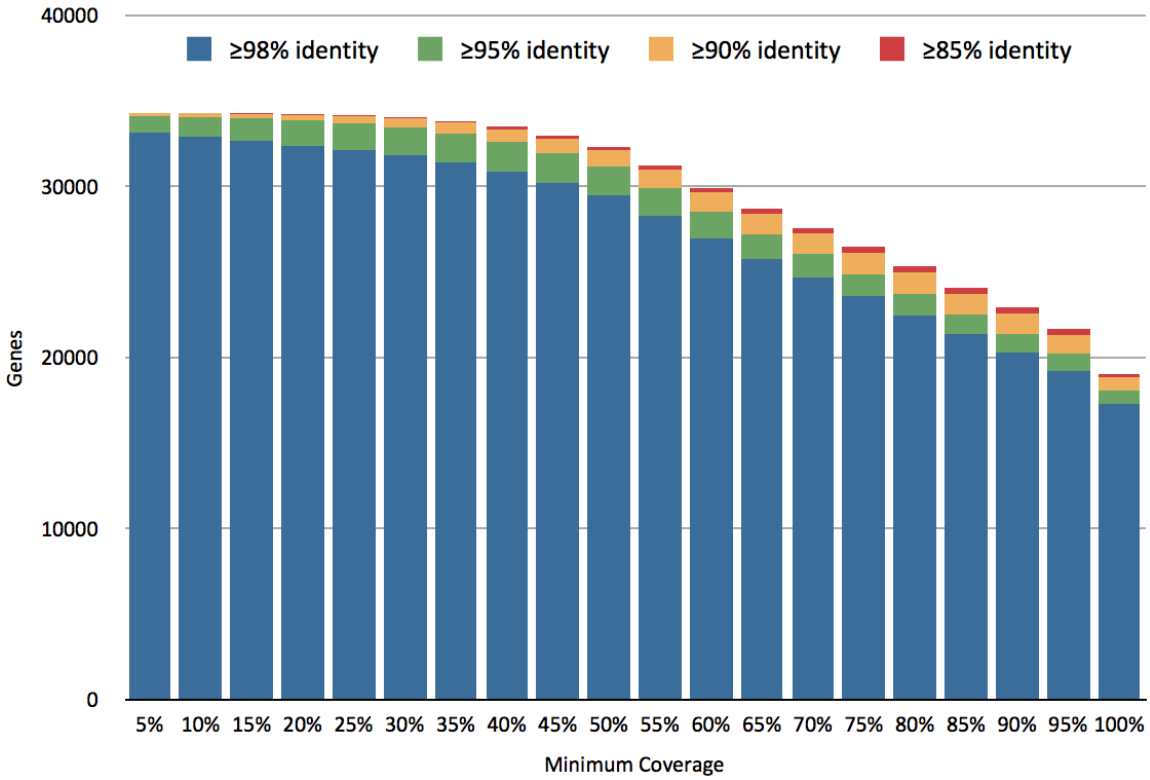
Supplementary figure 41. Age distribution of LTR-retrotransposon insertion in different species.



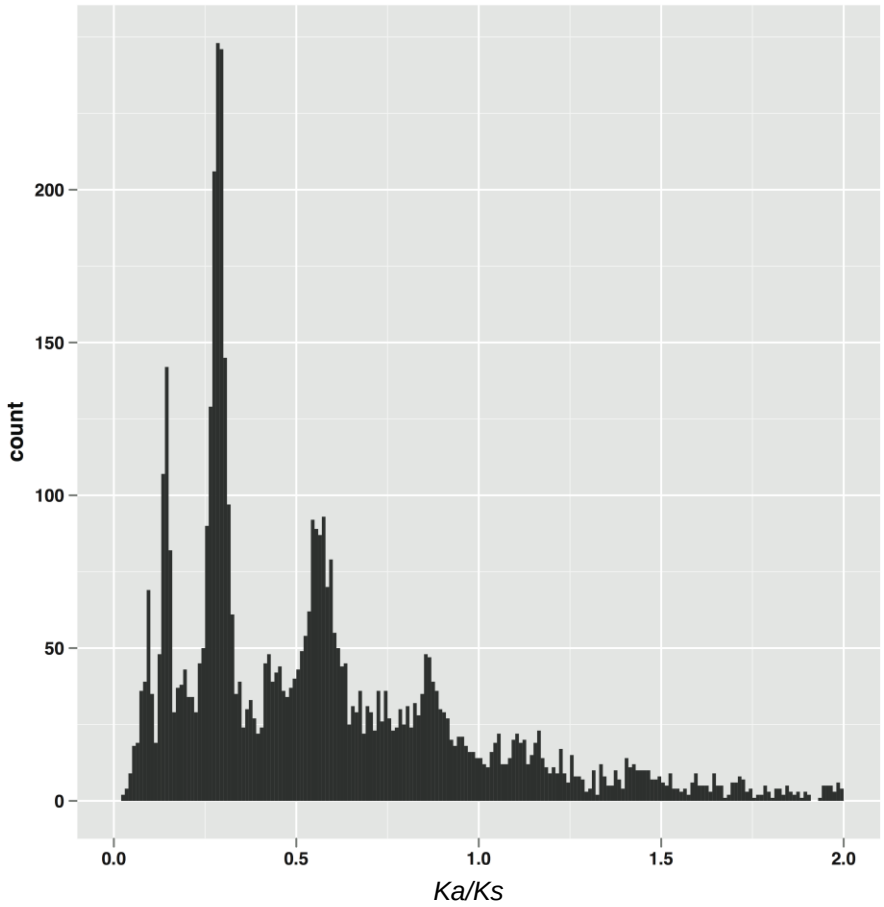
Supplementary Figure 42. Assessment of genome repetitivity *via* 16mer frequencies.
The cumulated coverage of the genome in base pairs is plotted against increasing 16mer frequencies. For instance all 16mers occurring ≥10 times account for 24% of the tomato and 22% of the potato genome. Both tomato and potato have a distinctly lower repetitive content compared to the similar sized Sorghum, where 41% of the genome are composed of 16mers with frequencies ≥10.



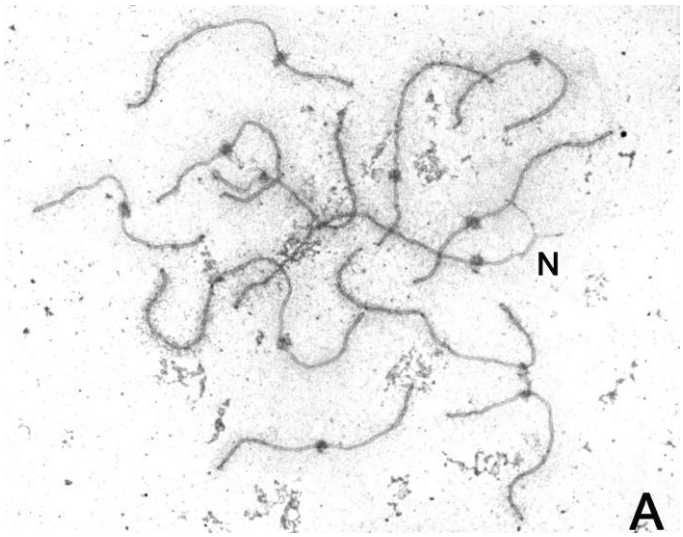
Supplementary Figure 43. Contig length distribution of the *de novo* assembly of *S. pimpinellifolium*.

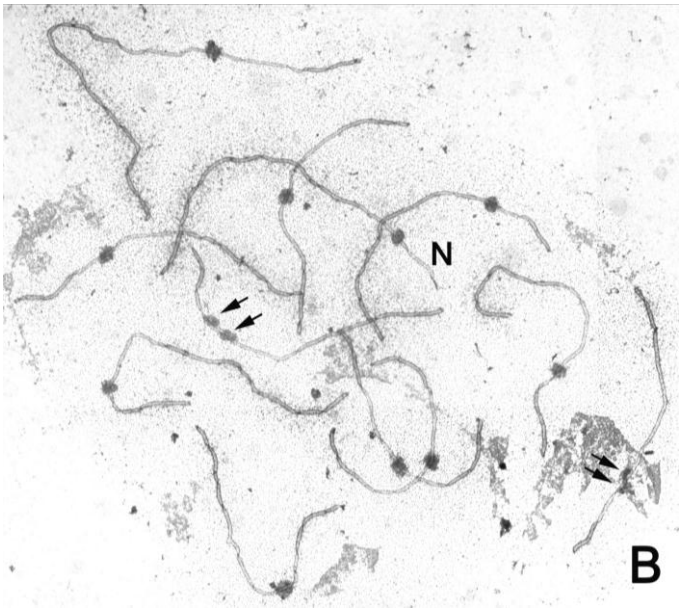


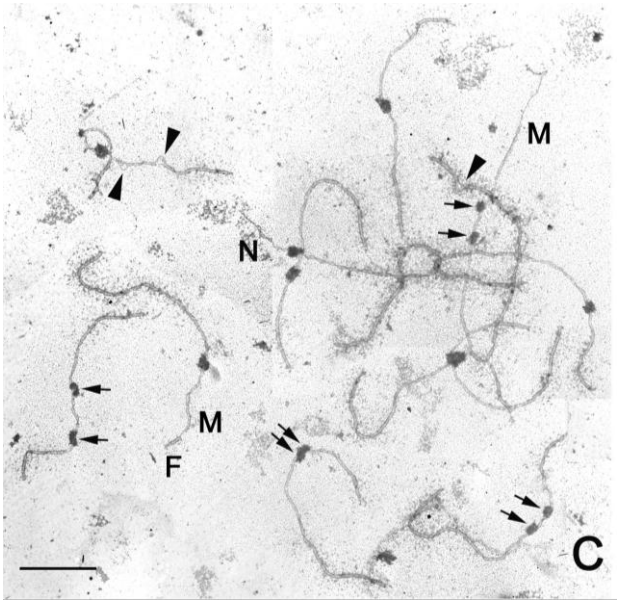
Supplementary Figure 44. Coverage of *S. lycopersicum* cv. 'Heinz 1706' genes (ITAG 2.3) in the *S. pimpinellifolium de novo* assembly.



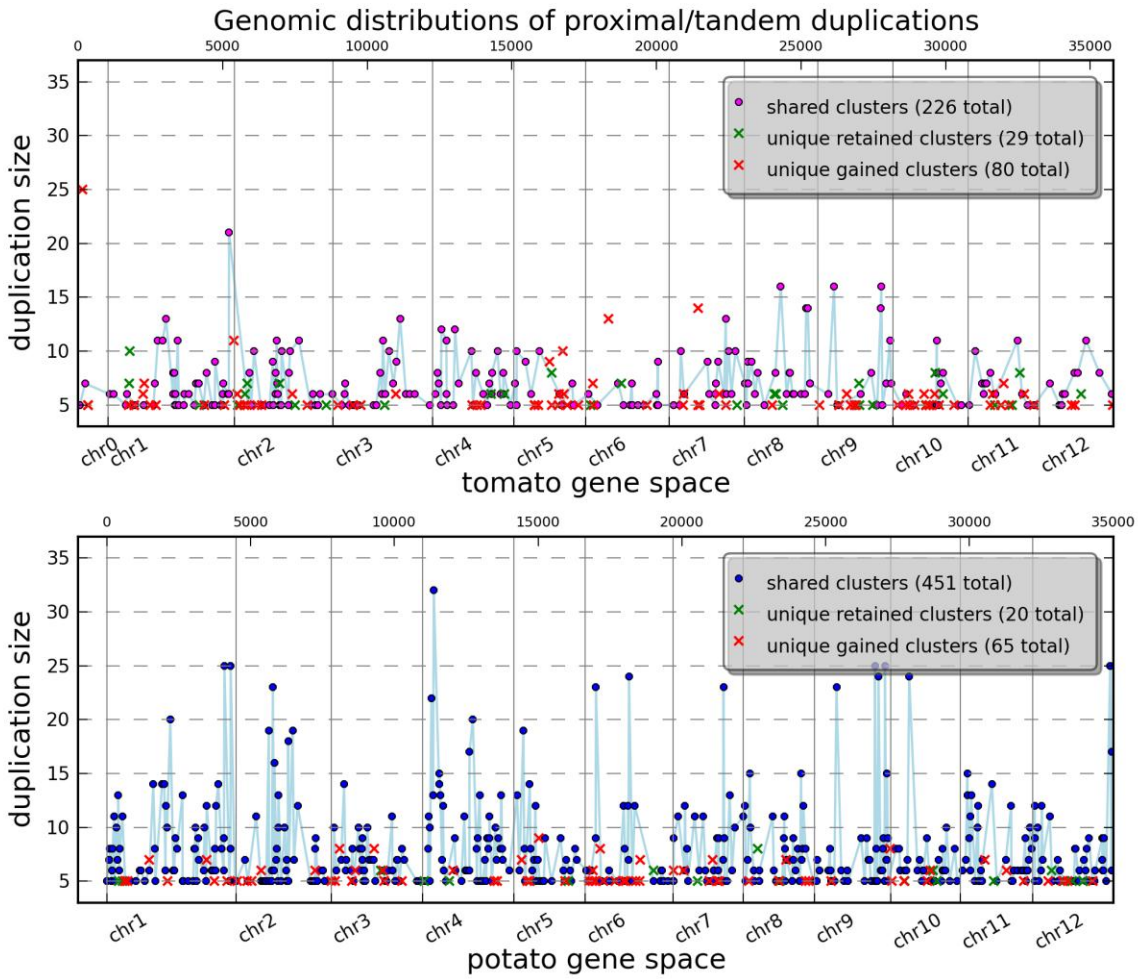
Supplementary Figure 45. Distribution of *Ka/Ks* for 16,467 genes with valid values, i. e. with non-zero Ks values, as a result of pairwise alignment and calculation of synonymous and non-synonymous changes between *S. pimpinellifolium* and *S. lycopersicum*.



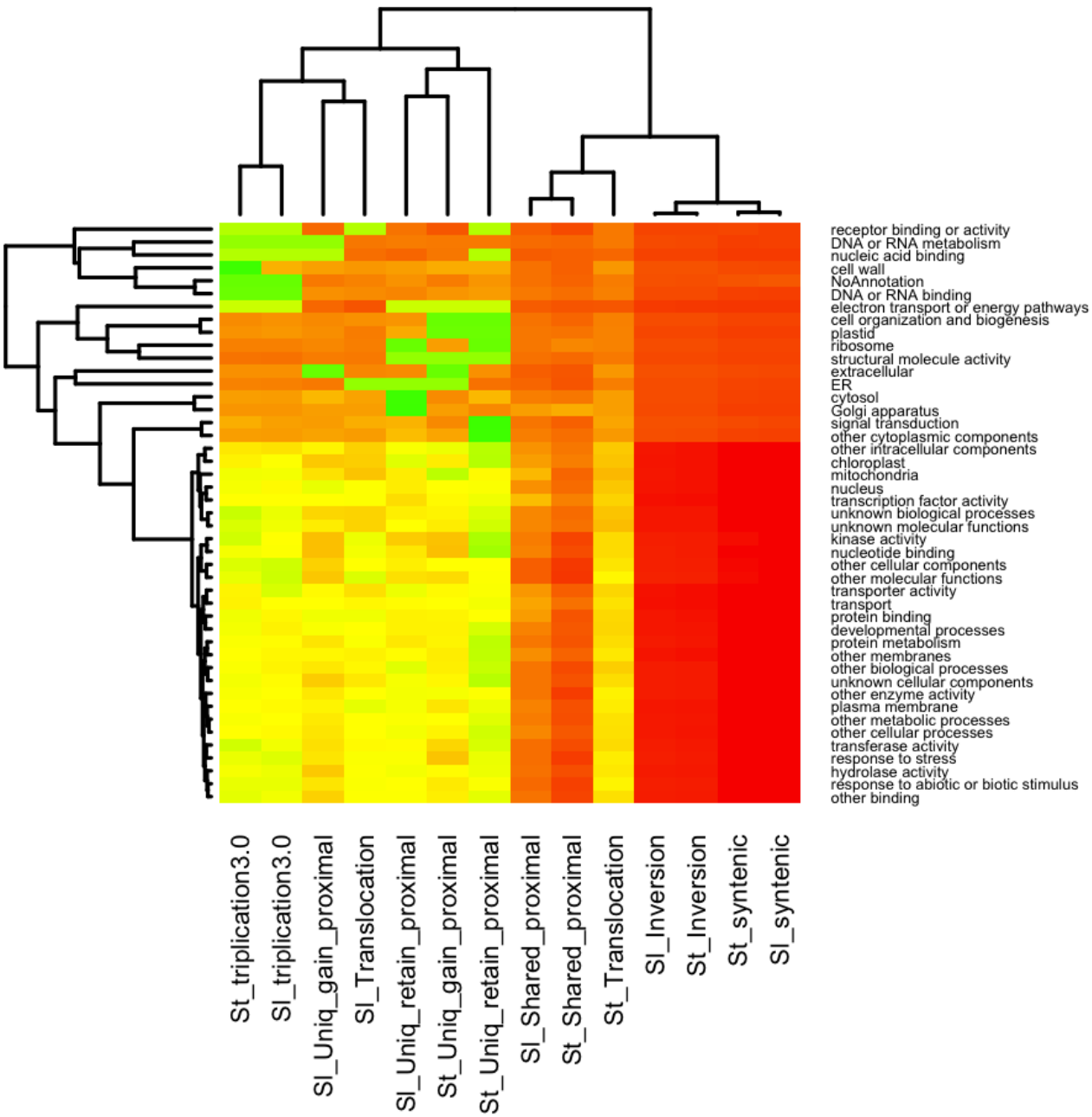




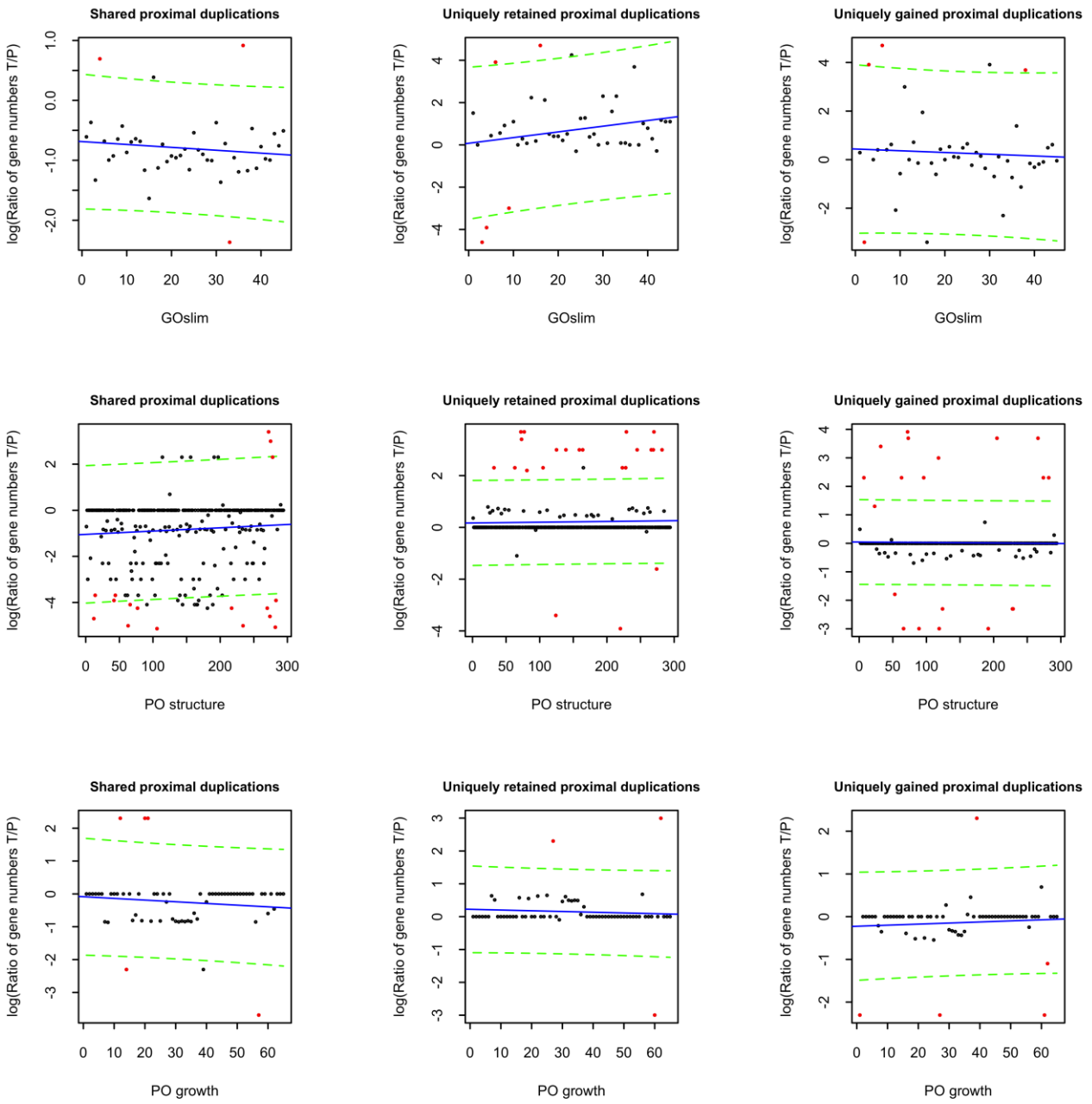
Supplementary Figure 46. Complete diploid sets of late pachytene synaptonemal complexes (SCs = pachytene chromosomes) from (**A**) *S. lycopersicum* (tomato), (**B**) *S. lycopersicum* X *S. pimpinellifolium* hybrid, and (**C**) *S. lycopersicum* X *S. pennellii* hybrid. Kinetochores are irregular, darkly stained ellipsoids about 1 µm in their longest dimension on each SC. (**A**) Synapsis is complete in all twelve bivalents, except the end of chromosome *2* where the NOR is often broken and/or desynapsed (N). (**B**) Synapsis is complete in all twelve bivalents, except the end of chromosome *2* where the NOR is often broken and/or desynapsed (N), and two of the bivalents show pairs of mismatched kinetochores (double arrows). (**C**) There are many synaptic irregularities in the twelve bivalents, including four pairs of mismatched kinetochores (double arrows – often five pairs are observed), asynapsed buckles (arrow heads), mismatched ends due to length differences (M), foldback synapsis (F), and irregular synapsis in the NOR (N). In early pachytene there is often a small inversion loop that is not visible here, probably due to adjustment to straight nonhomologous synapsis by late pachytene. The scale bar represents 5 µm.



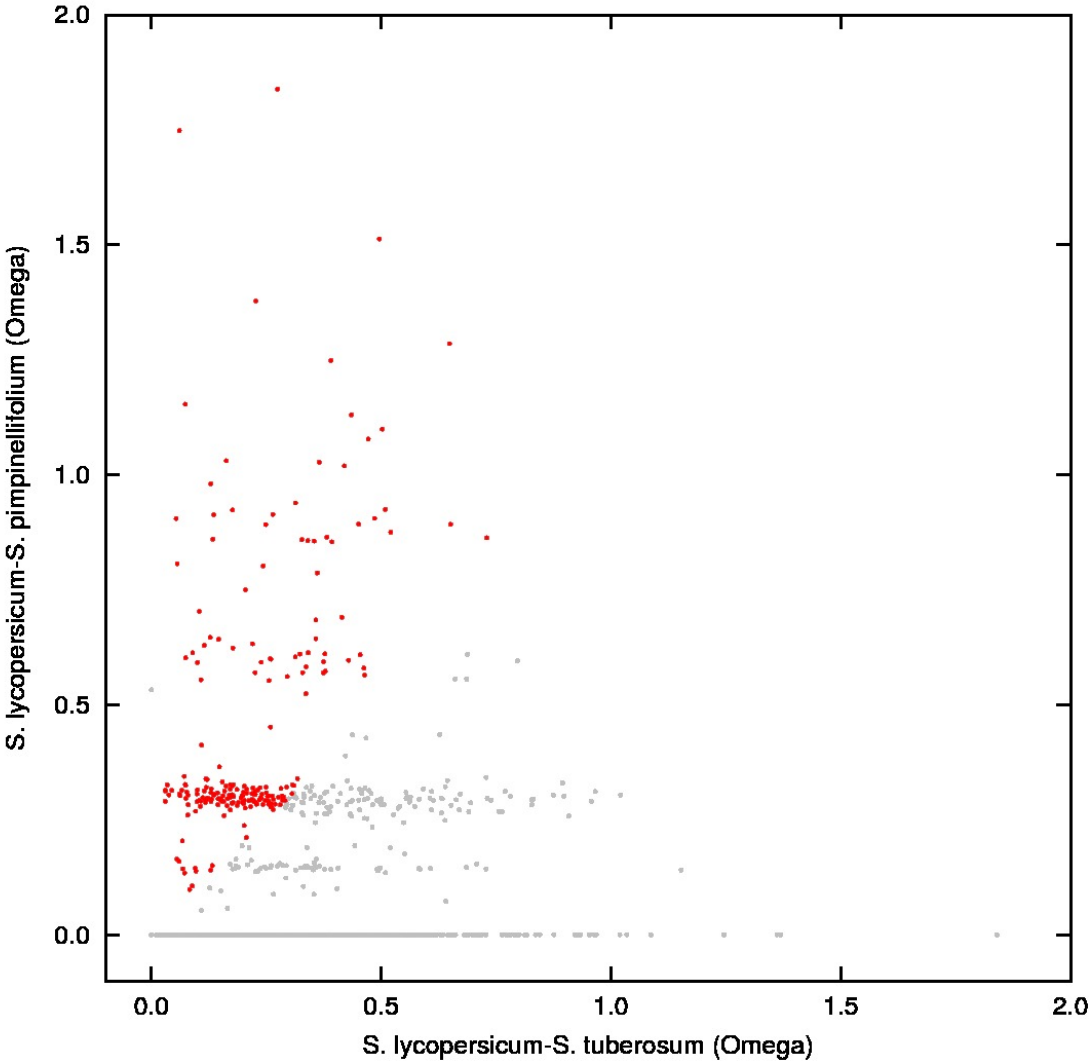
Supplementary Figure 47. Distribution of proximal/tandem duplicated gene clusters in the tomato and potato genomes. Proximally or tandemly duplicated genes are considered clustered if no more than 10 genes apart. Clusters are defined as containing 5 or more proximally duplicated genes. The clusters are plotted along the chromosomes in the tomato (upper panel) and potato (lower panel) genome. The x axis is genomic gene space in rank scale. The y axis is size of each duplication cluster. Most of the proximal/tandem duplicated clusters are shared between the two genomes (magenta dots in tomato and blue in potato), albeit possibly being in different locations and/or of different sizes. Unique clusters are either gained (red crosses) or retained (green crosses) in one of the genomes.



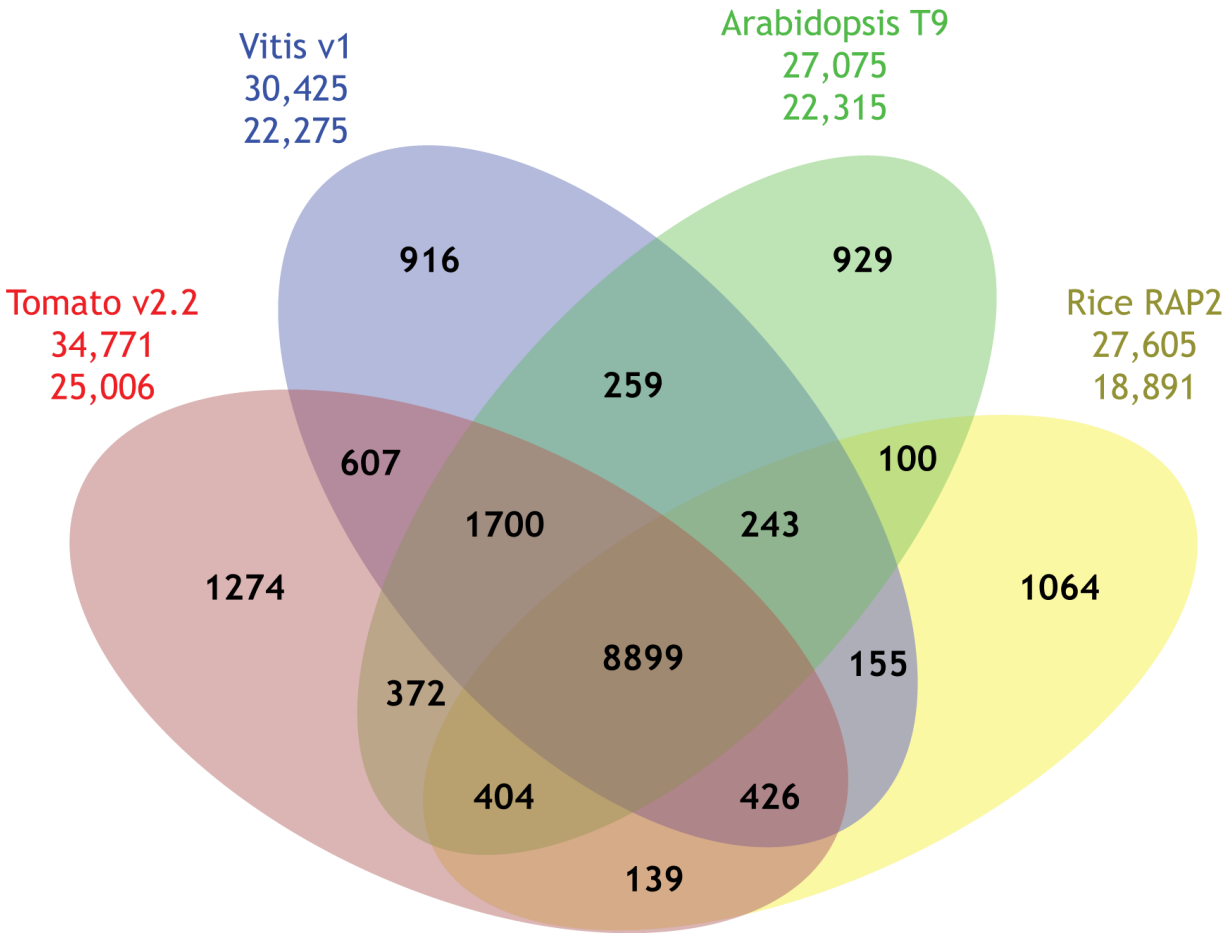
Supplementary Figure 48. Two-way clustering of structural and functional features in tomato and potato genomes. The heatmap shows different enrichment of GOslim terms (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/) in the shared and unique proximally/tandemly duplicated genes and other structural feature groups in tomato and potato genomes. Measurements were taken as the proportion of genes in each structural group (horizontal axis) out of all genes in the genome annotated to each GOslim group (vertical axis). Colour represents normalized gene counts, from low to high (green to red). Dendrograms are calculated by agglomerative hierarchical clustering. The clusters of top hierarchies are supported by bootstrapping (data not shown.



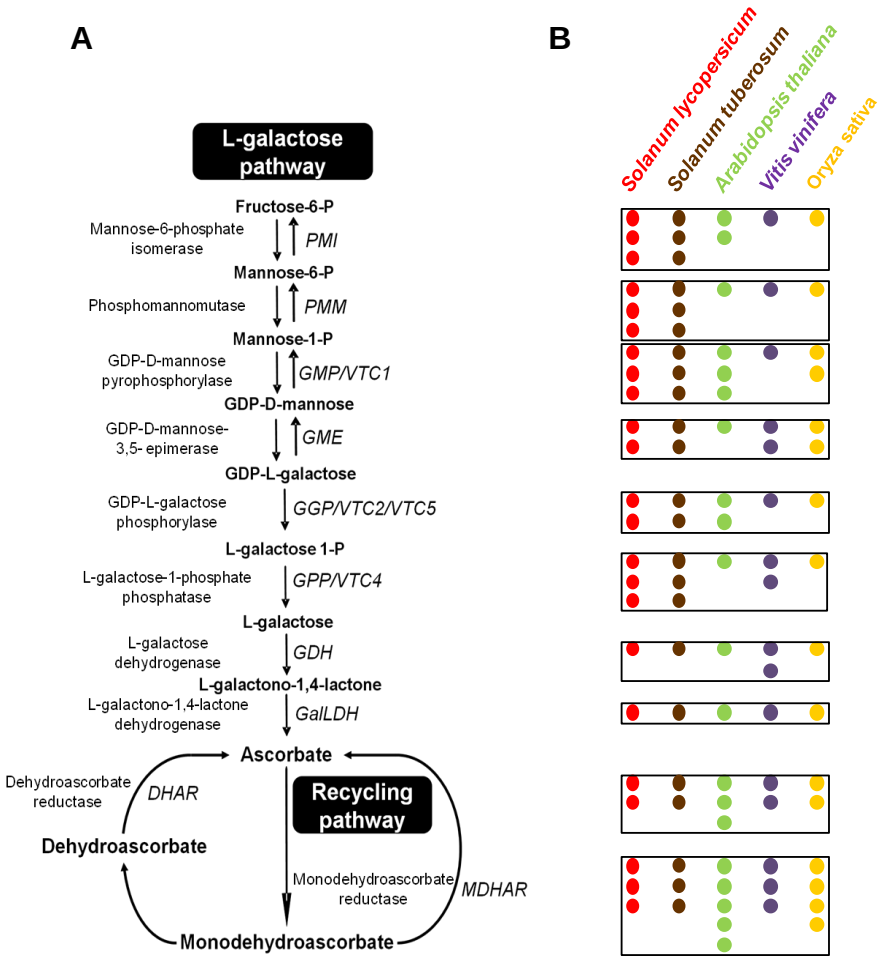
Supplementary Figure 49. GOslim, PO structural, and PO growth terms are differentially enriched in tomato and potato proximal/tandem duplications. Assuming no functional difference (in terms of GO/PO annotation) between tomato and potato genes belonging to each of the structural groups, we expect the ratios (tomato gene number/potato gene number) to be similar across functional groups (genomic distribution of functional groups are very similar in the two genomes, data not shown). In the panels the x axes are integer identifiers representing functional groups, while y axes are log ratio of gene numbers in each functional group (tomato/potato). Pseudocounts of 0.1 are added for original gene counts of 0. Blue lines plot the linear regression models. Green dashed lines mark 95% prediction interval. Red dots are outliers identified by Cook's distance with cutoff 4/data length.



Supplementary Figure 50. Scatter plot of omega values between *S. lycopersicum* (Sl) and *S. tuberosum* (St) and between *S. lycopersicum* (Sl) and *S. pimpinellifolium* (Sp) in 18,809 Sl-Sp-St ortholog groups. Red dots represent that Sl-Sp pairs of which omega values are larger than Sl-St pairs in the same ortholog group.



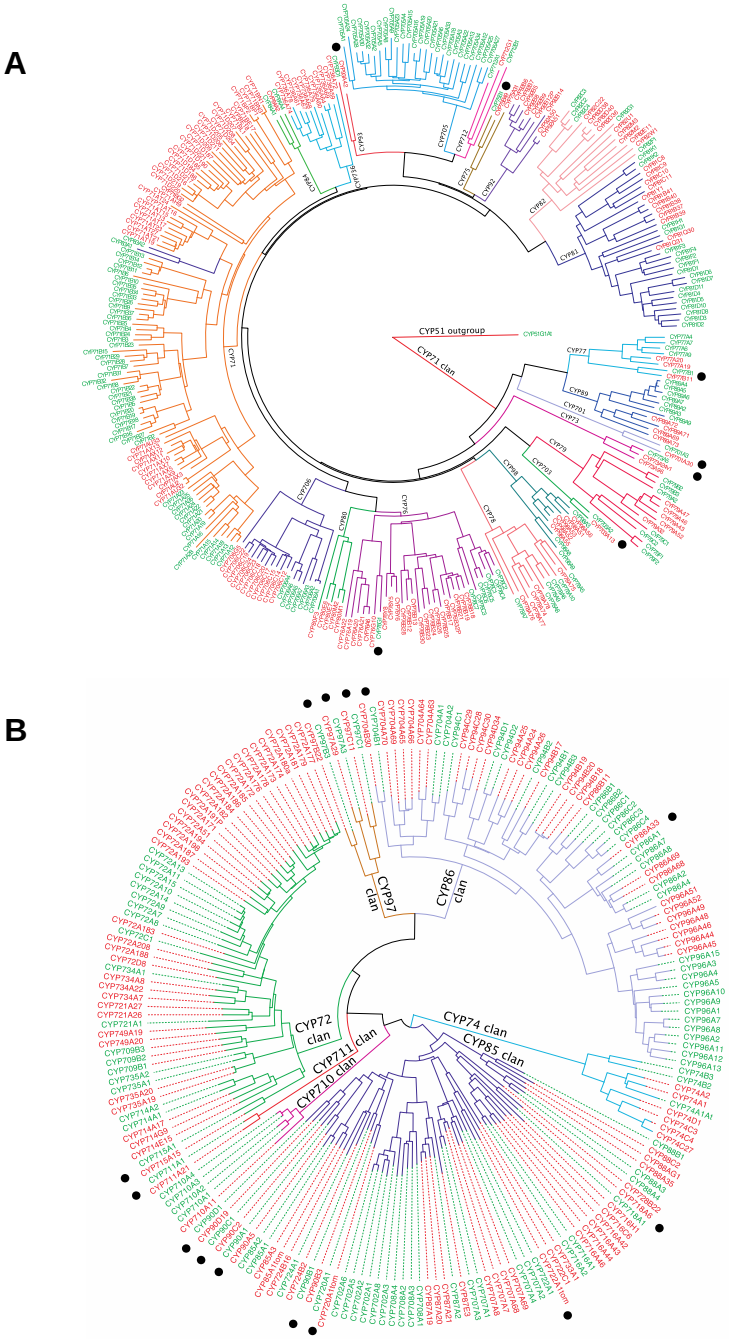
Supplementary Figure 51. Venn diagram of OrthoMCL group distribution in tomato, grape, *Arabidopsis* and rice. Numbers in the individual sections give the number of clusters/groups (not genes). The first number below the organism name marks the total number of genes of an organism that was used as an input for the software, the second number gives the number of genes that were found in clusters (the difference gives the number of singletons: genes that were not clustered at all).



Supplementary Figure 52. The L-galactose pathway of ascorbate biosynthesis and the recycling pathway in plants.

A. Biosynthetic and recycling pathways for L-ascorbic acid in plants[190]. Precursors from fructose-6-phosphate and oxidative forms of L-ascorbic acid (ascorbate or vitamin C) are shown with enzymes catalysing each step (left side). At the right side of the pathway, gene names coding for these enzymes have been determined for this study and gene names determined by Arabidopsis vtc mutations have been added. PMI, phosphomannose isomerase; PMM, phosphomannomutase; GMP, GDP-D-mannose pyrophosphorylase; GME, GDP-D-mannose 3′,5′-epimerase; GGP, GDP-L-galactose phosphorylase; GPP, L-galactose-1-P phosphatase; GDH, L-galactose dehydrogenase; GalLDH, L-galactono-1,4-lactone dehydrogenase; DHAR, dehydroascorbate reductase; MDHAR, monodehydroascorbate reductase.

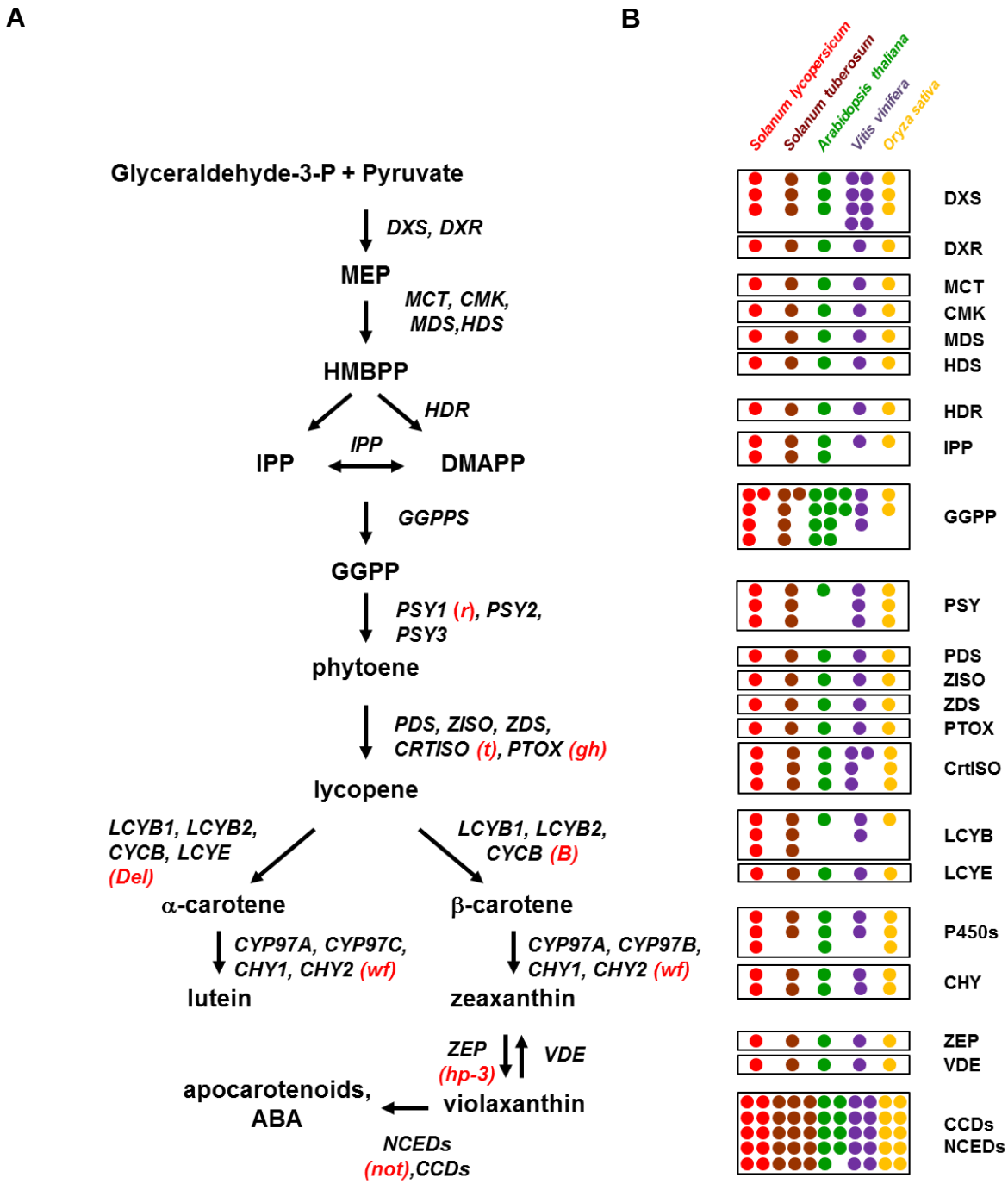
B. Orthologous proteins involved in vitamin C biosynthesis and recycling from tomato, potato, *Arabidopsis thaliana*, *Vitis vinifera* and *Oryza sativa* identified using the orthoMCL clusters (http://solgenomics.net/tools/genefamily/search.pl), in *Solanum lycopersicum* (red), *Solanum tuberosum* (brown), *Arabidopsis thaliana* (green), *Vitis vinifera* (purple) and *Oryza sativa* (yellow). Each circle represents one gene.



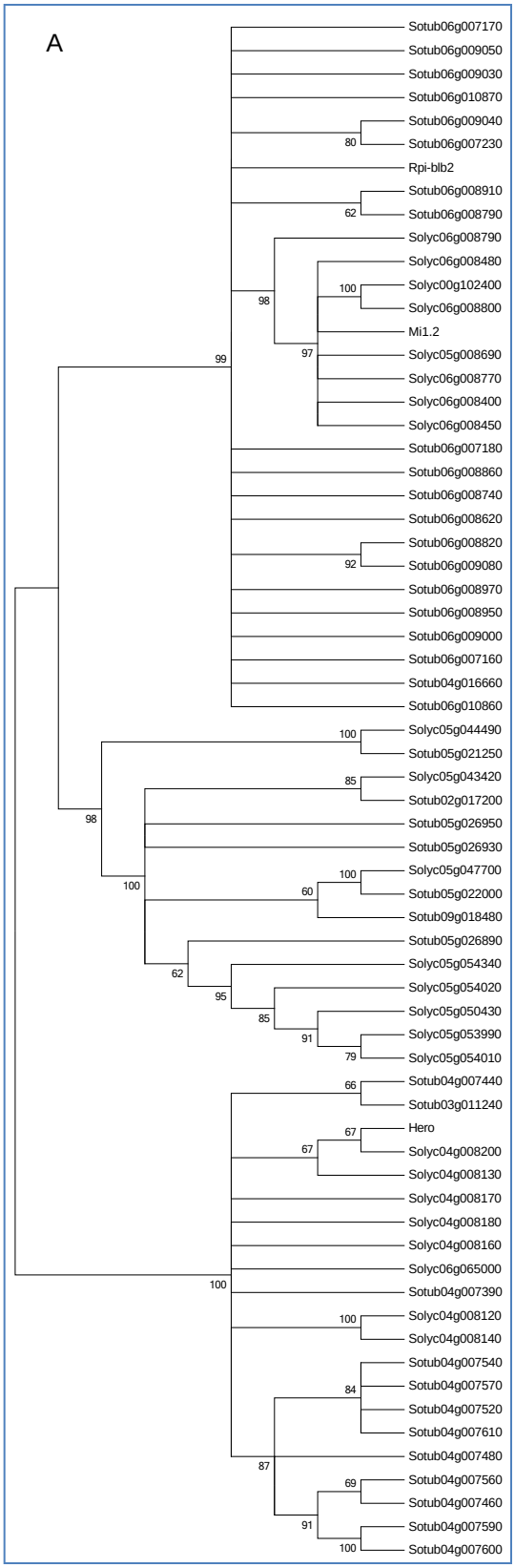
Supplementary Figure 53. The tomato cytochrome P450 family.

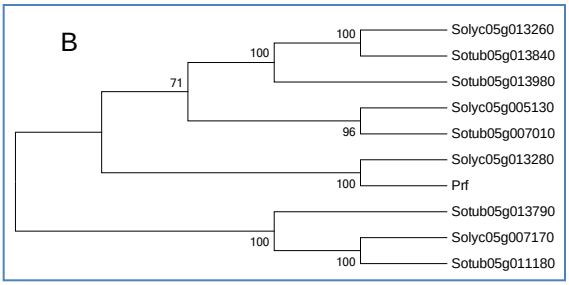
A. Neighbour-joining tree including 311 sequences representing the *Arabidopsis* and tomato CYP71 clan members, with CYP51 included as an outgroup. *Arabidopsis* names are in green and tomato in red. The 21 families are labeled and the branches are colored differently to make them easier to see.

B. Neighbour-joining tree with 202 sequences from the remaining seven CYP clans. Tomato and *Arabidopsis* are both lacking the CYP727 clan and CYP51 was included in Panel A. Black dots identify ortholog pairs. There are very few ortholog pairs due to the evolutionary distance between tomato (an asterid) and *Arabidopsis* (a rosid).This figure illustrates the size and diversity of the CYP72, CYP85 and CYP86 clans. These trees represent approximately 1% of the protein coding genes in these two species.

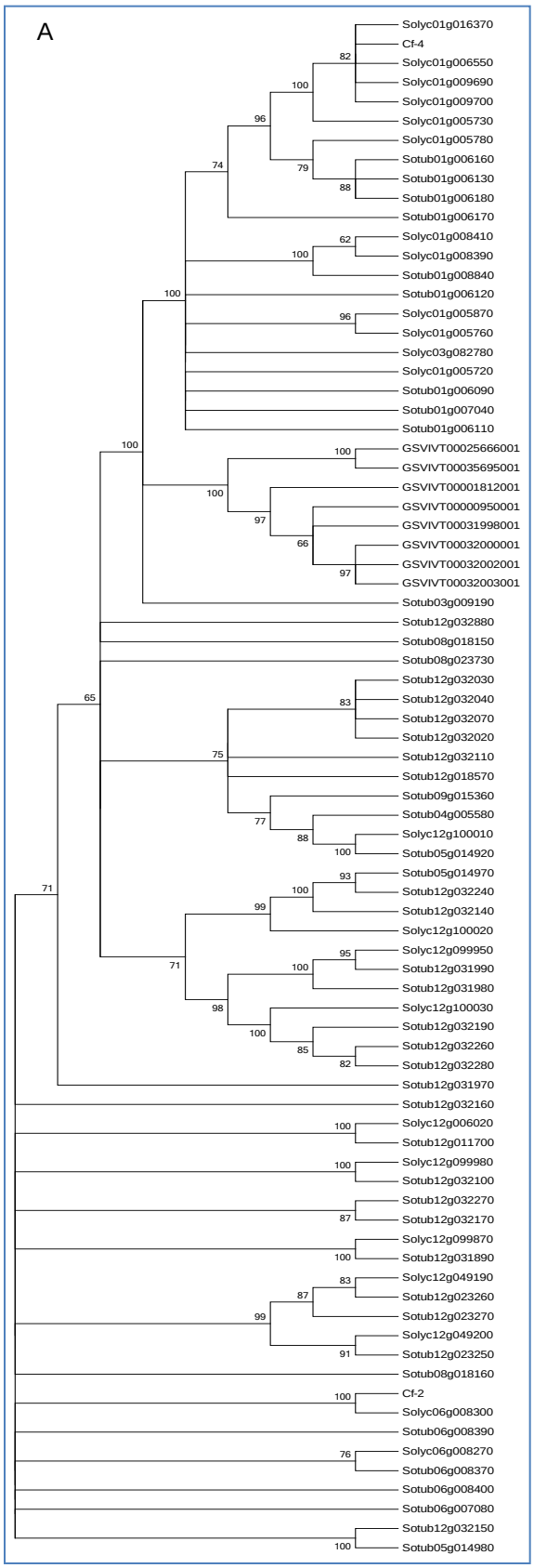


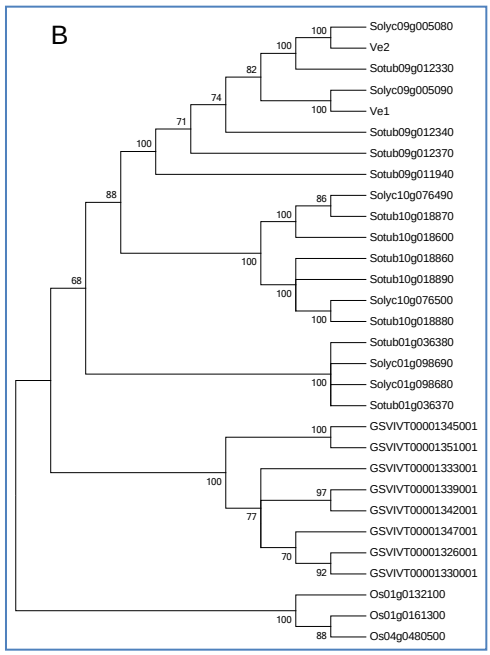
Supplementary Figure 54. Genes for the MEP/carotenoid pathway.
A. Simplified pathway. Mutant names are shown in red (see **Supplementary Table 14**).
B. Orthologous proteins involved in carotenoid biosynthesis from in *Solanum lycopersicum* (red), *Solanum tuberosum* (brown), *Arabidopsis thaliana* (green), *Vitis vinifera* (purple) and *Oryza sativa* (yellow) identified using the orthoMCL clusters (http://solgenomics.net/tools/genefamily/search.pl) and further manual curation. Each circle represents one gene.





Supplementary Figure 55. Phylogenetic analysis of tomato, potato, grape, rice and *Arabidopsis* CNL members belonging to the same orthoMCL group of: (A) Rpi-blb1, Mi 1-2 and Hero genes; (B) Prf genes. Bootstrap value is reported next to the branches. A different mutational rate was observed in these two families. (A) Three distinct *Solanum* groups were identified for this gene family. A clear expansion of potato members was observed. (B) Prf copies form almost perfect pairs in tomato and potato.





Supplementary Figure 56. Phylogenetic analysis of tomato, potato, grape, rice and *Arabidopsis* RLP members belonging to the same OrthoMCL group containing: (A) Cf2- Cf4 genes; (B) Ve1 and Ve2 genes; Bootstrap value is reported next to the branches. A) Cf genes were evidenced also in potato and grape; A gene number expansion was evidenced in potato. (B) Ve members were observed also in potato, grape and rice. A conspicuous number of copies was observed in potato (10) and grape (8).