The tomato genome sequence provides insights into fleshy fruit evolution

The Tomato Genome Consortium*

Tomato (Solanum lycopersicum) is a major crop plant and a model system for fruit development. Solanum is one of the largest angiosperm genera and includes annual and perennial plants from diverse habitats. Here we present a high-quality genome sequence of domesticated tomato, a draft sequence of its closest wild relative, Solanum pimpinellifolium, and compare them to each other and to the potato genome (Solanum tuberosum). The two tomato genomes show only 0.6% nucleotide divergence and signs of recent admixture, but show more than 8% divergence from potato, with nine large and several smaller inversions. In contrast to Arabidopsis, but similar to soybean, tomato and potato small RNAs map predominantly to gene-rich chromosomal regions, including gene promoters. The Solanum lineage has experienced two consecutive genome triplications: one that is ancient and shared with rosids, and a more recent one. These triplications set the stage for the neofunctionalization of genes controlling fruit characteristics, such as colour and fleshiness.

The genome of the inbred tomato cultivar ‘Heinz 1706’ was sequenced and assembled using a combination of Sanger and ‘next generation’ technologies (Supplementary Information section 1). The predicted genome size is approximately 900 megabases (Mb), consistent with previous estimates, of which 760 Mb were assembled in 91 scaffolds aligned to the 12 tomato chromosomes, with most gaps restricted to pericentromeric regions (Fig. 1A and Supplementary Fig. 1). Base accuracy is approximately one substitution error per 29.4 kilobases (kb) and one indel error per 6.4 kb. The scaffolds were linked with two bacterial artificial chromosome (BAC)-based physical maps and anchored/oriented using a high-density genetic map, intragenome and BAC fluorescence in situ hybridization (FISH).

The genome of S. pimpinellifolium LA1589 was sequenced and assembled de novo using Illumina short reads, yielding a 739 Mb draft genome (Supplementary Information section 3). Estimated divergence between the wild and domesticated genomes is 0.6% (5.4 million single nucleotide polymorphisms (SNPs) distributed along the chromosomes (Fig. 1A and Supplementary Fig. 1)).

Tomato chromosomes consist of pericentric heterochromatin and distal euchromatin, with repeats concentrated within and around centromeres, in chromomereres and at telomeres (Fig. 1A and Supplementary Fig. 1). Substantially higher densities of recombination, genes and transcripts are observed in euchromatin, whereas chloroplast insertions (Supplementary Information sections 1.22 and 1.23) and conserved microRNA (miRNA) genes (miRNA Information section 2.9) are more evenly distributed throughout the genome. The genome is highly syntenic with those of other economically important Solanaceae (Fig. 1B). Compared to the genomes of Arabidopsis* and Sorghum*, tomato has fewer high-copy, full-length long terminal repeat (LTR) retrotransposons with older average insertion ages (2.8 versus 0.8 million years (Myr) ago) and fewer high-frequency k-mers (Supplementary Information section 2.10). This supports previous findings that the tomato genome is unusual among angiosperms by being largely comprised of low-copy DNA.

The pipeline used to annotate the tomato and potato* genomes is described in Supplementary Information section 2. It predicted 34,727 and 35,004 protein-coding genes, respectively. Of these, 30,855 and 32,988, respectively, are supported by RNA sequencing (RNA-Seq) data, and 31,741 and 32,056, respectively, show high similarity to Arabidopsis genes (Supplementary Information section 2.1). Chromosomal organization of genes, transcripts, repeats and small RNAs (sRNAs) is very similar in the two species (Supplementary Figs 2–4). The protein-coding genes of tomato, potato, Arabidopsis, rice and grape were clustered into 23,208 gene groups (≥2 members), of which 8,615 are common to all five genomes, 1,727 are confined to eudicots (tomato, potato, grape and Arabidopsis), and 727 are confined to plants with fleshy fruits (tomato, potato and grape) (Supplementary Information section 5.1 and Supplementary Fig. 5). Relative expression of all tomato genes was determined by replicated strand-specific Illumina RNA-Seq of root, leaf, flower (two stages) and fruit (six stages) in addition to leaf and fruit (three stages) of S. pimpinellifolium (Supplementary Table 1).

sRNA sequencing data supported the prediction of 96 conserved miRNA genes in tomato and 120 in potato, a number consistent with other plant species (Fig. 1A, Supplementary Figs 1 and 3 and Supplementary Information section 2.9). Among the 34 miRNA families identified, 10 are highly conserved in plants and similarly represented in the two species, whereas other, less conserved families are more abundant in potato. Several miRNAs, predicted to target Toll interleukin receptor, nucleotide-binding site and leucine-rich repeat (TIR-NBS-LRR) genes, seemed to be preferentially or exclusively expressed in potato (Supplementary Information section 2.9).

Comparative genomic studies are reported in Supplementary Information section 4. Sequence alignment of 71 Mb of euchromatic tomato genomic DNA to their potato counterparts revealed 8.7% nucleotide divergence (Supplementary Information section 4.1). Intergenic and repeat-rich heterochromatic sequences showed more than 30% nucleotide divergence, consistent with the high sequence diversity in these regions among potato genotypes. Alignment of tomato–potato orthologous regions confirmed nine large inversions known from cytological or genetic studies and several smaller ones (Fig. 1C). The exact number of small inversions is difficult to determine due to the lack of orientation of most potato scaffolds.

A total of 18,320 clearly orthologous tomato–potato gene pairs were identified. Of these, 138 (0.75%) had significantly higher than average non-synonymous (Ks) versus synonymous (Ks) nucleotide substitution rate ratios (ωs), indicating diversifying selection, whereas 147 (0.80%) had significantly lower than average ωs, indicating purifying selection (Supplementary Table 2). The proportions of high and low ωs between sorghum and maize (Zea mays) are 0.70% and 1.19%, respectively, after 11.9 Myr of divergence, indicating that diversifying selection may have been stronger in tomato–potato. The highest densities of low-ωs genes are found in collinear blocks with average Ks > 1.5, tracing to a genome triplication shared with grape (see below) (Fig. 1C, Supplementary Fig. 6 and Supplementary Table 3). These genes, which have been preserved in paleo-duplicated locations for more than 100 Myr, are more constrained than ‘average’ genes and are

*Lists of participants and their affiliations appear at the end of the paper.
Figure 1 | Tomato genome topography and synteny. A. Multi-dimensional topography of tomato chromosome 1 (chromosomes 2–12 are shown in Supplementary Fig. 1). a, Left: contrast-reversed, 4',6-diamidino-2-phenylindole (DAPI)-stained pachytene chromosome; centre and right: FISH signals for repeat sequences on diagrammatic pachytene chromosomes (purple, TGR1; blue, TGR4; red, telomere repeat; green, Cot 100 DNA (including most repeats)). b, Frequency distribution of recombination nodules (RNs) representing crossovers on 249 chromosomes. Red stars mark 5 cM intervals starting from the end of the short arm (top). Scale is in micrometres. c, FISH-based locations of selected BACs (horizontal blue lines on left). d, Kazusa F2-2000 linkage map. Blue lines to the left connect linkage map markers on the BAC-FISH map (e), and to the right to heat maps (e) and the DNA pseudomolecule (f). e, f, From left to right: linkage map distance (cM/Mb, turquoise), repeated sequences (% nucleotides per 500 kb, purple), genes (% nucleotides per 500 kb, blue), chloroplast insertions; RNA-Seq reads from leaves and breaker fruits of S. lycopersicum and S. pimpinellifolium (number of reads per 500 kb, green and red, respectively), microRNA genes (transcripts per million per 500 kb, black), small RNAs (thin horizontal black and red lines, sum of hits-normalized abundances). Horizontal grey lines represent gaps in the pseudomolecule (f), f, DNA pseudomolecule consisting of nine scaffolds. Unsequenced gaps (approximately 9.8 Mb, Supplementary Table 13) are enriched for transcription factors and genes otherwise related to gene regulation (Supplementary Tables 3 and 4).

Sequence comparison of 31,760 Heinz 1706 genes with >5× S. pimpinellifolium read coverage in over 90% of their coding regions indicated by white horizontal lines. Tomato genes identified by map-based cloning (Supplementary Table 14) are indicated on the right. For more details, see legend to Supplementary Fig. 1. B, Syntenic relationships in the Solanaceae. COSII-based comparative maps of potato, aubergine (eggplant), pepper and Nicotiana with respect to the tomato genome (Supplementary Information section 4.5 and Supplementary Fig. 14). Each tomato chromosome is assigned a different colour and orthologous chromosome segment(s) in other species are shown in the same colour. White dots indicate approximate centromere locations. Each black arrow indicates an inversion relative to tomato and ‘+1‘ indicates a minimum of one inversion. Each black bar beside a chromosome indicates translocation breakpoints relative to tomato. Chromosome lengths are not to scale, but segments within chromosomes are. C, Tomato–potato syntenic relationships dot plot of tomato (T) and potato (P) genomic sequences based on collinear blocks (Supplementary Information section 4.1). Red and blue dots represent gene pairs with statistically significant high and low co (Kv/Ks) in collinear blocks, which average Kv ≤ 0.5, respectively. Green and magenta dots represent genes in collinear blocks which average 0.5 < Kv ≤ 1.5 and Ks > 1.5, respectively. Yellow dots represent all other gene pairs. Blocks circled in red are examples of pan-eudicot triplication. Inserts represent schematic drawings of BAC-FISH patterns of cytologically demonstrated chromosome inversions (also in Supplementary Fig. 15). revealed 7,378 identical genes and 11,753 with only synonymous changes. The remaining 12,629 genes had non-synonymous changes, including gains and losses of stop codons with potential consequences for gene function (Supplementary Tables 5–7). Several pericentric
regions, predicted to contain genes, are absent or polymorphic in the broader *S. pimpinellifolium* germplasm (Supplementary Table 8 and Supplementary Fig. 7). Within cultivated germplasm, particularly among the small-fruited cherry tomatoes, several chromosomal segments are more closely related to *S. pimpinellifolium* than to Heinz 1706 (Supplementary Figs 8 and 9), supporting previous observations on recent admixture of these gene pools due to breeding\(^1\). Heinz 1706 itself has been reported to carry introgressions from *S. pimpinellifolium\(^3\), traces of which are detectable on chromosomes 4, 9, 11 and 12 (Supplementary Table 9).

Comparison of the tomato and grape genomes supports the hypothesis that a whole-genome triplication affecting the rosid lineage occurred in a common eudicot ancestor\(^5\) (Fig. 2a). The distribution of \(K_x\) between corresponding gene pairs in duplicated blocks suggests that one polyploidization in the solanaceae lineage preceded the rosid–asterid (tomato–grape) divergence (Supplementary Fig. 10).

Comparison with the grape genome also reveals a more recent triplication in tomato and potato. Whereas few individual tomato/ potato gene models remain triplicated (Supplementary Tables 10 and 11), 73% of tomato gene models are in blocks that are orthologous to one grape region, collectively covering 84% of the grape gene space. Among these grape genomic regions, 22.5% have one orthologous region in tomato, 39.9% have two, and 21.6% have three, indicating that a whole-genome triplication occurred in the *Solanum* lineage, followed by widespread gene loss. This triplication, also evident in tomato, 39.9% have two, and 21.6% have three, indicating that a whole-genome triplication occurred in the *Solanum* lineage, followed by widespread gene loss. This triplication, also evident in tomato and potato. Whereas few individual tomato/ potato gene models remain triplicated (Supplementary Tables 10 and 11), 73% of tomato gene models are in blocks that are orthologous to one grape region, collectively covering 84% of the grape gene space. Among these grape genomic regions, 22.5% have one orthologous region in tomato, 39.9% have two, and 21.6% have three, indicating that a whole-genome triplication occurred in the *Solanum* lineage, followed by widespread gene loss. This triplication, also evident in tomato and potato.

The number of euasterid lineages that have experienced the explosive radiation of angiosperm plants on Earth\(^15\). This radiation is apparent for 2,687 promoters, including those of cell-wall-related functions (Fig. 3). These include transcription factors and enzymes necessary for ethylene biosynthesis (*RIN, CNR, ACS*) and perception (*ETR3/NR, ETR4\(^17\), red light photoreceptors influencing fruit quality (PHYB1/PHYB2) and ethylene- and light-regulated genes mediating lycopene biosynthesis (PSY1/PSY2). Several cytochrome P450 subfamilies associated with toxic alkaloid biosynthesis show contraction or complete loss in tomato and the extant genes show negligible expression in ripe fruits (Supplementary Information section 5.4).

Fruit texture has profound agronomic and sensory importance and is controlled in part by cell wall structure and composition\(^14\). More than 50 genes showing differential expression during fruit development and ripening encode proteins involved in modification of cell wall architecture (Fig. 4a and Supplementary Information section 5.7). For example, a family of xyloglucan endotransglycosylases/hydrolases (XTHs) has expanded both in the recent whole-genome triplication and through tandem duplication. One of the triplicated members, XTH10, shows differential loss between tomato and potato (Fig. 4a and Supplementary Table 12), suggesting genetically driven specialization in the remodelling of fruit cell walls.

Similar to soybean and potato and in contrast to *Arabidopsis*, tomato sRNAs map preferentially to euchromatin (Supplementary Fig. 2). sRNAs from tomato flowers and fruits\(^19\) map to 8,416 gene promoters. Differential expression of sRNAs during fruit development is apparent for 2,687 promoters, including those of cell-wall-related genes (Fig. 4b) and occurs preferentially at key developmental transitions (for example, flower to fruit, fruit growth to fruit ripening, Supplementary Information section 2.8).

The genome sequences of tomato, *S. pimpinellifolium* and potato provide a starting point for comparing gene family evolution and sub-functionalization in the *Solanaceae*. A striking example is the *SELF PRUNING* (*SP*) gene family, which includes the homologue of *Arabidopsis FT*, encoding the mobile flowering signal *florigen*\(^20\) and its antagonist *SP*, encoding the orthologue of *TFL1*. Nearly a century ago, a spontaneous mutation in *SP* spawned the ‘determinate’ varieties.

Fleshy fruits (Supplementary Fig. 12) are an important means of attracting vertebrate frugivores for seed dispersal\(^16\). Combined orthology and synteny analyses indicate that both genome triplications added new gene family members that mediate important fruit-specific functions (Fig. 3). These include transcription factors and enzymes necessary for ethylene biosynthesis (*RIN, CNR, ACS*) and perception (*ETR3/NR, ETR4\(^17\), red light photoreceptors influencing fruit quality (PHYB1/PHYB2) and ethylene- and light-regulated genes mediating lycopene biosynthesis (PSY1/PSY2). Several cytochrome P450 subfamilies associated with toxic alkaloid biosynthesis show contraction or complete loss in tomato and the extant genes show negligible expression in ripe fruits (Supplementary Information section 5.4).

Figure 2 | The *Solanum* whole genome triplication. a Speciation and polyploidization in eudicot lineages. Confirmed whole-genome duplications and triplications are shown with annotated circles, including ‘T’ (this paper) and previously discovered events ‘a’, ‘b’, ‘c’, ‘d’, ‘e’. Dashed circles represent one or more suspected polyploids reported in previous publications that need further support from genome assemblies\(^22–23\). Grey branches indicate unpublished genomes. Black and red error bars bracket indicate the likely timings of divergence of major asterid lineages and of ‘T’, respectively. The post-‘T’ subgenomes, designated T1, T2, and T3, are further detailed in Supplementary Fig. 10. b On the basis of alignments of multiple tomato genome segments to single grape genome segments, the tomato genome is partitioned into three non-overlapping ‘subgenomes’ (T1, T2, T3), each represented by one axis in the three-dimensional plot. The ancestral gene order of each subgenome is inferred according to orthologous grape regions, with tomato chromosomal affinities shown by red (inner) bars. Segments tracing to pan-eudicot triplication (‘γ’) are shown by green (outer) bars with colours representing the seven putative pre-‘γ’ eudicot ancestral chromosomes\(^29\), also coded a–g.

©2012 Macmillan Publishers Limited. All rights reserved
that now dominate the tomato mechanical harvesting industry. The genome sequence has revealed that the SP family has expanded in the Solanum lineage compared to Arabidopsis, driven by the Solanum triplication and tandem duplication (Supplementary Fig. 13). In potato, Solanum families are shown in Supplementary Figs 16 and 17.

that now dominate the tomato mechanical harvesting industry. The genome sequence has revealed that the SP family has expanded in the Solanum lineage compared to Arabidopsis, driven by the Solanum triplication and tandem duplication (Supplementary Fig. 13). In potato, Solanum families are shown in Supplementary Figs 16 and 17.

Figure 3 | Whole-genome triplications set the stage for fruit-specific gene neo-functionalization. The genes shown represent a fruit ripening control network regulated by transcription factors (MADS-RIN, CNR) necessary for production of the ripening hormone ethylene, the production of which is regulated by ACC synthase (ACS). Ethylene interacts with ethylene receptors (ETRs) to drive expression changes in output genes, including phytoene synthase (PSY), the rate-limiting step in carotenoid biosynthesis. Light, acting through phytochromes, controls fruit pigmentation through an ethylene-independent pathway. Paralogous gene pairs with different physiological roles (MADS1/RIN, PHYB1/PHYB2, ACS2/ACS6, ETR3/ETR4, PSY1/PSY2) were generated during the eudicot (T, red) or the more recent Solanum (T, red) triplications. Complete dendrograms of the respective protein families are shown in Supplementary Figs 16 and 17.

The genome sequences of tomato and S. pimpinellifolium - Aroma - Texture - Pathogen susceptibility

Figure 4 | The tomato genome allows systems approaches to fruit biology.

a. Xyloglucan transglucosylase/hydrolases (XTHs) differentially expressed between mature green and ripe fruits (Supplementary Information section 5.7). These XTH genes and many others are expressed in ripening fruits and are linked with the Solanum triplication, marked with a red circle on the phylogenetic tree. Red lines on the tree denote paralogues derived from the Solanum triplication, and blue lines are tandem duplications.

b. Developmentally regulated accumulation of sRNAs mapping to the promoter region of a fruit-regulated cell wall gene (pectin acetyl-esterase, Solyc08g005800). Variation of abundance of sRNAs (left) and messenger RNA expression levels from the corresponding gene (right) over a tomato fruit developmental series (T1, bud; T2, Bower; T3, fruit 1–3 mm; T4, fruit 5–7 mm; T5, fruit 11–13 mm; T6, fruit mature green; T7, breaker; T8, breaker + 3 days; T9, breaker + 7 days). The promoter regions are grouped in 100-nucleotide windows. For each window the size class distribution of sRNAs is shown (red, 21; green, 22; orange, 23; blue, 24). The height of the box corresponding to the first time point shows the cumulative sRNA abundance in log scale. The height of the following boxes is proportional to the log offset fold change (offset = 20) relative to the first time point. The expression profile of the mRNA is shown in log2 scale. The horizontal black line represents 1 kb of the promoter region. 0 to 12 represent arbitrary units of gene expression.

the 16th century, and the intensive breeding that followed. Charles Rick pioneered the use of trait introgression from wild tomato relatives to increase genetic diversity of cultivated tomatoes. Introduction lines exist for seven wild tomato species, including S. pimpinellifolium, in the background of cultivated tomato. The genome sequences presented here and the availability of millions of SNPs will allow breeders to revisit this rich trait reservoir and identify domestication genes, providing biological knowledge and empowering biodiversity-based breeding.

METHODS SUMMARY

A total of 21 gigabases (Gb) of Roche/454 Titanium shotgun and mate pair reads and 3.3 Gb of Sanger paired-end reads, including ~200,000 BAC and fosmid end sequence pairs, were generated from the ‘Heinz 1706’ inbred line (Supplementary Information sections 1.1–1.7), assembled using both Newbler and CABOG and integrated into a single assembly (Supplementary Information sections 1.17 and 1.18). The scaffolds were anchored using two BAC-based physical maps, one high density genetic map, overgo hybridization and genome-wide BAC FISH (Supplementary Information sections 1.8–1.16 and 1.19). Over 99% of BAC/fosmid end pairs mapped consistently on the assembly and over 98% of EST sequences could be aligned to the assembly (Supplementary Information section 1.20). Chloroplast genome insertions in the nuclear genome were validated using a mate pair method and the flanking regions were identified (Supplementary Information sections 1.22–1.24). Annotation was carried out using a pipeline based on EuGene that integrates de novo gene prediction, RNA-Seq alignment and rich function annotation (Supplementary Information section 2). To facilitate interspecies comparison, the potato genome was re-annotated using the same pipeline. LTR retrotransposons were detected de novo with the LTR-STRUC program and dated by the sequence divergence between left and right solo LTR (Supplementary Information section 2.10). The genome of S. pimpinellifolium was sequenced to 40X depth using Illumina paired-end reads and assembled using ABYSS (Supplementary Information section 3). The tomato and potato genomes were aligned using LASTZ (Supplementary Information section 4.1). Identification of triplicated regions was done using BLASTP, in-house-generated scripts and three-way comparisons between tomato, potato and S. pimpinellifolium using MCSCAN (Supplementary Information sections 4.2–4.4). Specific gene families/
groups (genes for ascorbate, carotenoid and jasmonate biosynthesis, cytochrome P450s, genes controlling cell wall architecture, hormonal and transcriptional regulators, resistance genes) were subjected to expert curation/analysis (Supplementary Information section 5). PHYML and MEGA were used to reconstruct phylogenetic trees and MCSCAN was used to infer gene collinearity (Supplementary Information section 5.2).

Received 17 August 2011; accepted 3 April 2012.

1. Frodin, D. G. History and concepts of big plant genera. Taxon 53, 753–776 (2004).

2. Peratta, I. E., Spooner, D. M. & Knapp, S. Taxonomy of tomatoes: a revision of wild tomatoes (Solanum lycopersicum) and their outgroup relatives in sections Juglandifolia and Lycopersicoides. Syst. Bot. Monogr. 84, 1–186 (2008).

3. Michaelson, M. J., Price, H. J., Ellison, J. R. & Johnston, J. S. Comparison of plant DNA contents determined by Feulgen microspectrophotometry and laser flow cytometry. J. Hered. 78, 193–198 (1987).

4. The Arabidopsis Genome Initiative. The Arabidopsis Genome Initiative reveals the complexity of the Arabidopsis thaliana genome. Nature 408, 796–815 (2000).

5. Paterson, A. H. et al. The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–556 (2009).

6. Zamir, D. & Tanksley, S. D. Tomato genome is comprised largely of fast-evolving, low copy-number sequences. Synteny and collinearity in plant genomes. Science 320, 486–488 (2008).

7. 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).

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

9. Swigonˇova´, Z. et al. Flowering plant cytometry. Am. J. Bot. 88, 1916–1923 (2001).

10. Stockey, R. A., Graham, S. W. & Crane, P. R. Introduction to the Darwin special issue: Rep. Tomato Genet. Coop. 189–195 (2001).

11. Howe, H. F. & Smallwood, J. Ecology of seed dispersal. Annu. Rev. Ecol. Syst. 14, 201–228 (1983).

12. Navarro, C. Self-pruning drives heterosis for yield in tomato. Nature Genet. 42, 459–463 (2010).

13. Jelion, O. et al. The grapevine genome sequence reveals ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467 (2007).

14. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic relationships of flowering plants: a combined analysis of the gene sequences as numbers CM0001064–CM0001705. The RNA-Seq data are available in the Sequence Read Archive under accession number SRAM9915, GSE33507, SRA050797, and SRA048114. Further information on data access can be found in Supplementary Information section 2.2. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Readers are welcome to comment on the online version of this paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D. Zamir (dizamir@agri.huji.ac.il) or G. Giovanngiliano (giovanni@enea.it).

The Tomato Genome Consortium

Kazusa DNA Research Institute Shusei Sato (Principal Investigator), Satoshi Tabata (Principal Investigator), Hidemi Harakawa1, Erka Asai2, Kenta Sugasawa3, Sachiko Isobe4, Takazaka Kaneko5, Yasuakza Nakamura6, Daisuke Shibata7, Koh Aoki8, 454 Life Science, a company Michael Egholm9, James Knight10, Amplicon Express11, Bogden12, Beiji13, Liuhua Yan13, Hongling Jiang13, Bao Wang13, Hongshuang Li13, Zhenjun Li13, Yongyou Fu14, Minyun Huang8, Jonathan A. Leto8, Naama Menda8, Susan Strickler8, Willem J. Stiekema9, Antonio J. Matas20, Ryan McQuinn20, Steven D. Tanksley (Principal Investigator)18, Zachary B. Lippman19, W. Richard McCombie (Principal Investigator)17, Doreen Ware (Principal Investigator)16, Yves Van de Peer (Principal Investigator)22; Yves Van de Peer (Principal Investigator)22; Jianchang Gao12, Yanmei Guo12, Xiaoxuan Wang12, Ying Li12, Jun He12; Truck-It EXPRESS; C. R. Buell from the NSF-funded Potato Genome Sequence and Annotation project; construction of the Whole Genome Profiling (WGP) method was supported by EnzaZaden, RijkZwaan, Vilmorin & Cie, and Takii & Co. Keygene N.V. owns patents and patent applications covering its AFLP and Whole Genome Profiling technologies; AFLP and Keygene are registered trademarks of Keygene N.V. The following individuals are also acknowledged for their contribution to the work described: J. Park, B. Wang, C. Niu, D. Liu, F. Cojotti, S. Pescarolo, A. Zambon, G. Xiao, J. Chen, J. Shi, L. Zhang, Z. M. Caccamo, D. Bolser, D. Martin, M. Gonzalez, P. A. Bedinger, P. A. Covey, P. Farchi, R. R. Pousada, S. Hakim, S. Elyashiv, B. Lu, W. Long, X. Tan, C. Du, X. Liu, D. Martin, R. Lizardal, E. M. Ya and Z. Li. M. Zabeau is thanked for his support and encouragement and S. van den Brink for her secretarial support. We dedicate this work to the late C. Rick who pioneered tomato genetics, collection of wild germplasm and the distribution of seed and knowledge.

Author Contributions For full details of author contributions, please see the Supplementary Information.

Author Information The genomic data generated by the project are available in GenBank as accession numbers CM0001064–CM0001705, and the genome sequences as numbers CM0001064–CM0001705. The RNA-Seq data are available in the Sequence Read Archive under accession number SRAM9915, GSE33507, SRA050797, and SRA048114. Further information on data access can be found in Supplementary Information section 2.2. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Readers are welcome to comment on the online version of this paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D. Zamir (dizamir@agri.huji.ac.il) or G. Giovanngiliano (giovanni@enea.it).

This project has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 602773 (EU-RTG GenoPOM). This work was supported by: Argentina: INTA and CONICET. Belgium: Flemish Institute for Biodiversity and Ghent University. China: The State Key Laboratory of Developmental Genetics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences; Ministry of Science and Technology (2006A10A111, 2004CB720405, 2006CB1101970, 2007DFB30080) Ministry of Agriculture (9348 Program: 2007-25); National Natural Science Foundation (36171319); Postdoctoral Foundation (2007042044); European Union: FP6 Integrated Project EU-SOL PL 016214. France: Institute National de la Recherche Agronomique and Agence Nationale de la Recherche. Germany: the Max Planck Society. India: Department of

©2012 Macmillan Publishers Limited. All rights reserved
