the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **80,** 136–148 doi: 10.1111/tpj.12616

RESOURCE

Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing

The 100 Tomato Genome Sequencing Consortium, Saulo Aflitos[1,2], Elio Schijlen[1], Hans de Jong[3], Dick de Ridder[2], Sandra Smit[2], Richard Finkers[4], Jun Wang[5], Gengyun Zhang[6], Ning Li[7], Likai Mao[5], Freek Bakker[8], Rob Dirks[9], Timo Breit[10], Barbara Gravendeel[11], Henk Huits[12], Darush Struss[13], Ruth Swanson-Wagner[14], Hans van Leeuwen[15], Roeland C.H.J. van Ham[16], Laia Fito[17], Laëtitia Guignier[18], Myrna Sevilla[19], Philippe Ellul[20], Eric Ganko[21], Arvind Kapur[22], Emannuel Reclus[23], Bernard de Geus[24], Henri van de Geest[1], Bas te Lintel Hekkert[1], Jan van Haarst[1], Lars Smits[1], Andries Koops[1], Gabino Sanchez-Perez[1], Adriaan W. van Heusden[3], Richard Visser[3], Zhiwu Quan[5], Jiumeng Min[7], Li Liao[7], Xiaoli Wang[7], Guangbiao Wang[7], Zhen Yue[7], Xinhua Yang[7], Na Xu[7], Eric Schranz[8], Erik Smets[9], Rutger Vos[9], Johan Rauwerda[10], Remco Ursem[11], Cees Schuit[12], Mike Kerns[14], Jan van den Berg[15], Wim Vriezen[15], Antoine Janssen[16], Erwin Datema[16], Torben Jahrman[17], Frederic Moquet[18], Julien Bonnet[21] and Sander Peters[1]*

[1]*Plant Research International, Business Unit of Bioscience, Cluster Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands,*
[2]*Bioinformatics, Department of Plant Sciences, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands,*
[3]*Laboratory of Genetics, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands,*
[4]*Laboratory of Plant Breeding, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands,*
[5]*Beijing Genomics Institute, Shenzhen 518083, China,*
[6]*State Key Laboratory of Agricultural Genomics, Beijing Genomics Institute, Shenzhen 518083, China,*
[7]*BGI Europe, DK-2200 Copenhagen, Denmark,*
[8]*Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands,*
[9]*Department of R&D Bioinformatics, Rijk Zwaan Breeding BV, PO Box 40, 2678 ZG De, Lier, The Netherlands,*
[10]*RNA Biology & Applied Bioinformatics, Faculty of Science, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands,*
[11]*Netherlands Biodiversity Center Naturalis, Darwinweg 2, 2333 CR, Leiden, The Netherlands,*
[12]*Bejo Zaden BV, Research Centre, Trambaan 2A, 1749 CZ, Warmenhuizen, The Netherlands,*
[13]*East West Seed, Hortigenetics Research, Chiang Mai-Praw Road, Sansai, Chiang Mai 50290, Thailand,*
[14]*Monsanto Holland BV, Leeuwenhoekseweg 52, 2660 BB, Bergschenhoek, The Netherlands,*
[15]*Nunhems Netherlands BV, PO Box 4005, 6080 AA, Haelen, The Netherlands,*
[16]*Keygene NV, PO Box 216, 6700 AE, Wageningen, The Netherlands,*
[17]*Semillas Fito, Cetro de Biotecnologia, Riera d'Agell 11, 08349 Cabrera de Mar, Spain,*
[18]*Gautier Semences, Route d'Avignon, 13630 Eyragues, France,*
[19]*BHN Research, PO Box 3267, Immokalee, FL 34143, USA,*
[20]*Consultative Group on International Agricultural Research, Avenue Agropolis, 34394 Montpellier, France,*
[21]*Syngenta Biotechnology Inc., 3054 East Cornwallis Road, PO Box 12257, Research Triangle Park, NC 27709–2257, USA,*
[22]*Rasi Seeds, 273 Kamarajanar Road, Attur 636* 102, Salem District, Tamilnadu, India,
[23]*ADNid Cap Alpha, Avenue de l'Europe, 34830 Clapiers, France, and*
[24]*Technologisch Top Instituut Groene Genetica, Vossenburchkade* 68A, 2805 PC, Gouda, The Netherlands

Received 24 January 2014; revised 23 June 2014; accepted 1 July 2014; published online 12 July 2014.
*For correspondence (e-mail sander.peters@wur.nl).

SUMMARY

We explored genetic variation by sequencing a selection of 84 tomato accessions and related wild species representative of the *Lycopersicon*, *Arcanum*, *Eriopersicon* and *Neolycopersicon* groups, which has yielded a huge amount of precious data on sequence diversity in the tomato clade. Three new reference genomes were reconstructed to support our comparative genome analyses. Comparative sequence alignment revealed group-, species- and accession-specific polymorphisms, explaining characteristic fruit traits and growth habits in the various cultivars. Using gene models from the annotated Heinz 1706 reference genome, we observed differences in the ratio between non-synonymous and synonymous SNPs (dN/dS) in fruit diversification and plant growth genes compared to a random set of genes, indicating positive selection and differences in selection pressure between crop accessions and wild species. In wild species, the number of single-nucleotide polymorphisms (SNPs) exceeds 10 million, i.e. 20-fold higher than found in most of the crop accessions, indicating dramatic genetic erosion of crop and heirloom tomatoes. In addition, the highest levels of heterozygosity were found for allogamous self-incompatible wild species, while facultative and autogamous self-compatible species display a lower heterozygosity level. Using whole-genome SNP information for maximum-likelihood analysis, we achieved complete tree resolution, whereas maximum-likelihood trees based on SNPs from ten fruit and growth genes show incomplete resolution for the crop accessions, partly due to the effect of heterozygous SNPs. Finally, results suggest that phylogenetic relationships are correlated with habitat, indicating the occurrence of geographical races within these groups, which is of practical importance for *Solanum* genome evolution studies.

Keywords: *Solanum* genetic diversity, comparative sequence analysis, phylogenomics, introgression, chromosome evolution, *Solanum lycopersicum*, *Solanum pennellii*, *Solanum arcanum*, *Solanum habrochaites*, heterozygosity.

INTRODUCTION

The *Solanaceae* or nightshade family consists of more than 3000 species with great diversity in terms of habit, habitat and morphology. Its species occur worldwide, and range from large forest trees in wet rain forests to annual herbs in deserts (Knapp, 2002). *Solanum* is the largest genus in the family, and includes tomato (*Solanum lycopersicum*) and various other species of economic importance. Tomato breeding over recent decades has focused on higher productivity and adaption to different cultivation systems. Its economic success is reflected by the fact that, on a global scale, tomato is one of the most important vegetable crops, with a worldwide production of 161 million tonnes covering some 4 800 000 ha (http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor). However, domestication of tomato is clearly distinct from the species divergence by natural selection as a consequence of selecting for a limited set of traits, including fruit shape and size (Rodríguez *et al.*, 2011). As a result, its genetic basis has been seriously narrowed, known as the 'domestication syndrome' (Hammer, 1985; Doebley *et al.*, 2006; Bai and Lindhout, 2007; Bauchet and Causse, 2012). In more recent times, tomato has been adapted to different growing systems by adjustment of a small number of traits, including self-pruning, plant height, earliness, fruit morphology and fruit color (Rodríguez *et al.*, 2011; Bauchet and Causse, 2012). The relative small genetic variation became apparent in the face of rapidly changing environmental conditions, competing claims for arable lands, and new consumer requests. These challenges have pushed tomato breeding efforts towards better biotic and abiotic stress tolerance, higher productivity, and increased sensory and nutritional value. However, the reduced genetic variation that resulted from extensive inbreeding has decelerated tomato crop improvement. To enlarge the genetic basis, breeders now focus on introgression of desirable genes from wild relatives into the elite cultivars, but so far, this has been quite limited (Bai and Lindhout, 2007; Singh, 2007).

The first step of introgressive hybridization involves crosses of the cultivated tomato with heirloom species, wild relatives or more distant species of the tomato clade. Introgression breeding is possible as cultivated tomato and related wild species are intra-crossable, and most of wild species are also inter-crossable (Rick, 1979, 1986; Spooner *et al.*, 2005) despite the fact that diverse mating systems have evolved, varying from allogamous self-incompatible (SI) and facultative allogamous, to autogamous self-compatible (SC). Especially at the geographic margins of the distributions, inter-species changes in incompatibility systems that promote inbreeding over outcrossing have been documented (Peralta *et al.*, 2008;

Grandillo *et al.*, 2011). Species boundaries and genetic diversity have been extensively studied in tomato using a wide range of molecular data (Peralta *et al.*, 2008 and Grandillo *et al.*, 2011). For example, RFLP analysis showed that genetic diversity for SI species far exceeds that of SC species, estimated at 75% versus 7% (Miller and Tanksley, 1990). Furthermore, 'within-accession' genetic variation was estimated at 10% of the 'between-accession' variation, in contrast to the genetic variation of the modern cultivars, which was estimated at <5%. This further illustrates the dramatic erosion of genetic diversity in cultivated tomato crops.

Selection of crossing parents for inter-specific hybridization requires insight into phylogenetic relationships for the tomato clade, but trees based on morphological and molecular data have not been undisputed. Four informal species groups have been proposed for the tomato clade (*Lycopersicon*, *Arcanum*, *Eriopersicon* and *Neolycopersicon*; Peralta *et al.*, 2008), which are thought to have evolved from the most recent common ancestor approximately 7 million years ago (Nesbitt and Tanksley, 2002; Spooner *et al.*, 2005; Moyle, 2008; Peralta *et al.*, 2008). Despite these studies, evolutionary relationships between the 13 species in the *Lycopersicon* clade have not been fully resolved; for example, the dichotomy between *Solanum pennellii* and *Solanum habrochaites* (Peralta *et al.*, 2005, 2008; Spooner *et al.*, 2005). The evolutionary history of *Solanum* genomes has also been investigated from the perspective of chromosome organization. The study by Szinay *et al.* (2012) involving cross-species BAC FISH painting of *Solanum* species revealed few large rearrangements in the short arm euchromatin of chromosomes 6, 7 and 12, whereas Anderson *et al.* (2010) demonstrated pairing loops, multivalents and kinetochore shifts in synaptonemal complex spreads of hybrids between members of the tomato clade, suggesting paracentric and pericentric inversions and translocations between the homeologous chromosomes. Furthermore, comparative genomics suggest a *Solanum* genome landscape in which chromosome evolution for the majority of the 12 chromosomes has been far more dynamic than currently appreciated (Peters *et al.*, 2012). Collectively, these findings demonstrate that evolutionary relationships among the wild relatives should be considered provisional (Peralta *et al.*, 2008).

The availability of high-throughput sequencing technologies has provided unprecedented power to determine genome variation across entire clades, at both the structural and genotype level. Initiatives such as the 1001 Genomes Project for *Arabidopsis thaliana*, the *Drosophila* sequence project, and the 1000 Genomes Project for human have demonstrated the existence of a vast amount of intra-species specific polymorphic sequence features such as insertion/deletion events (InDels), repeats and single-nucleotide polymorphisms (SNPs) for hundreds of genes (Weigel and Mott, 2009; The 1000 Genomes Project Consortium, 2010; MacKay *et al.*, 2012), and have illustrated that there is no such thing as 'the genome' for a particular species. Rather, the range of physiological and developmental traits appears to be reflected in the tremendous amount of sequence variants contributing to intra-specific variation. Considering the overwhelming inter-species genetic variability, tomato germplasm collections represent a gene pool with unprecedented possibilities to address new breeding demands imposed by climate change, world population increase, and consumer needs. Here we have studied this genetic variation by genome sequencing a selection of representative tomato accessions, which has become possible through the recent development of the *S. lycopersicum* Heinz 1706 reference genome (The Tomato Genome Consortium, 2012). In addition to this reference genome for the *Lycopersicon* species, we describe construction of reference genomes for three other related species representing the *Arcanum*, *Eriopersicon* and *Neolycopersicon* groups, respectively, providing an expanded resource for detailed comparative genomic studies in the near future. We also present results for robust/high-confidence detection and identification of sequence polymorphisms, heterozygosity levels and introgressions, and assess the genetic diversity within the tomato clade from a phylogenetic and evolutionary perspective. This study provides an invaluable dataset for advanced 'omics' studies on sequence trait relationships and the molecular mechanisms of tomato genome evolution, as well as the development of genotyping-by-sequencing breeding approaches.

RESULTS

Selection of tomato accessions

We selected 84 accessions of the *Solanum* clade section *Lycopersicum* for shallow whole-genome sequencing (36-fold coverage). The first set of 54 accessions consists of tomato landraces and heirloom cultivars of *S. lycopersicum* and *S. lycopersicum* var. *cerasiforme*, which have been selected from the EU–SOL tomato core collection (https://www.eu-sol.wur.nl). The second set of 30 accessions comprises wild relatives of tomato, representing the full range of expected genetic variation within *S. lycopersicum*. Their selection was based on previous use in genetic research and previous utilization of quality or (a)biotic stress traits (Grandillo *et al.*, 2011). We also chose *S. arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716 for *de novo* sequencing and whole-genome reconstruction, aiming to produce a reference genome for each of the four main phylogenetic groups in the tomato clade. An important selection criterion was the self-compatibility of these accessions, allowing inbreeding for several generations to minimize heterozygosity, and so reduce *de novo* genome assembly problems. A complete list of the

phenotyped selected accessions used in this study is presented in Table S1 and at http://www.tomatogenome.net.

De novo assembly of three wild tomato relatives and Heinz

Comparisons of molecular data have indicated relatively low DNA sequence diversity between genetically related species within the phylogenetic groups of the tomato clade (Miller and Tanksley, 1990). Furthermore, preliminary analysis indicated that the SNP frequencies for *S. pimpinellifolium* and *S. pennellii* compared to *S. lycopersicum* were 1 and 10%, respectively. As *S. pimpinellifolium* and *S. pennellii* are phylogenetically among the closest and most distantly related species to tomato, respectively, we assumed the same range of SNP frequencies for other species. Our strategy to determine the proportion of polymorphic loci across the entire *Lycopersicon* clade therefore comprised *de novo* sequencing and assembly of three new reference genomes, followed by shallow sequencing of the bulk of the accessions and subsequently mapping them to a reference genome. For reference genome reconstruction, we relied on massive parallel sequencing using Illumina HiSeq 2000 (http://www.illumina.com/systems.ilmn) and 454 FLX technology (http://454.com/products/gs-flx-system/). Two paired-end libraries with insert sizes of 170 and 500 bp and one MP library of 2 kbp were sequenced using the Illumina HiSeq 2000 at 25-, 25- and 30-fold coverage, respectively (assuming a genome size of 950 Mbp), and at least 80% of the bases had a *Q* value >30 (error rate ≤1/1000). For the 454 FLX sequencing, large-insert size libraries of 8 and 20 kbp were created, each at 0.6-fold coverage. For *S. pennellii* LA0716, we used an additional 8 kbp Illumina MP library at 0.4-fold coverage. We discarded unpaired reads, resulting in 205-fold coverage. For *de novo* assembly, we aimed to maximize short-range contiguity, long-range connectivity, completeness and quality by following the strategy outlined by Gnerre *et al.* (2011). Our assembly statistics show a total contig length for *S. arcanum*, *S. habrochaites* and *S. pennellii* of approximately 760 Mb (Figure 1 and Table S2). The unique portion is comparable in size in these genomes, consistent with widespread research, including comparative mapping studies, revealing a high level of synteny among species of the *Solanaceae* (Paterson *et al.*, 2000). However, previous estimates on DNA content and flow cytometry analyses suggest a considerable variation in total genome size among species in the tomato clade (Grandillo *et al.*, 2011). For example, the DNA content of cultivated tomato varies from 1.87 to 2.07 pg/2C, indicating a genome size of approximately 950 Mbp, whereas that of *S. pennellii* is substantially larger and corresponds to a DNA content of 2.47–2.77 pg/2C, corresponding to 1200 Mbp. Furthermore, we assume that most of the estimated 35 000 genes reside on the approximately 220–250 Mb of DNA in the euchromatic regions (http://www.rbgkew.org.uk/cval/; Arumuganathan and Earle, 1991; Van der Hoeven *et al.*, 2002; The Tomato Genome Consortium, 2012). For the most part, the increased genome size of *S. pennellii* may probably be explained by expansion of the repetitive portion of the genome. Repeats are known to impede genome reconstruction, resulting in a more fragmented assembly and a lower N50 contig size. This is consistent with the *S. pennellii* LA0716 assembly statistics (Table S2). The re-assembled *de novo* *S. lycopersicum* cv. Heinz reaches the assembly size plateau more slowly, and also appears to be more fragmented than the published reference genome, probably due to previous use of older sequencing platforms and a BAC-by-BAC sequencing strategy.

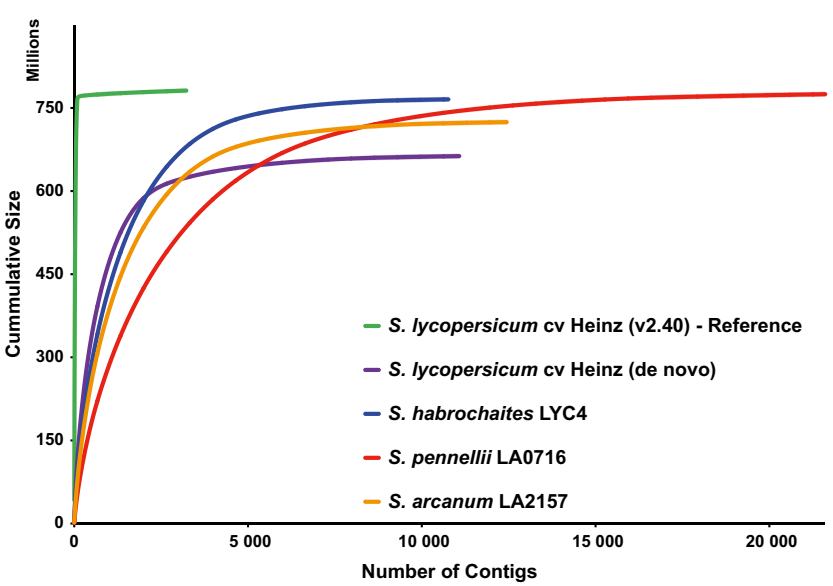

Figure 1. Evaluation of genome assemblies.
De novo assemblies for *S. lycopersicum* Heinz 1706, *S. habrochaites* LYC4, *S. pennellii* LA0716 and *S. arcanum* LA2157 were generated using the AllPaths-LG assembler and scaffolded with the Scarpa scaffolder using 454 data. The number of contigs (*x* axis) is plotted against the cumulative contig size (*y* axis) with contigs ordered by size (largest first). Values for the gold standard assembly *S. lycopersicum* cv. Heinz 1706 version 2.40 are plotted in green.

To determine the extent of sequence diversity, read pairs from *de novo* sequenced genomes were mapped to the *S. lycopersicum* cv. Heinz 1706 version 2.40 reference genome. *S. lycopersicum* has the lowest number of unmapped reads (11%), which probably consist of low-quality sequences; this value increases for *S. arcanum* (17%), *S. pennellii* (22%) and *S. habrochaites* (25%). Given the comparable sequence quality, we assume an equal percentage of low-quality reads for the *de novo* sequenced genomes, while sequence diversity, introgressions and genome expansion contribute to the remainder of the unmapped reads.

Sequencing and mapping of the 84 accessions and wild species

For the 84 accessions, 2.9×10^{12} base pairs were sequenced, giving a mean coverage of 36.7 $\pm$ 2.3-fold per accession (32.5 $\pm$ 2.1-fold for $Q \geq 30$). All individuals were mapped against *S. lycopersicum* cv. Heinz 1706 version 2.40 to assess the diversity in both crop and wild species, resulting in 96.4 $\pm$ 0.88 and 52.9 $\pm$ 2.93% of the reads correctly mapping for crops and wild species, respectively (Figure 2). These numbers improved when reads from wild species were mapped against a reference genome from a closer relative. For *S. arcanum*, *S. habrochaites* and *S. pennellii*, 72.87 $\pm$ 7.87, 78.74 $\pm$ 15.63 and 55.37 $\pm$ 9.29% of the reads correctly mapped against the *S. arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716 reference genomes, respectively. These results illustrate the large genetic erosion within the crop tomatoes, and the large sequence diversity among the wild species. Moreover, they emphasize the need for multiple reference genomes to support unbiased interpretations of genetic variation consequences among species in the tomato clade.

Whole-genome sequence diversity

To further assess the sequence diversity in *Solanum* section *Lycopersicon*, we quantified and classified the SNPs for each of the 84 accessions using read mappings against Heinz. The SNP counts for tomato cultivars are relatively low and gradually increase for *S. galapagense*, *S. cheesmaniae* and *S. pimpinellifolium* accessions. The SNP numbers for specific members of the *Arcanum*, *Eriopersicon* and *Neolycopersicon* groups increase sharply (Figures 3 and S1), which correlates with their more distant position in the phylogenetic tree in the tomato clade (Peralta *et al.*, 2008).

When compared to the Heinz annotated genome, we consistently observed a significantly higher SNP frequency in intergenic regions than in genic regions for all accessions: 89.47 $\pm$ 3.03% of the polymorphisms fall into intergenic regions, while 7.55 $\pm$ 2.19% map to introns and 2.33 $\pm$ 0.68% map to exons (Figure 4). Of the polymorphisms in exons, 55.17 $\pm$ 11.54% are synonymous while 44.83 $\pm$ 21.03% are non-synonymous (Figure S2).

The mean number of SNPs in wild species is 20 times higher than in crop tomatoes. These results are consistent with the notion that crop tomato genomes are extensively genetically eroded compared to the large genetic diversity found among wild species. A striking trend is the genome-wide ratio between synonymous and non-synonymous SNPs (dN/dS). For crops, non-synonymous SNPs outnumber synonymous SNPs, while the opposite is generally true for wild species (Figure S2). Although we currently have no clear explanation for the higher dN frequency in crop accessions, it may partly be the result of the artificial selection pressure imposed by breeding, maintaining a relatively small number of SNP determining traits under positive selection, while allowing fixation of many

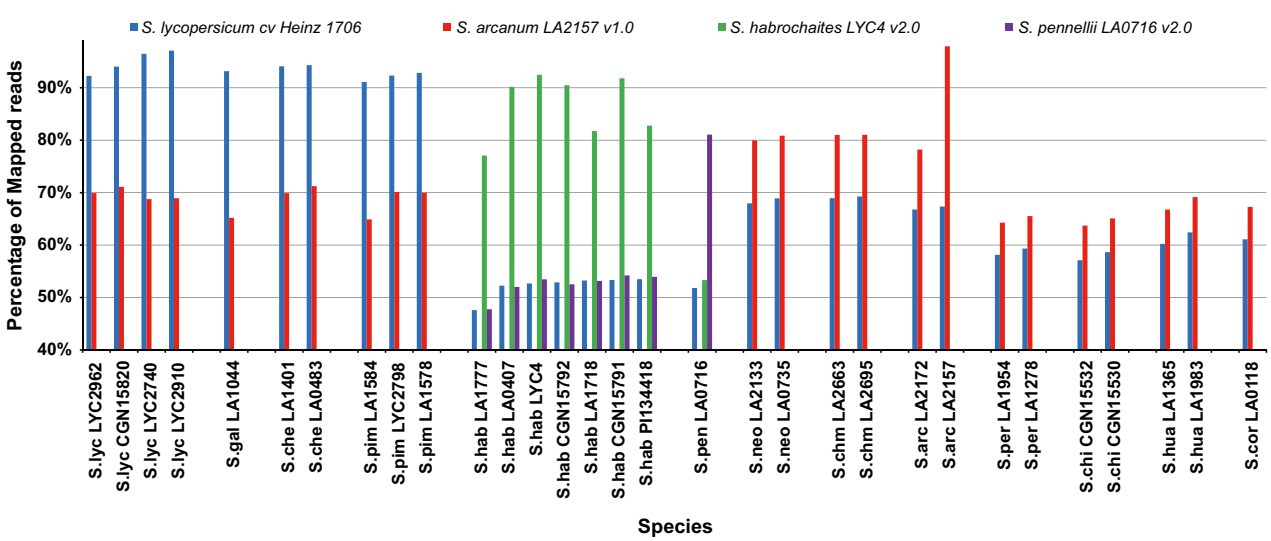

Figure 2. Percentage reads from *S. lycopersicum*, *S. arcanum*, *S. habrochaites* and *S. pennellii* accessions mapped against reference genomes. Species are indicated on the *x* axis. Bar color codes correspond to the reference genome indicated in the legend that was used for mapping.

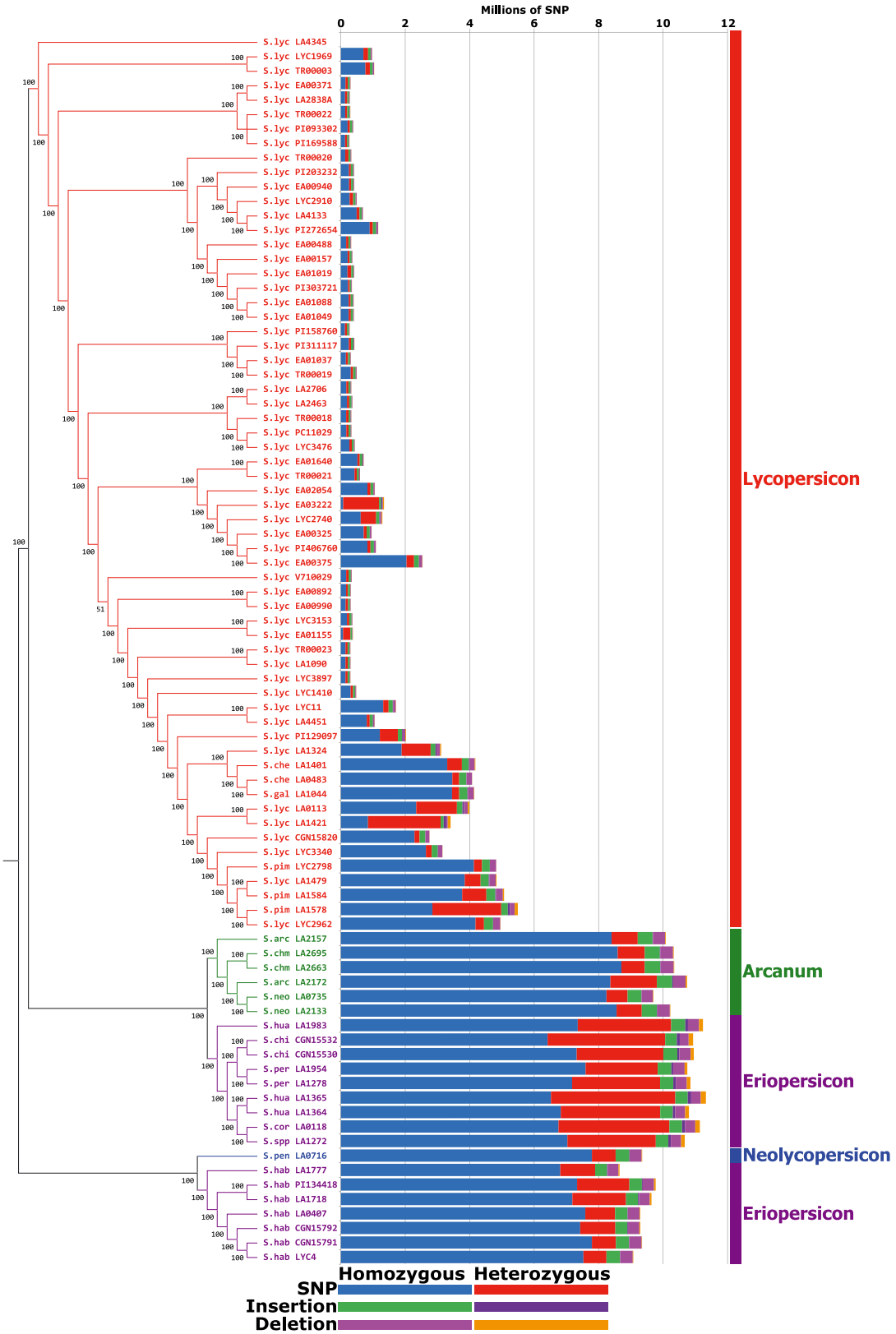

Figure 3. Strict consensus tree based on whole-genome homozygous SNPs from 84 accessions with overlaid bootstrap values obtained by maximum-likelihood analysis. Bars show the numbers of SNPs (in millions) for the various classes of polymorphisms per accession.

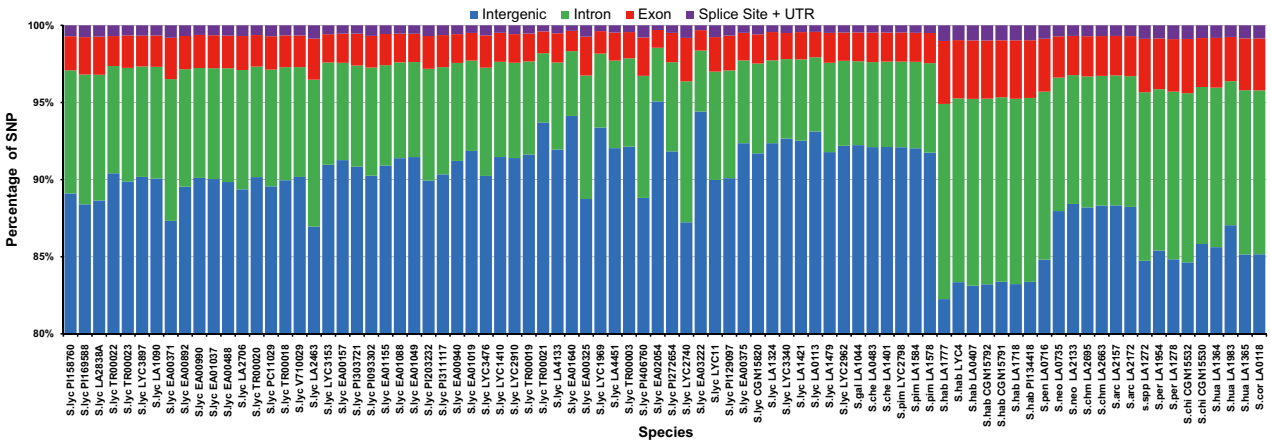

Figure 4. Genome-wide SNP ratio for 84 accessions. SNP classes are color-coded as indicated. Accession IDs and the percentage of SNPs are indicated on the *x* and *y* axes, respectively.

deleterious non-synonymous SNPs, as previously reported for tomato crop accessions (Sim *et al.*, 2011, 2012; Koenig *et al.*, 2013).

An overview of the SNP and InDel variation in the 84 accessions supported by JBrowse (Skinner *et al.*, 2009; Westesson *et al.*, 2013) and downloadable consensus VCF files may be accessed in the tomato 100+ variant browser that is publicly accessible via http://www.tomatogenome.net/VariantBrowser/.

Heterozygosity and introgressions

For the *Lycopersicum* accessions, the highest heterozygosity levels were observed for beef-type accessions *S. lycopersicum* EA03222 and EA01155 (Dana), as shown in Figure 3. With respect to mating type, highest ratios were found for allogamous SI wild species, while facultative SC wild species display an intermediate heterozygosity ratio (Figure S3). On average, autogamous SC species have a slightly lower heterozygosity level compared to facultative SC species, of which the autogamous SC wild species *S. neorickii* LA0735 has the lowest (Figure S3). The 'northern' *Eriopersicon* species group displays a higher level of heterozygosity than the 'southern' *Eriopersicon* and *Arcanum* species group.

Surprisingly, some tomato accessions display very high SNP counts (Figures 3 and S1), which may be attributed to introgressions. Indeed, additional footprints for inter-specific introgressions in 54 *lycopersicum* accessions compared to Heinz were found by testing the SNP distributions in 1 Mbp-sized bins along each of the 12 chromosomes. Bins with SNP counts deviating significantly from the mean (*z* test, $P < 0.05$) were considered as either introgressed in crop accessions or natural highly divergent in wild *S. lycopersicum* species after subtracting the number of SNPs found when mapping Heinz reads against itself and correcting for chromosome position- and species-specific effects by median polish. Introgressions occurred in $5.56 \pm 7.98\%$ of the 767 1 Mbp bins, with *S. lycopersicum* PC11029 having the smallest number (0.13%), *S. lycopersicum* LA0113 having the highest number (31.42%), and *S. lycopersicum* PI272654 having 0.91% of its bins marked as introgression. Cherry, giant and beef tomatoes have a higher number of introgressions among the crops, while wild *S. lycopersicum* species are even more divergent (Figure S4).

The specific chromosome locations of divergent SNP intervals are shown in Figure S4. Here, similar patterns of SNP concentrations may be observed between crop accessions, which most probably are introgressions originating from the same donor accessions. In some cases, the most likely source of introgressions may be deduced from the SNP identity and the phylogenetic distance inferred from the SNP alignment. Indeed, when plotting the chromosomal SNP distribution, we found a 2.2 Mb introgressed segment in the long arm of chromosome 6 roughly between Tomato-EXPEN2000 genetic markers C2_At4g10030 (44 cM) and TG365 (50 cM) for accessions LA2838A (Alisa Craig), LA2706 (MoneyMaker), LA2463 (All Round) and CGN15820. Phylogenetic distance analysis revealed that a 2.2 Mb segment in the heirloom open-pollinated tomato accession MoneyMaker is most closely related to the wild species *S. pimpinellifolium* LYC2798 (Figure S5). Interestingly, the Heinz ITAG 2.4 annotation of this segment indicates several loci that have been implicated in hormone-induced stress responses, fruit development, flavonoid phytonutrient production, and mitogen-activated protein kinase-mediated production of reactive oxygen species involved in innate immunity to *Phytophtera infestans*-induced late blight.

Sequence diversity and phylogenetic relationships

SNPs in genes related to fruit and growth diversification. To analyze diversity in specific genes and loci that underlie a phenotypic effect on fruit diversification and

plant growth (FDPG), we determined the orthologs for *ovate* (Solyc02g085500), *fw2.2* (Solyc02g090740), *ls* (Solyc07g066250), *og*/*beta* (Solyc06g074240), *lcy1* (Solyc04g040190), *lfy* (Solyc03g118160), *rin* (Solyc05g012020), *sp* (Solyc06g074350), *fer* (Solyc06g051550), *style* (Solyc02g093580), *psy1* (Solyc03g031860), *lin5* (Solyc09g010080) and locus *lc* (gb|JF284941). Based on the ITAG 2.4 annotation of the tomato reference genome (The Tomato Genome Consortium, 2012), the polymorphisms in the orthologs were classified as coding or non-coding, and as non-synonymous or synonymous (silent) SNPs to compare intra- and inter-species sequence diversity and SNP effects among the 84 accessions. FDPG genes in many cases underlie a phenotypic trait that is determined by a few SNPs or sometimes a single one (see below). While tomato breeding has primarily been directed toward selection of these traits, it is conceivable that a SNP determining a single trait was subjected to positive selection but the bulk of the genes were subjected to more relaxed selection. Non-synonymous SNP counts in FDPG genes from wild species are consistently higher than observed for a randomly selected set of genes. In contrast, synonymous SNP are consistently lower. Nevertheless, both counts are just below one standard deviation from the mean (Figure S6). Perhaps this observation reflects a higher selection pressure in wild species than in crop accessions against deleterious mutations in FDPG genes.

Lycopersicon-, Aracanum-, Eriopersicon- and Neolycopersicon-specific SNPs. Several characteristic SNPs were found to be distinctive for the *Lycopersicon*, *Arcanum*, *Eriopersicon* and *Neolycopersicon* sections. For example, the red or orange/yellow-fruited *Lycopersicon* group accessions have a GTC codon for valine at position 23 in the *og*/*beta* gene on tomato chromosome 6 encoding chromoplast-specific lycopene β–cyclase, whereas the green-fruited *Arcanum*, *Eriopersicon and Neolycopersicon* species have a non-synonymous TTC (Phe) substitution at the same position. In all accessions in the *Arcanum* group, the GAG codon for the Gln30 residue of lycopene β–cyclase 1 encoded by the *lcy1* gene on chromosome 4 has been substituted. *S. chmielewskii* accessions have a GTG codon (Val), while the *S. neorickii* and *S. arcanum* accessions have a CTG codon(Leu).

Species-specific SNPs. In addition to group-specific polymorphisms, we also found species-specific SNPs. For example, further downstream in the *og*/*beta* orthologous gene of *S. corneliomulleri*, a GCT (Ala437) → ACT (Thr) SNP and a TTG (Leu464) → TTT (Phe) SNP occur, and the AAA codon for amino acid Lys277 is replaced by ATA (Ile) for the *S. chmielewskii* accessions. In the *lfy* gene orthologs, a synonymous SNP TTA (Leu25) → CTA is shared by the *huaylasense* accessions, whereas a CCA (Pro122) → CAA (Gln) nucleotide substitution is characteristic of *S. chmiemlewskii* accessions. *S. peruvianum* accessions have a CGATGA insertion (Asp Asp) downstream of Asp89 in the *fer* gene. *S. arcanum* and *S. chilense* accessions share a GCC (Ala107) → GCA synonymous SNP in the *ovate* gene ortholog, and we detected several intron SNPs that are specific for *S. neorickii* in the *sp* orthologous gene. Finally, *S. chilense* accessions have a TTT (Phe80) → TTC substitution in the *style* gene.

Accession-specific SNPs related to fruit traits. We also observed accession-specific polymorphisms related to specific fruit traits. The cultivar Black Cherry has a single nucleotide deletion in the coding sequence of the chromosome 6 *B* gene (Figure S7). This specific deletion occurs in the *old-gold-crimson* (og^c) null allele (Ronen *et al.*, 2000). The resulting frameshift causes the loss of lycopene β–cyclase function, resulting in accumulation of lycopene and dark red/purple appearance of the tomato fruits. Galina, lidi and T1039 are yellow-skinned cherry tomatoes, and have a single nucleotide deletion resulting in a frameshift causing a Lys389 → Ser substitution and a premature TGA stop codon directly downstream, resulting in a truncated *psy1* protein lacking the terminal 23 amino acids (Figure S7). In this respect, it is interesting to note that the r^y mutant allele, which encodes a phytoene synthase lacking these terminal amino acids, has been shown to underlie the yellow-colored fruit skin phenotype in tomato mutants (Fray and Grierson, 1993).

Fruit shape and size in tomato are influenced by locule number. Two QTLs, *lc* and *fas*, have major effects on these traits, and may act synergistically, leading to extremely high locule numbers (Cong *et al.*, 2008; Munos *et al.*, 2011). *Fas* is the major gene responsible for increasing locule numbers (from two to more than six), while *lc* has a weaker effect, increasing locule numbers to three or four. Two T → C and A → G SNPs are associated with the high locule number allele (lc^h), while an extreme high locule number caused by down-regulation of a YABBY-like transcription factor is associated with a 6–8 kb insertion in the first intron of the *fas* gene (Cong *et al.*, 2008). Sequence analysis revealed that all bi-locule accessions have the low locule number allele (lc^l), while accessions with three to four locules (except Cal J TM VF and Dana) have the lc^h allele. Whether tomato fruit are pear-shaped is controlled by the quantitative trait locus *OVATE*. The allelic interactions at the *ovate* locus have been described as recessive, but their expression depends on the genetic background (Ku *et al.*, 1999). Liu *et al.* (2002) showed that a GAA (Glu279) → TAA nonsense mutation in the second exon causes an early stop codon and a premature translation termination, resulting in a 75 amino acid truncated ovate protein (AAN17752), leading to formation of pear-shaped fruit. All accessions with pear-shaped fruits have the

premature stop codon, but the mutational effect is less pronounced in the ovate-fruited accession 'Porter' (Figure S8).

Phylogenetic relationships

Cladistics based on molecular data resulted in clear grouping of species within the *Solanum* genus section *Lycopersicon* (Peralta *et al.*, 2008). However, at the species level, relationships are still unresolved. For example, while *S. pennellii* was placed in its own group (*Neolycopersicon*) as a sister to the rest of the section *Lycopersicon*, it nonetheless appears as sister to *S. habrochaites* in the main trichotomy (Spooner *et al.*, 2005; Peralta *et al.*, 2008; Grandillo *et al.*, 2011). Our SNP analysis indicates that many polymorphisms are distinct for *habrochaites* species, whereas *S. pennellii* LA0716 shares many SNPs with accessions of the *Arcanum* and *Eriopersicon* groups. This suggests the existence of a complicated phylogenetic relationship for *S. pennellii* and *S. habrochaites*.

We applied the vast amount of multi-locus molecular data obtained in this study to shed more light on the species and accession relationships in the tomato clade. First, we used the polymorphisms detected in the FDPG gene orthologs to assess species boundaries and relationships within the tomatoes and wild relatives. A strict consensus tree for ten concatenated genes (Figure S9) revealed that all *S. habrochaites* species cluster into a monophyletic group, while *S. pennellii* LA0716 is sister to *S. habrochaites*. The *S. chilense* accessions also group together and cluster with *S. corneliomulleri* and *S. peruvianum* accessions, which are representatives of the erstwhile *S. peruvianum* 'southern group', and with accession LA2172. The green-fruited self-compatible (SC) *S. chmielewskii* and *S. neorickii* species, which are representatives of the *Arcanum* group (Peralta *et al.*, 2008), are resolved into two monophyletic groups and cluster with two *S. arcanum* species into a larger clade. Furthermore, all red- or orange-fruited SC species of the *Lycopersicon* group (*S. cheesmaniae*, *S. galapagense*, *S. lycopersicum* and *S. pimpinellifolium*) form a well-supported clade. In particular, the orange-fruited *S. cheesmaniae* and *S. galapagense* cluster into a sub-group, illustrating the very close relationship between both species. These relations are in agreement with previously presented phylogenetic studies (Grandillo *et al.*, 2011).

Next we excluded heterozygous SNPs from the analysis, as they are arbitrarily converted into a single nucleotide call for FASTA-converted sequences, thereby introducing noise and possible bias in the data. SNPs in introns were also excluded as they are likely to be under less selective pressure than exon SNPs and probably carry less phylogenetic information and introduce more noise. Figure S10 shows that the homozygous SNPs in the FDPG genes have sufficient power to resolve the phylogenetic placement consistent with the grouping at the sectional level as previously described (Peralta *et al.*, 2008). We noticed a slight increase in resolution when comparing the gene trees based on unfiltered and filtered SNPs (Figures S9 and S10). Nevertheless, at the species level, placement of *lycopersicum*, *pimpinellifolium*, *galapagense* and *cheesmaniae* accessions appeared largely unresolved. In this analysis, *S. pennellii* is a sister species to the *Eriopersicon* group. We also assessed the clustering using genome-wide homozygous SNPs aiming to increase resolution. The whole-genome SNP cladogram in Figure 3 shows complete resolution into separate branches, with high bootstrap values for each of the *Lycopersicon* accessions and wild species. Although phylogenetic relationships may be influenced by SNPs that arise from introgressions, the genome-wide SNP information generates sufficient resolution power, and enables inter-species and intra-species identification of all 84 individuals in monophyletic groups. Based on our phylogenetic analysis and SNP sequences, we propose that accession LYC2740 should be considered as an *S. lycopersicum* species rather than a *S. pimpinellifolium* species. We also observed that several *S. lycopersicum* accessions grouped with *S. pimpinellifolium*, *S. galapagense* and *S. cheesmaniae*. These *S. lycopersicum* accessions probably are hybrids or carry substantial *S. pimpinellifoulium* introgressions. Additional analysis should be performed to substantiate this hypothesis. In addition, *S. pennellii* appears to be a sister species to the *S. habrochaites* species group in the whole-genome SNP tree, suggesting that *S. pennellii* may be considered an intermediate species between *S. habrochaites* and *S. arcanum*, which coincides with its intermediate geographical distribution.

DISCUSSION

Multiple reference genomes and sequence diversity

Our study has yielded a huge amount of precious data on sequence diversity in wild species of the tomato clade. The reads for *S. habrochaites* (78%), *S. arcanum* (73%) and *S. pennellii* (53%) were mapped onto the corresponding species reference genome, illustrating the large inter-species sequence variation in the *Lycopersicon* clade. We also demonstrated dramatic genetic erosion in cultivated tomatoes. There is an increasing demand for broadening the genetic base of this crop, and our study provides pivotal information for future tomato breeding programs. The Heinz reference genome is not simply only partly representative of the genetic and structural information in the related wild species, it also emphasizes the need to reconstruct additional reference genomes. The three *de novo* sequenced genomes presented here thus constitute a valuable resource in addition to the currently available genomic tools in support of studies on evolution, domestication and the genetic basis underlying important traits such as disease resistance and abiotic stress tolerance.

Our sequencing and mapping strategy effectively supports the detection and identification of high-confidence sequence polymorphisms and is explanatory for the rich phenotypic diversity among a large set of cultivated tomato accessions and wild relatives. We observed group-, species- and accession-specific polymorphisms, some of which may be attributed to economically important fruit and growth traits. Such information may easily be translated into array- or PCR-based assays to genotype genetic variants across extensive populations as well as progeny populations. Assuming that gene models from the Heinz annotated reference genome are also applicable for other species, 8–10% of all sequence polymorphisms are located in the genic portion of the genome. Non-synonymous and synonymous SNPs occur each at 1% of the total number of SNPs. As a conserved estimate for wild species, such a percentage equates to approximately 1×10^5 non-synonymous SNPs, but little is known about how much of the phenotypic diversity may be attributed to this. Given that traits such as fruit color and shape are determined by a single SNP, the total number of SNPs most likely represents a wealth of diversity that awaits further exploration.

Relationships of tomatoes and wild species relatives

The past few decades have seen the publication of several phylogenetic studies of *Solanum* species in the *Lycopersicon* section, but use of phenotypic characteristics, markers and sequencing data for their construction resulted in dissimilar trees, with provisional species groupings lacking fully resolved relationships (Grandillo *et al.*, 2011). In this study, we reconstructed intra- and inter-species relationships for a large number of tomato accessions and related wild species, taking advantage of whole-genome sequence data to maximize tree resolution. Initially, our phylogenetic analysis focused on genes controlling economically important traits that have been subject to inter-specific hybridization breeding. As a subset of these genes originated from wild species, cladistics potentially may result in skewed relationships. However, our maximum-likelihood consensus cladograms for the targeted genes and the whole-genome SNP data show comparable tree topology down to the sub-sectional (species group) level, suggesting that phylogenetic relationships between fruit and growth diversification genes are not particularly skewed. While the strict consensus cladogram for the concatenated fruit and growth diversification genes displays unresolved relationships at the species level for some of the cultivated tomato and *S. peruvianum* accessions, the use of whole-genome SNP data allowed increased tree resolution. Indeed, the whole-genome SNP dataset supports the placement of taxa into separate branches with high bootstrap values for each of the accessions and wild species, including corrected placement of several previously putatively typed accessions.

Ecological differences probably have resulted in dramatic genome evolutionary consequences. Moreover, there is evidence that mating system shifts have a large impact on complex multigene-based traits such as floral and fruit development (Moyle, 2008), which may further account for large intra-species variation. Large intra-species variation trends have been observed for *S. chilense* species, which have been grouped into geographic races that may be distinguished both morphologically and genetically (Grandillo *et al.*, 2011). Other examples involve remarkable levels of morphological and genetic diversity, as found in *S. peruvianum* populations (Rick, 1986; Städler *et al.*, 2005), which may apply also to *S. arcanum* explaining the distinct phylogenetic positions for accessions LA2157 and LA2172. Here, we placed both accessions into the *Arcanum* group with 'northern' species of the *peruvianum* complex. Accession LA2172, which is allogamous SI, appears to be sister to the monophyletic *S. neorickii* clade, while LA2157 is facultative SC and sister to the monophyletic *S. chmielewskii* clade. Furthermore, it is important to note that AFLP cladistics previously resulted in grouping of *S. arcanum* LA1984 with southern *S. peruvianum* species, while the other *S. arcanum* accessions grouped with *S. huaylasense* (Spooner *et al.*, 2005). Interestingly, it has been speculated that accessions such as LA1984 may represent a 'crossing bridge' between morphologically and genetically distinct populations (Rick, 1986; Grandillo *et al.*, 2011).

Detection of introgressions in crop accessions and genome structure

While marker-assisted introgressions focus on the relationships between traits and allelic variants, they are mostly used for indirect selection of genetic determinants for a trait of interest, and are restricted to alleles that may be diagnosed. Based on genome-wide SNP data, introgressions from *S. pimpinellifolium* into chromosomes 4, 9, 11 and 12 of *S. lycopersicum* Heinz 1706 have been reported previously (The Tomato Genome Consortium, 2012). Following the same strategy, the bulk of our introgression detection was based on SNP distributions divergent from the reference genome, targeting introgressions not present in Heinz. Our approach shows that both the location and size of introgressed segments may be inferred from the SNP distribution. By testing the phylogenetic distance for 84 accessions, we found a 2.2 Mb chromosome 6 introgressed segment in *S. lycopersicum* acceptor accessions LA2838A, LA2706, and CGN15820, and assigned *S. pimpinellifolium* LYC2798 as the closest related donor accession. These results add a new perspective to future introgression hybridization breeding.

The success of introgressive hybridization breeding depends, among other things, on proper identification of colinear chromosome segments in donor and recipient

genomes, which, in turn, is dependent on the consistency and completeness of their assembled genomes. The genome structure of the parental species influences crossing success, and a difference in genomic colinearity has a direct effect on chromosome pairing at meiosis, and hence determines the rate of alien chromatin transfer into a recipient crop. However, the proper ordering and orientation of contigs into megabase-sized scaffolds depends on the availability of genetic and physical maps, which are currently lacking for the three *de novo* sequenced genomes. Furthermore, the N50 contig sizes for the *de novo* assemblies of *S. arcanum*, *S. habrochaites* and *S. pennellii* do not exceed 400 kb. Although advances in next-generation sequencing technology for the use of extant germplasm resources now allow assembly of large numbers of complex genomes relatively rapidly and cheaply, they do not yet allow full genome reconstruction of the *Solanaceae* family, and hence are of limited use in introgression breeding. The identification of compatible genomes for introgression breeding, the rearrangement phylogeny within the *Solanaceae*, and reconstruction of the ancestral *Solanum* karyotype all require additional physical mapping information in addition to the genome sequence information to properly order contigs along the chromosome arms. Therefore, we suggest integration of next-generation sequencing and new technological platforms, such as optical mapping, to advance *Solanum* genome reconstruction.

EXPERIMENTAL PROCEDURES

Selection of tomato accessions

We genotyped the 7000 accessions in the EU–SOL project (https://www.eu-sol.wur.nl) on the basis of 20 traits and markers, followed by denser genotyping of a subset of 1000 accessions using 384 SNP markers, and a final selection of 200 accessions covering the full genetic diversity of the crop. We also included a set of old cultivars that were selected on the basis of previously documented trait identifications of wild tomato relatives (Grandillo *et al.*, 2011).

DNA isolation

Young leaves were collected from the first plant of each plot (self-compatible accessions) or from the pollen acceptor (self-incompatible accessions) for DNA extraction. Approximately 100 mg frozen leaf material was ground using an M300 mixer mill (Retsch, www.retsch.com). Subsequently, genomic DNA was extracted by a standard DNA isolation protocol (Van der Beek *et al.*, 1992), using a nuclear lysis buffer containing sarkosyl. The DNA was quantified using a Qubit 2.0 fluorometer (Lifetechnologies, www.lifetechnologies.com/qubit.html). For each accession, 1.5–2.0 μg DNA was used for library construction.

Sequencing and read mapping of Illumina and 454 libraries

Shallow sequencing of 500 bp inserts was performed using an Illumina HiSeq 2000 sequencer to generate a 100 bp paired-end (PE) library at a mean coverage of 36-fold. Bases with $Q < 20$ were trimmed before read mapping with BWA (Li and Durbin, 2009, 2010) against *S. lycopersicum* cv. Heinz version 2.40 with a maximum insert size of 750 bp (50% deviation), taking into account at most 30 hits and removing PCR duplicates. SAMTOOLS (Li *et al.*, 2009) was used for variant calling without skipping InDels, a minimum gap distance of 5 bp, a minimum alignment quality of 20, a minimum depth of 4, and default parameters otherwise. The same protocol was used to map the wild species to their closest *de novo* version 1.0 assembled counterpart. Contamination with *Escherichia coli*, human, insect, mouse, ΦX174 phage, yeast and phytoviral genomes (Adams and Antoniw, 2006) was checked using BOWTIE (Langmead *et al.*, 2009).

De novo assembly of the three wild species genomes and Heinz

For *de novo* sequencing of *S. arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716, we sequenced an overlapping PE library with 170 bp insert size, at 93.2-, 76.4- and 80.8-fold coverage, re-used the 500 bp insert size PE library at 35.7-, 35.6- and 28.2-fold coverage, using 100 bp Illumina HiSeq 2000 reads, and used a mate pair (MP) library with 2 kbp insert size at 33.8-, 38.0- and 31.2-fold coverage, respectively. Using the 454 FLX system, a long mate pair library of 8 kbp insert size and an extra-long mate pair library of 20 kbp insert size were sequenced at 0.55 ± 0.10- and 0.47 ± 0.07-fold coverage, respectively. For *S. pennellii* LA0716, we sequenced an additional short mate pair library of 3 kbp insert size at 0.4-fold coverage. On average, reads produced from 454 libraries contained 35 ± 7% adaptamers. For *S. lycopersicum* cv. Heinz 1706, we used a reduced set of reads with 14.78-fold 250 PE, 17.54-fold 300 PE, 37.42-fold 500 PE, 6.25-fold 2 kb MP, 6.51-fold 3 kb MP, 5.94-fold 4 kb MP and 6.02-fold 5 kb MP, with a total of 69.74-fold coverage for PE libraries, 24.73-fold coverage for MP libraries and 94.47-fold coverage overall.

The *S. pennellii*, *S. habrochaites* and *S. lycopersicum* data were assembled using ALLPATHS-LG (assembly version SL 2.0) as described by Gnerre *et al.* (2011) with a ploidy of 2, while *S. arcanum* was assembled using CLC GENOMICS WORKBENCH version 7 (CLC Inc., www.clcbio.com) with a bubble size of 300, a minimum contig length of 200, and a word size of 64 (assembly version 1.0). Subsequently, *S. arcanum* was assembled using ALLPATHS-LG (assembly version 2.0). ALLPATHS-LG-generated scaffolds were further scaffolded using the 454 FLX data in the SCARPA scaffolder (Donmez and Brudno, 2013). The *de novo* assembly statistics were compared to the tomato reference genome *S. lycopersicum* cv. Heinz version SL 2.40 (Table S2). The CLC, ALLPATHS-LG and ALLPATHS-LG plus SCARPA assembly are referred to as versions 1.0, 2.0 and 3.0, respectively. *S. arcanum* version 1.0, *S. habrochaites* version 2.0 and *S. pennellii* version 2.0 were used for mapping of the 84 accessions. Version 3.0 was used to assess genome sizes and rearrangements for all species.

Sequence diversity and phylogenetic relationships of 84 accessions

To assess sequence diversity in FDPG genes, orthologs in 84 accessions were obtained from reciprocal best BLASTN hits of CLC-assembled contigs and tomato ITAG 2.4 annotated sequences (http://solgenomics.net/gb2/gbrowse/ITAG2.4_genomic/) and aligned using CLUSTAL W (Thompson *et al.*, 1994). SNPs were then called using the quality-based variant detection algorithm in CLC. Optimal substitution models for CLUSTAL W -aligned gene and concatenated gene sequences were calculated using MEGA 5.1 (Tamura *et al.*, 2011). Maximum-likelihood trees for each individual gene as

well as concatenated gene sequences were inferred using a neighbor-joining initial tree followed by nearest-neighbor interchange. Phylogenies were tested using 1000 random genes separated into 100 sets of ten genes (Figure S6). Finally, strict consensus trees for individual genes and concatenated genes were calculated using a cut-off value of 50%.

For each species, we used a concatenation of all homozygous non-unique SNPs (Van Gent *et al.*, 2011) with quality above 20, which were obtained from the VCF files generated using BWA and SAMTOOLS. Multiple nucleotide polymorphisms and InDels were disregarded due to their low frequency and the low alignment speed. We used ITAG version 2.3 gene models with FASTTREE 2.1.7 software (Price *et al.*, 2010) for heuristic neighbor-joining as input to the maximum-likelihood algorithm, thus reducing the number of trees with a mix of nearest-neighbor interchange and subtree-prune-regraft. A Jukes–Cantor generalized time-reversible model, weighted neighbor joining, 100 bootstrap re-sampling, and gamma fitting for reported likelihood were used in the analysis.

Annotation of SNP calls

All VCF files from the mapped individuals were processed using SNPEFF 3.4 (Cingolani *et al.*, 2012) based on the ITAG 3.1 annotation and default parameters. SNPEFF annotates SNPs in the VCF files based on their position and the reference annotation, including their effects and report statistics such as the rates of synonymous and non-synonymous SNPs (Figures S2 and S6), heterozygosity levels (Figures 3 and S3), the number of SNPs per 1 Mbp bins (Figure S4) and the location of the SNP (Figure 4).

Introgression estimation in *S. lycopersicum*

To estimate the level of introgression in the *S. lycopersicum* species, we used the median polish procedure (Mosteller and Tukey, 1977; Xie *et al.*, 2009) on the SNP counts per 1 Mbp bin along each chromosome (Figure S4, left panel) to remove species- and bin-specific effects (species or bins naturally having a higher number of SNPs). The residuals were tested using a *z* test, and bins from crop accessions with residuals significantly different from the mean ($P < 0.05$) were labeled as introgressions (Figure S4, right panel). Note that, in wild *S. lycopersicum*, we are unable to discriminate between natural variance and inter-specific crossings.

Variant browser

JBROWSE 1.10.12 (Skinner *et al.*, 2009) was used to visualize the detected structural variants. The SL2.40 genome assembly and ITAG 2.31 genome annotation were loaded in the tomato 100+ variant browser, together with the VCF files of the 84 accessions.

Sequence repository

Sequence reads and associated analyses have been deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession numbers PRJEB5226 (*S. arcanum* LA2157), PRJEB5227 (*S. habrochaites* LYC4), PREB5228 (*S. pennellii* LA0716) and PRJEB5235.

ACKNOWLEDGEMENTS

This research was supported by the Technologisch Top Instituut Groene Genetica (TTI GG), with financial aid from the Dutch Ministry of Economic Affairs, Agriculture and Innovation, the Centre for BioSystems Genomics, and additional funding from the industrial partners listed in the affiliations section.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Genome-wide SNP counts in 84 accessions.

Figure S2. Non-synonymous and synonymous SNPs in tomato accessions and related wild species.

Figure S3. Heterozygosity level by mating type.

Figure S4. SNP counts along 12 chromosomes for 54 *S. lycopersicum* accessions.

Figure S5. Sequence distance graph for tomato accessions.

Figure S6. Distribution of non-synonymous and synonymous SNPs in FDPG genes compared to a random set of genes.

Figure S7. Accession-specific fruit color traits.

Figure S8. Accession-specific fruit shape traits.

Figure S9. Strict consensus tree based on ten FDPG gene sequences from 84 accessions.

Figure S10. Strict consensus tree based on homozygous SNPs in the exons of ten FDPG gene sequences from 84 accessions.

Table S1. Selected tomato and wild species accessions.

Table S2. *De novo* assembly statistics for three reference genomes.

REFERENCES

Adams, M.J. and Antoniw, J.F. (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res.* **34**, D382–D385.

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

Arumuganathan, K. and Earle, E. (1991) Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Rep.* **9**, 208–218.

Bai, Y. and Lindhout, P. (2007) Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Ann. Bot.* **100**, 1085–1094.

Bauchet, G. and Causse, M. (2012) Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. In *Environmental Sciences* (Çaliskan, M., ed.). Rijeka, Croatia: InTechOpen, pp. 133–162.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso–2; iso–3. *Fly (Austin)*, **6**, 80–92.

Cong, B., Barrero, L.S. and Tanksley, S.D. (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* **40**, 800–804.

Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

Donmez, N. and Brudno, M. (2013) SCARPA: scaffolding reads with practical algorithms. *Bioinformatics*, **29**, 428–434.

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

Gnerre, S., MacCallum, I., Przybylski, D. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

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

Hammer, K. (1985) Das Domestikationssyndrom. *Kulturpflanze*, **32**, 11–34.

Knapp, S. (2002) Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the *Solanaceae*. *J. Exp. Bot.* **53**, 2001–2022.

Koenig, D., Jiménez-Gómez, J.M., Kimura, S. *et al.* (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc. Natl Acad. Sci. USA*, **110**, E2655–E2662.

Ku, H.M., Doganlar, S., Chen, K.Y. and S.D. (1999) The genetic basis of pear-shaped tomato fruit. *Theor. Appl. Genet.* **9**, 844–850.

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

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–6170.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu, J., Van Eck, J., Cong, B. and Tanksley, S. (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl Acad. Sci. USA*, **99**, 13302–13306.

MacKay, T.F.C., Richards, S., Stone, E.A. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.

Miller, J.C. and Tanksley, S.D. (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *Theor. Appl. Genet.* **80**, 437–448.

Mosteller, F. and Tukey, J. (1977) *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Moyle, L.C. (2008) Ecological and evolutionary genomics in the wild tomatoes (*Solanum* sect. *Lycopersicon*). *Evolution*, **62**, 2995–3013.

Munos, S., Ranc, N., Botton, E. *et al.* (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. *Plant Physiol.* **156**, 2244–2254.

Nesbitt, T.C. and Tanksley, S.D. (2002) Comparative sequencing in the genus *Lycopersicon*: implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics*, **162**, 365–379.

Paterson, A.H., Bowers, J.E., Burow, M.D. *et al.* (2000) Comparative genomics of plant chromosomes. *Plant Cell*, **12**, 1523–1539.

Peralta, I.E., Knapp, S. and Spooner, D.M. (2005) New species of wild tomatoes (*Solanum* Section *Lycopersicon*: *Solanaceae*) from Northern Peru. *Syst. Bot.* **30**, 424–434.

Peralta, I.E., Spooner, D.M. and Knapp, S. (2008) Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect *Jugandifolia*, sect. *Lycopersicon*; *Solanaceae*). In *Systematic Botany Monographs*, Vol. 84 (Anderson, C., ed.). Ann Arbor, MI: American Society of Plant Taxonomists, pp. 1–186.

Peters, S.A., Bargsten, J.W., Szinay, D., Van de Belt, J., Visser, R.G.F., Bai, Y. and De Jong, H. (2012) Structural homology in the *Solanaceae*: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *Plant J.* **71**, 602–614.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**, 3.

Rick, C.M. (1979) Biosystematic studies in *Lycopersicon* and closely related species of *Solanum*. In *The Biology and Taxonomy of Solanacae* (Hawkes, J.G., Lester, R.N. and Skelding, A.D., eds). New York, NY: Academic Press, pp. 667–677.

Rick, C.M. (1986) Reproductive isolation in the *Lycopersicum peruvianum* complex. In *Solanaceae: Biology and Systematics* (D'Arcy, W.G.D., ed.). New York, NY: Columbia University Press, pp. 477–495.

Rodríguez, G.R., Muños, S., Anderson, C., Sim, S.C., Michel, A., Causse, M., Gardener, B.B., Francis, D. and van der Knaap, E. (2011) Distribution of *SUN*, *OVATE*, *LC*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiol.* **156**, 275–285.

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

Sim, S.C., Robbins, M.D., Van Deynze, A., Michel, A.P. and Francis, D.M. (2011) Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). *Heredity*, **106**, 927–935.

Sim, S.C., Van Deynze, A., Stoffel, K. *et al.* (2012) High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE*, **7**, e45520.

Singh, R.J. (2007) *Genetic Resources, Chromosome Engineering, and Crop Improvement. Volume 3: Vegetable Crops*. Boca Raton, FL: CRC Press.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.* **19**, 1630–1638.

Spooner, D.M., Peralta, I.E. and Knapp, S. (2005) Camparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. *Taxon*, **54**, 43–61.

Städler, T., Roselius, K. and Stephan, W. (2005) Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution*, **59**, 1268–1279.

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

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) 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.

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

Van der Beek, J.G., Verkerk, R., Zabel, P. and Lindhout, P. (1992) Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: Cf9 (resistance to *Cladosporium fulvum*) on chromosome 1. *Theor. Appl. Genet.* **84**, 1–2.

Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G. and Tanksley, S. (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell*, **14**, 1441–1456.

Van Gent, M., Bart, M.J., Van der Heide, H.G.J. *et al.* (2011) SNP-based typing: a useful tool to study *Bordetella pertussis* populations. *PLoS ONE*, **6**, e20340.

Weigel, D. and Mott, R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107.

Westesson, O., Skinner, M. and Holmes, I. (2013) Visualizing next-generation sequencing data with JBrowse. *Brief. Bioinformatics*, **14**, 172–177.

Xie, W., Chen, Y., Zhou, G., Wang, L., Zhang, C., Zhang, J., Xiao, J., Zhu, T. and Zhang, Q. (2009) Single feature polymorphisms between two rice cultivars detected using a median polish method. *Theor. Appl. Genet.* **119**, 151–164.