Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives

Susan R Strickler[1]§, Aureliano Bombarely[1], Jesse D Munkvold[1,2], Naama Menda[1], Gregory B Martin[1,3], Lukas A Mueller[1]

[1]Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853 USA

[2]Keygene Inc., 155 Gibbs St. Suite 405A, Rockville, MD 20850

[3]Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, NY 14853, USA

§Corresponding author

Email addresses:

SRS: srs57@cornell.edu

AB: ab782@cornell.edu

JDM: jesse.munkvold@keygene-inc.com

NM: nm249@cornell.edu

GBM: gbm7@cornell.edu

LAM: lam87@cornell.edu

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Abstract

Background

Studies of ancestry are difficult in tomato because it crosses with many wild relatives and species in the tomato clade have diverged very recently. As a result, the phylogeny in relation to its closest relatives remains uncertain. By using coding sequence from *Solanum lycopericum*, *S. galapagense*, *S. pimpinellifolium*, *S. corneliomuelleri*, and *S. tuberosum* and genomic sequence from two of cultivated tomato's closest relatives, *S. galapagense* and *S. pimpinellifolium*, as well as an heirloom line, *S. lycopersicum* 'Yellow Pear', we have aimed to resolve the phylogenies of these closely related species as well as identify phylogenetic discordance in the reference cultivated tomato.

Results

Divergence date estimates suggest divergence of *S. lycopersicum*, *S. galapagense*, and *S. pimpinellifolium* happened less than 0.5 MYA. Phylogenies based on 8,857 coding sequences support grouping of *S. lycopersicum* and *S. galapagense*, although two secondary trees are also highly represented. A total of 25 genes in our analysis showed evidence of selection along the *S. lycopersicum* lineage. Whole genome phylogenies showed that while incongruence is prevalent in genomic comparisons between these accessions, likely as a result of incomplete lineage sorting and introgression, a primary phylogenetic history was strongly supported.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Conclusions

Based on analysis of these accessions, *S. galapagense* appears to be closely related to *S. lycopersicum*, suggesting they had a common ancestor prior to the arrival of an *S. galapagense* ancestor to the Galápagos Islands, but after divergence of the sequenced *S. pimpinellifolium.* Genes showing selection along the *S. lycopersicum* lineage may be important in domestication. Further analysis of intraspecific data in these species will help to establish the evolutionary history of cultivated tomato. The use of an heirloom line is helpful in deducing true phylogenetic information of *S. lycopersicum* and identifying regions of introgression from wild species.

Keywords

phylogeny, tomato, Solanum, genome, incomplete lineage sorting, introgression, selection

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

PeerJ PrePrints

Background

Identifying and exploiting diversity present in wild tomato species has been crucial for the improvement of production traits in cultivated tomato [1]. Useful traits, such as ease of harvest, shelf life, pathogen resistance, and abiotic stress tolerance have been introduced through introgressions from wild species. The wild tomatoes, along with the cultivated tomato, *Solanum lycopersicum,* comprise the 13 members of *Solanum* sect. *Lycopersicon*, and are native to western South America. All members of the clade can be crossed to cultivated tomato with varying degrees of ease [1] and breeding programs for cultivated tomato have widely utilized this property since the 1940s [1], allowing for the introgression of traits desirable in fruit production. Interestingly, most wild species are green-fruited except for three, *S. pimpinellifolium, S. galapagense*, and *S. cheesmaniae*. These species are also thought to be the closest relatives to cultivated tomato [2,3].

S. pimpinellifolium is native to areas of low elevation on the western slopes of the Andes in Peru and Equador [1]. It is the proposed nearest wild relative to cultivated tomato [2] and is the only red-fruited wild species. *S. pimpinellifolium* has been used to introduce traits such as disease resistance and improved soluble solids into cultivated tomato [1]. The other two wild species, *Solanum galapagense* along with the closely-related *Solanum cheesmaniae,* are perennials endemic to the Galápagos Islands and comprise the only two orange-fruited tomato species. *S. galapagense* was only recently recognized as a separate species from *S. cheesmaniae.* It was previously classified as *S. lycopersicon cheesmaniae* L. Riley var. *minor* (Hook.f) [3] and there is debate based on genetic variation between the species that questions whether they should

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

be classified as morphotypes rather than separate species [4]. *S. galapagense* and *S. cheesmaniae* have been used to a limited degree in breeding programs, mainly to improve salt tolerance and soluble solids [1]. Orange fruit color in these two species is due to a dominant variant of the *B* gene that results in 5 to 10-fold increase in β-carotene in comparison to red fruit [5]. These species have other phenotypic differences from cultivated tomato including scent, pathogen response, trichomes, and leaf morphology [6].

The relative phylogenetic positions of *S. lycopersicum, S. galapagense, S. cheesmaniae,* and *S. pimpinellifolium* are currently unresolved [1]. Several different tree topologies have been inferred for the species in recent literature using various methods [1]. In recently diverged species such as these, phylogenetic discordance can be prevalent [7], due to both incomplete lineage sorting of ancestral polymorphism and introgression from other species. Introgression is expected to make an especially strong contribution to phylogenetic discordance in tomato species, due to the use of wild species in the development of various *S. lycopersicum* cultivars. In particular, the sequenced tomato, *S. lycopersicum* 'Heinz 1706' (H1706) is known to have *S. pimpinellifolium* in its parentage [2,8,9]. Interspecific hybridization also occurs in wild populations of tomato along hybrid zones [10] and also as evidenced by *S. lycopersicum* var *cerasiforme*, which is purportedly the result of crossing between *S. lycopersicum* and *S. pimpinellifolium* [11]. In contrast, *S. galapagense* and *S. cheesmaniae* have likely evolved in relative isolation, although *S. pimpinellifolium* and *S. lycopersicum* have been introduced to the Galápagos Islands in the past few decades [6]. Heirloom lines, which have existed prior to the implementation of major breeding programs, have been perpetuated mainly from lines of *S.*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

lycopersicum often by home gardeners decreasing their likelihood of containing introgressions from wild species.

For this study, we have sequenced *S. galapagense* and the heirloom line *Solanum lycopersicum* 'Yellow Pear' (YP-1) [12]. Given the close relationship between *S. galapagense* and *S. cheesmaniae, S. galapagense* was chosen as a representative sample for the current study. These data were used in conjunction with coding sequence data from *S. pimpinellifolium* [2], *S. corneliomuelleri* [13], and *S. tuberosum* [14]. Positively selected genes along the *S. lycopersicum* lineage were of interest since they may relate to domestication phenotypes. Using whole genome sequence from YP-1, *S. galapagense*, *S. pimpinellifolium, S. tuberosum*, and two types of genome sequence from H1706 [2], regions of divergence from cultivated tomato including structural variation were identified and the placement of *S. galapagense* on the *Solanum* phylogenic tree was resolved for these accessions. Also, a survey of genomic discordance was performed to gain a greater understanding of phylogenetic incongruence in newly diverged plant species. *S. lycopersicum* YP-1, a heirloom line that predates major tomato breeding programs was included as a negative control for introgressions from wild tomato species. All data are available at the Sol Genomics Network site (http://solgenomics.net/) [15].

Results

Assembly statistics

Quality filtering and trimming of the paired-end reads yielded 462.7 million *S. lycopersicum* H1706, 466.3 million *S. lycopersicum* YP-1 reads, 363.9 million *S. galapagense* reads, and 281.5 million *S. pimpinellifolium* (Table 1). Approximately 92.1% of the *S. lycopersicum* H1706 reads,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

93.5% of the *S. lycopersicum* YP-1, 89% of the *S. galapagense*, 88% of the *S. pimpinellifolium* mapped to the *S. lycopersicum* version 2.40 genome assembly giving 39x, 45x, 32x, and 25x coverage and covering 99.2%, 99.3%, 95.4%, and 95% of the tomato genome respectively, after mapping quality filtering and duplicate read removal (Additional file 1). Gaps were calculated as regions without read coverage that were not gaps in the *S. lycopersicum* H1706 scaffolds. The H1706 assembly had 76,276 gaps totaling 5.9 megabase pairs (Mbp) and the YP-1 assembly had 51,980 gaps totaling 5.4 Mbp of sequence. A total of 227,699 gaps were found in the *S. galapagense* assembly, totaling 36.1 Mbp (Additional file 1). The *S. pimpinellifolium* assembly had 209,919 gaps totaling 38.9 Mbp of sequence.

In addition, *de novo* assemblies were produced for each non-reference genome. By comparing assemblies generated from a range of k-mer values, the best k-mer values were found to be 63, 57, and 51, for YP-1, *S. galapagense*, and *S. pimpinellifolium* respectively. Contigs greater than 200 bp were used for further analysis. The YP-1 assembly produced the largest contigs with an N50 of 25.2 kb totaling 716.7 Mbp of sequence while *S. pimpinellifolium* had the shortest with an N50 of 5 kb totaling 669.3 Mbp of sequence (Additional file 2).

SNP and indel detection and effect on the genome

Over 500,000 single nucleotide polymorphisms (SNPs) were found between YP-1 and H1706 (Additional file 3). *S. galapagense* was found to have approximately 4.7 million SNPs whereas *S. pimpinellifolium* had 6 million when compared to H1706 (Additional file 3). Variation in SNP density was found across the genome and was found to differ between chromosomes and accessions (Figure 1 and Additional file 4). In particular, regions on chromosomes 4 (~59 Mbp)

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

and 11 (~4 Mbp) show reduced SNP density in *S. pimpinellifolium* and elevated density in YP-1 (Figure 1 and Additional file 4. A large assembly coverage gap in *S. pimpinellifolium* located at approximately 11 Mbp on chromosome 1 is found at the position of the tomato self-incompatibility locus [16] (Additional file 4). Large assembly coverage gaps were also detected in *S. pimpinellifolium* on chromosomes 3 (~37 Mbp), 8 (~40 Mbb), 10 (~30 Mbp) , and *S. galapagense* chromosomes 8 (~16 Mbp), and 12 (~60 Mbp) (Figure 1 and Additional file 4). As expected, more SNPs were found in noncoding regions than coding regions (Table 5a). SNPs were found in approximately 0.05%, 0.5%, and 0.8% of the YP-1, *S. galapagense*, and *S. pimpinellifolium* genomes respectively, while affecting only 0.04%, 0.3%, and 0.4% of the coding regions of these genomes (Additional file 3). A total of 3,418 YP-1, 20,447 *S. galapagense,* and 12,143 *S. pimpinellifolium* genes were found to have nonsynonymous SNPs associated with them. Additionally, 242,165 SNPs were identified using the aligned Illumina data from H1706 to the reference H1706 v 2.40 assembly of which 225,625 were predicted to be heterozygous with the reference genome (Additional file 5).

Approximately 50,000 indels were found between YP-1 and H1706, 350,000 between *S. galapagense* and H1706, and 520,000 between *S. pimpinellifolium* and *S. lycopersicum* H1706 (Additional file 6). Indels were more prevalent in noncoding regions (Additional file 6). Indels in coding sequence were found in a total of 595 YP-1 genes, 3,493 *S. galapagense* genes, and 3,645 *S. pimpinellifolium* genes. Additionally, 41,776 indels were identified between the H1706 sequence and H1706 v 2.40 (Additional file 5), 4,716 of which were heterozygous.

Structural variation

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

To determine the nature of regions where reads from YP-1, *S. galapagense*, or *S. pimpinellifolium* could not map to the H1706 genome, but H1706 reads could map, these regions were further analyzed for each species. Regions lacking coverage in the H1706 mapping are likely repetitive regions or regions containing an error in the reference genome. Gap size distribution was similar between *S. galapagense* and *S. pimpinellifolium* with less gaps found in YP-1 (Figure 2), with all accessions having a peak at 90 bp. Since gaps could be missing regions or divergent regions where short reads cannot map, *de novo* contigs assembled from the wild and heirloom species reads were mapped to the reference genome to determine if they covered gap regions. Approximately 3.3% of YP-1, 3.7% of *S. galapagense*, and 6.0% of *S. pimpinellifolium* contigs did not map with greater than 90% id. A small number of these contigs contained many repeats or matched plastid, mitochondrial, or vector DNA (Additional file 7). After removal of gaps covered by *de novo* contigs, a total of 2.4 Mbp of YP-1, 13.8 Mbp of *S. galapagense*, and 21.6 Mbp of *S. pimpinellifolium* was putatively deleted relative to H1706. The largest gap in each species was 12.7 kbp on chromosome 12 for YP-1, 41 kb on chromosome 12 of *S. galapagense,* and 38.7 kbp on chromosome 10 of *S. pimpinellifolium* (Additional file 8). Deleted genes were determined as genes that were at least 90% contained in putative gaps and had no matches in de novo contig assemblies. A total of 13 genes from YP-1, 87 genes in *S. galapagense*, and 157 in *S. pimpinellifolium* were found to have no coverage in either the small read mapping or contig mapping (Additional file 9). Many of these genes were classified as disease resistance-related proteins or lacked a predicted function (Additional file 9).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Two small insertions of 130 bp were predicted in *S. pimpinellifolium* in reference to H1706 on chromosomes 4 and 10, but these were not well supported (Additional file 10). No insertions larger than 20 bp could be predicted in the other accessions relative to H1706.

Patterns of gene evolution in *Solanum*

To determine the average nucleotide substitution rate amongst coding sequences, aligned sequence from 32,982 *S. galapagense* genes and 32,795 *S. pimpinellifolium* genes was used to generate estimates of nonsynonymous (dN) and synonymous (dS) substitution rates in reference to YP-1 (Table 1). H1706 was not considered in the analysis since introgression from wild species could bias the analysis. Missing genes or genes containing stop codons were removed from the analysis. *S. pimpinellifolium* had a larger dS than *S. galapagense* (Table 1). The number of synonymous substitutions per synonymous site ranged from 0 to 0.5655 for *S. galapagense*, and 0 to 0.3403 for *S. pimpinellifolium* (Table 1). Nonsynonymous substitutions per nonsynonymous site ranged from 0 to 0.2106 in *S.galapagense* and 0 to 0.1105 in *S. pimpinellifolium* (Table 1).

Coding sequence from 8,857 orthologous genes that could be aligned with confidence between YP-1, *S. galapagense, S. pimpinellifolium, S. corneliomuelleri,* and *S. tuberosum* were analyzed to infer gene tree topology using maximum likelihood. The majority of trees (3,611) supported tree topology 1 which groups *S. lycopersicum* and *S. galapagense,* suggesting these two species may be more closely related, although two other tree topologies were also well supported, albeit to a lesser degree (2,344 and 2,037 trees) (Figure 3). The genes were then

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

subjected to site-branch selection tests along the *S. lycopersicum* lineage. Stop codons were found in at least one of the species for 288 genes and these were removed from further analysis. A total of 25 genes showed evidence of a faster rate of evolution along the *S. lycopersicum* lineage (Additional file 11). Many of these genes have predicted function in adaptive or domestication phenotypes such as pathogen and abiotic stress response, cell division, and carbohydrate metabolism (Additional file 11).

Species divergence time estimates calculated based on dS values from 3,611 genes fitting topology 1 suggest a divergence estimate for *S. lycopersicum* and *S. pimpinellifolium* of 0.44 MYA (Table 2). Using coalescence-based divergence estimates of 8 genes fitting topology 1, a similar divergence of 0.45 was obtained (Table 2), although hybridizations signatures were apparent between the species (Additional file 11). These signatures support the hypothesis of recent hybridizations between the following groups: 1) *S. lycopersicum* and *S. galapagense*, 2) *S. lycopersicum* and *S. pimpinellifollium,* and 3) *S. galapagense* and *S. pimpinellifollium*. A more recent divergence of 0.19 MYA, using dS values, and 0.25 ± 0.01, using the coalescence method, was estimated for *S. lycopersicum* and *S. galapagense* (Table 2). Our results are similar to a previous estimate of ~1 MYA which was based on a smaller gene sample size [17].

Genomic phylogenetic discordance

To look at genome wide phylogenetic discordance, whole genome alignments were created with H1706, YP-1, *S. galapagense, S. pimpinellifolium,* and *S. tuberosum.* A total of 781.5 Mbp of the H1706 genome was represented in the alignment. The alignments were then partitioned using the minimum description length (MDL) principle [7] resulting in 7,828 loci, and an

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

average size of 201.8 kbp for a total of 784 mb, since some partitions overlapped. The partitioning was based on 1 - 100 kbp windows.

Trees for each genome partition were constructed using Bayesian phylogenetic analysis. A total of 4,541 loci covering 53% of the H1706 genome supported topology 1 with a posterior probability of 0.9 or greater (Figure 4, Additional file 12) grouping *S. galapagense* closer to *S. lycopersicum* than to *S. pimpinellifolium,* (Figure 4 and Additional file 12). Topology 3, which clusters the two *S. lycopersicum* accessions more closely to *S. pimpinellifolium*, was found with a posterior probability of 0.9 or greater at 191 loci and covered 6% of the H1706 genome (Figure 4 and Additional file 12). Overall, the predominant tree topology was topology 1 which was the best supported topology at 72% of loci. Topology 2 was the second most prevalent tree and supported at 17% while topology 3 was found at 13% of the genome.

A total of 190 loci consisting of 2.8 mb constituting 0.4% of the H1706 genome best supported topology 5 indicative of introgression in H1706, placing *S. pimpinellifolium* closer to H1706 than YP-1. This includes an introgression of 141 kb on chromosome 9 containing 67 genes which contain *Ve1* and *Ve2*, involved in Verticillium wilt resistance [18] and a 1.1 mb region on chromosome 11 with 109 predicted gene models that is associated with the *I* gene, which confers resistance to Fusarium wilt [19] (Additional files 12 and 13). An additional 1.6 mb region on chromosome 4 containing 216 gene models was found to be introgressed in H1706, although, to date, no known disease resistance genes are found here (Figure 4 and Additional file 13).

By using functional predictions for the gene models predicted within the chromosome 11 introgression, four TIR-NB-LRR resistance proteins were identified (Additional file 13). Since

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

I2 is known to be a protein of this type, these genes are likely candidates for *I* [19]. One of the candidates, Solyc11g011080, was found to have a frameshift mutation and possible splice site mutation in YP-1, while retaining the H1706 reading frame in *S. pimpinellifolium*.

Discussion

Here, we present two new genome assemblies: the wild tomato species, *S. galapagense,* and an heirloom variety, *S. lycopersicum* 'Yellow Pear'. We determined variation by comparing these two assemblies, as well as the published assemblies of *S. pimpinellifolium* and the reference H1706 genome. While a difference in SNP count was found between this study and a previous study for *S. pimpinellifolium* [2], the same SNP calling pipeline was used for all accessions in this study, so estimates of variations across species should not be biased. The homozygous SNPs and indels, which were identified by mapping reads from H1706 to the reference genome from this accession, are likely errors in the reference sequence. Heterozygous sites in H1706 were identified since the reference is based on a collapsed chromosome. It is also possible that 454 sequencing used in the reference assembly introduced indel errors. Slightly more gaps were found in the H1706 reference-guided assembly than in the YP-1 assembly, which could be related to newer technologies used for library preparation for the sequencing of YP-1. Putative divergent regions in the assemblies are likely not repetitive regions or other regions where reads map poorly to the reference genome, since these regions would have been removed from further analysis based on gaps in the H1706 reference-guided assembly. Large insertions could not be predicted with accuracy, likely due to the use of only short insert size paired-end libraries. Based on the total length of the de novo assemblies and divergent regions, it is likely that the genome

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

size of *S. galapagense* is comparable to *S. lycopersicum*, although earlier studies suggested a possibly smaller genome [20]. *S. pimpinellifolium* may have a smaller genome than H1706, based on gap sizes and kmer assessment, although further sequencing is necessary to obtain a higher coverage for this analysis.

Omega values tend to decrease with presumed evolutionary distance in closely related species [21] and a similar result was obtained based on the coding sequence analysis of *S. lycopersiucm*, *S. galapagense*, *S. pimpinellifolium*, *S. corneliomuelleri*, and *S. tuberosum*. Since omega is a ratio derived from scaling dN by dS, omega values can be artificially inflated if synonymous mutations are not neutral, and also as a factor of short branch length. The latter is a likely explanation for our results, as species in the tomato clade have a very recent divergence. For example, one of the more divergent wild tomato species, *S. pennellii,* has an estimated divergence from *S. lycopersicum* of only 7 MYA [17]. Our results suggest a more recent divergence of tomato from its closest wild relatives, giving further evidence of short branch length. Our results are in agreement with a previous estimate of ~1 MYA based on a smaller sample size of genes [17]. Interestingly, a much larger number of *S. galapagense* genes are affected by nonsynonymous substitutions, likely due to the fixation of slightly deleterious alleles due to drift acting strongly on a small population size, during initial colonization of the Galápagos Islands. Nonsynonymous substitutions per nonsynonymous site has a larger range in *S. galapagense* also likely due to drift2

Only 25 genes were detected in this study that are candidates for selection along the *S. lycopersicum* lineage. Many factors relevant to this data set likely play a role in these results. Genes that are more divergent may also be more rapidly evolving as seen in the results of all

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

coding sequence analysis versus the subset analysis (Table 1). The reduced gene set only includes genes with putative orthologs that fit strict criteria and contained matches in all species studied. In particular, the *S. corneliomuelleri* transcriptome dataset contained only 50% of the total number of expected genes based on tomato annotations. As a result, this data set is likely a biased sample including predominantly genes that are more conserved across Solanum species. Indeed, a study analyzing a larger sample of 11,221 genes, found a total of 51 genes to be positively selected [22]. Moreover, the short branch length of members of the tomato clade impedes detection of differential rates of evolution. An average dS of 0.05 is necessary for detection of lineage-specific selection, meaning there is little statistical power to detect selection along a lineage in this group [23]. In the near future, availability of sequence of more divergent Solanaceae members may improve the ability to detect differential selection rates, albeit not within the tomato clade. To detect selection within the tomato clade, alternative selection detection methods, such as McDonald-Kreitman tests [24] involving intraspecific data may prove useful, as well as larger sample size of genes.

Whole genome phylogenies proved useful to detect topological discordance in these recently diverged plant species. Since a greater number of SNPs occur in non-coding genomic regions, higher phylogenetic signal may be achieved with genomic alignments, rather than only coding sequence. In our study, most regions of the H1706 genome where phylogenies do not fit the majority rules species tree, did fit a pattern expected from incomplete lineage sorting, for example, grouping both *S. lycopersicum* accessions closer to *S. pimpinellifolium*. Incomplete lineage sorting is also supported by the nearly equal frequency of two secondary trees in the coding sequence phylogenies [25]. These results are expected when speciation has occurred in a

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

short period of time from an ancestral population with greater diversity, which is likely the case in this study. A secondary cause of phylogenetic discordance was found in regions of introgression from *S. pimpinellifolium* in the H1706 genome and could be ascertained by the inclusion of the heirloom YP-1 as a control. Genome-wide phylogenies, as well as SNP density patterns on chromosomes 4, 9, and 11, support introgression of a *S. pimpinellifolium* in the H1706 genome. Additionally, an overlapping region on chromosome 4 was found in comparisons to an inbred line to H1706 further supporting an introgression in H1706 at this location [26]. These regions are in concordance with previous introgression predictions [2] and the known H1706 pedigree [8]. There are also several regions of high SNP density across the chromosomes that do not correspond to regions identified as introgressions in the tree topologies, suggesting these are regions of high variability. For example, in *S. pimpinellifolium* chromosome 1 has a region of high SNP density on either side of the self-incompatibility locus. Self-incompatibility loci are known to exhibit high polymorphism and rearrangement [27] and this is evidenced by the lack of read coverage in the immediate area between *S. pimpinellifolium* and H1706. This result would also suggest the sequenced *S. pimpinellifolium* has a self-incompatibility haplotype that is very different from H1706, YP-1, and *S. galapagense*. While H1706 is known to have *S. pimpinellifolium* in its parentage, the specific accession analyzed here may be different than the H1706 parental accession.

Despite extensive phylogenetic discordance, by using coding sequence and whole genome sequence data, we were able to ascertain a predominant species tree for the accessions in this study. *S. galapagense* is more closely related to *S. lycopersicum* than *S. pimpinellifolium* as supported in some previous studies, one of which includes a different *S. galapagense* accession

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

than what was used in our study [1,22]. It is possible by sampling the spectrum of variation in *S. pimpinellifolium* a more closely related accession to cultivated tomato may be found. Sequencing of additional accessions of these species will help shed light on the evolution of domesticated tomato.

Conclusions

H1706 provides an excellent reference for genome assembly of its nearest wild relatives and allow for efficient genome analysis. Using this reference genome, we have determined areas of variation across closely related tomato species and found candidate genes involved in domestication. Genome-wide phylogenies support this *S. galapagense* accession as the closest wild relative of cultivated tomato in our study. The sequenced tomato is expected to have wild introgressions and we have successfully delimited candidate introgression regions from wild species. This method may also be useful in detecting candidate regions for breeding purposes as well as conservation biology, since wild species may be threatened due to introgression from cultivated tomatoes [3,6].

Methods

Solanum lines and libraries

S. galapagense accession LA0436 was obtained from the Tomato Genetic Resource Center (TGRC; http://tgrc.ucdavis.edu/) and *S. lycopersicum* 'Yellow Pear' (YP-1) was obtained from the Martin Lab. Genomic DNA was prepared using a modified version of a protocol described previously [28] using precipitation and CsCl purification instead of agarose bead imbedding.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Samples for *S. galapagense* were sent to the Life Science Core Laboratory Center at Cornell University (Ithaca, NY) for library preparation and sequencing. YP-1 was sent to Genomics Resources Core facility at Weill Cornell Medical College (New York, NY) for library preparation and sequencing. *S. pimpinellifolium* accession LA1589 and *S. corneliomuelleri* accession LA0103 were sequenced by the Lippman Lab at Cold Spring Harbor [2,11]. H1706 and *Solanum tubersosum* sequence is publicly available [2,13]. H1706 Illumina paired-end data from libraries 090617, 090619, and 090701_SNPSTER5B was provided by Syngenta.

llumina sequencing

Sheared genomic DNA from *S. galapagense* was run on 2 lanes of an Illumina HiSeq 2000. Read length was 100 base pairs (bp) and insert size was 200 bp. In addition, sheared genomic DNA was run on 7 lanes of an Illumina GA II using the mate pair module. Genomic DNA from YP-1 was run on 1 lane of an Illumina HiSeq 2000 and the resulting sequence was 100 bp in length with an insert size of 300 base pairs. *S. galapagense* and YP-1 sequence was submitted to the NCBI Small Read Archive (SRA) as experiment numbers SRX520161 and SRX521582. Data and output from this study can be accessed through Solgenomics at ftp://ftp.solgenomics.net/genomes/.

Sequence assembly

Reads were inspected for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and rechecked after cleaning. Cleaning was performed with fastq-mcf (https://code.google.com/p/ea-utils/wiki/FastqMcf). Reads were mapped to the H1706 reference

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

assembly v 2.40 using a tiered approach with an initial round of BWA [29] mapping followed by Novoalign [30] for the remaining discordant and unmapped reads. Duplicate reads and reads with a mapping quality less than 30 were removed for variation analysis with Picard (http://picard.sourceforge.net) and Samtools [31] respectively.

Whole genome *de novo* assemblies of *S. galapagense*, *S. pimpinellifolium,* and YP-1 were created using SOAP *de novo* version 1.05 [32]. Assemblies were produced using a kmer range between 25 and 63. Scripts supplied with the SOAP *de novo* package were used for error correction and gap filling of the scaffolds. Reads that did not map or did not pair properly in the reference-guided assembly were mapped to the *de novo* assembled contigs for each genome. Contigs that had an above average number of reads mapped to them were further analyzed (see next section).

Variation discovery

SNPs and indels 15 base pairs and smaller were detected using the GATK recommended best practices [33]. Since a suitable dataset was not available for base quality calibration, one was generated by pooling high quality SNPs from both *S. galapagense* and *S. pimpinellifolium*. Snpeff was used to determine the effect of each SNP and indel in the genome and determine zygosity [34]. Putative deleted regions were detected by finding regions that had no sequence coverage and did not overlap with gaps in the reference assembly using Bedtools [35]. Only gaps greater than 15 base pairs and not found on chromosome 0 were used for further analysis. These regions were compared to the mapping assembly of H1706 and matching gaps were removed from further analysis. BLAT [36] with default values (sequence identity 90%) was used

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

to map *de novo* assembled contigs greater than 200 bp from each accession to the reference genome. The best hit was determined by using scripts included with the BLAT package. Unmapped contigs were processed by Seqclean [37] to identify matches to *S. lycopersicum* plastid or mitochondrial DNA [2], plant pathogen sequence found in Comprehensive Phytopathogen Genomics Resource (CPGR) [38], vector sequence found in UniVec database [39], or contigs that were low complexity. Putative deletions were confirmed if *de novo* assembled contigs did not map to regions not covered in the reference-guided assemblies. Bedtools [35] was used to identify genes found at least 90% in deleted regions. Blat [36] was used to search for orthologs of these genes in the *de novo* assemblies. Genes with hits covering less than 50% of the gene and not the top match in reciprocal BLAT [36] output were considered deleted. Breakdancer v1.1 [40] was used to predict insertions greater than 15 base pairs for insertion analysis.

Coding sequence analysis

Predicted coding sequence from *S. galapagense,* and *S. pimpinellifolium* was used for pairwise comparisons to YP-1. Only genes with no stop codons predicted within the gene sequence were used. Coding sequence was predicted using H1706 annotation version ITAG2.3 [2]. Coding regions were first reverse translated and aligned using ClustalW [41]. Alignments containing premature stop codons were discarded. Pairwise maximum likelihood comparisons were performed to determine nonsynonymous and synonymous substitution rates using the codeml package of PAML version 4.5 [23].

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Predicted coding sequence of genes from YP-1, *S. galapagense*, *S. pimpinellifolium*, *S. corneliomuelleri*, and *S. tuberosum* were subjected to phylogenetic analysis. Coding sequence with at least 50% *S. lycopersicum* gene coverage was selected as input. BLAST [42] was used to find putative *S. lycopersicum* orthologs in *S. tuberosum* coding sequence. These matches were then used as a query for a reciprocal BLAST [42] to the *S. lycopersicum* genome. Any hits that were not one-to-one matches were discarded. Alignments were calculated as above. The underlying phylogeny was calculated for each gene using DNAml with the Kimura model and 100 bootstrap replicates using PhygOmics [43]. Pairwise estimates of ω were calculated using the codeml package of PAML [23]. Codeml [23] was also used to perform a branch-site test to detect positive selection along the *S. lycopersicum* lineage. The maximum likelihood value from the alternative model allowing sites to evolve under positive selection was compared to the value from the null model in which no selection occurs. The null model was rejected if 2 times the difference between the log likelihood values was larger than 2.71 at the 5% significance level.

Divergence dating was estimated by assuming a nuclear gene substitution rate of 6.03 x 10-9 dS per site per year and dividing dS by 2 times the substitution rate [17]. These estimates were compared to coalescent-based estimates using *BEAST [44]. Only genes fitting the predominant gene tree topology were used in the calculations. Eight gene clusters were used for this analysis (homologous genes to the reference gene models: Solyc02g081560, Solyc02g093130, Solyc04g054810, Solyc04g078200, Solyc05g010810, Solyc06g009630, Solyc09g013140, Solyc11g069330). Based in the small divergence between species a conservative substitution model was chosen, JC. Monte Carlo Markov Chains (MCMC) of

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

100,000 generations were used to perform this analysis. DensiTree [45] was used to visualize tree set output.

Whole Genome phylogeny

Genomes for YP-1, *S. galapagense,* and *S. pimpinellifolium* were created by substituting SNPs and masking gaps in coverage into the reference assembly. Repeat masking was performed using RepeatMasker [46] and a tomato-specific repeat dataset [2]. Whole genome multiple sequence alignment were generated for H1706*,* YP-1, *S. galapagense*, *S. pimpinellifolium*, and *S. tuberosum* using Mercator and Mavid [47]. Potential recombination breakpoints were deduced using PAUP [48] as implemented through the MDL program as described in a previous study, using 1000 bp partitions as a starting point [7]. MrBayes [48] was run for 2,000,000 generations on each partition with 1 hot and 1 cold chain. Tree locations were mapped to H1706 genomic coordinates.

List of abbreviations

YP-1: Yellow Pear; H1706: Heinz 1706; Mbp: megabase pairs; SNPs: single nucleotide polymorphisms; dN; nonsynonymous substitutions per nonsynonymous site; dS: synonymous substitutions per synonymous site; MDL: minimum description length; TGRC: Tomato Genetic Resource Center; bp: base pairs; SRA: Small Read Archive; CPGR: Comprehensive Phytopathogen Genomics Resource; MCMC: Monte Carlo Markov Chains.

Competing interests

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014
PeerJ PrePrints

The authors declare that they have no competing interests.

Authors' contributions

SRS conceived and designed the study, carried out the comparative genomic and phylogenetic analyses, and drafted the manuscript. AB assisted in writing, experimental design, and provided scripting assistance. JDM and GBM conceived the idea of *S. galapagense* sequencing, grew *S. galapagense* plants and provided genomic DNA. NM provided scripting assistance. LAM assisted with drafting the manuscript. All authors read and approved the final manuscript.

Description of additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a figure of SNPs and read coverage over all chromosomes for YP-1, *S. galapagense*, and *S. pimpinellifolium*. Additional data file 2 is a table of putative gaps greater than 20 bp in YP-1, *S. galapagense*, and *S. pimpinellifolium*. Additional data file 3 is a table of putatively missing genes in YP-1, *S. galapagense*, and *S. pimpinellifolium.* Additional data file 4 is a table of putative insertions in *S. pimpinellifolium*. Additional data file 5 is a vcf of H1706 SNPS and indels. Additional data file 6 is a graph of phylogenetic topologies over all other chromosomes. Additional file 7 is de novo assembly mapping results. Additional file 8 is a list of putative gaps in sequenced accessions. Additional file 9 is a list of deleted genes and predicted functions. Additional file 10 is putative insertions in LA1589. Additional file 11 is PAML site-branch test analysis results. Additional file 12 is Beast results. Additional file 13 is whole genome

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

phylogeny results for all chromosomes. Additional file 14 lists genes predicted to be in introgressions in H1706 genome.

Acknowledgments

We thank Diane M. Dunham for growing YP-1 plants, Julia Vrebalov for isolating genomic DNA for YP-1 sequencing, and Joanne Labate for critical review of the manuscript. Seed material for *S. galapagense* was developed by and/or obtained from the UC Davis/C.M. Rick Tomato Genetics Resource Center and maintained by the Department of Plant Sciences, University of California, Davis, CA 95616.

Funding

This study was supported by startup funds from the Boyce Thompson Institute for Plant Research (LAM) and by National Science Foundation grant IOS-1025642 (GBM).

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

References

1. Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta I, Cammareri M, Perez P, Termolino P, Chiusano ML, Ercolano MR, Frusciante L, Monti L, Pignone D: From *Solanum* sect. Lycopersicon. In *Wild Crop Relatives: Genomic and Breeding Resources. Vegetables* 1st edition. Edited by Kole C. Berlin; Heidelberg; New York: Springer; 2011:129-215.
2. The Tomato Genome Consortium: The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012, 485:635-641.
3. Darwin SC, Knapp S, Peralta IE: Taxonomy of tomatoes in the Galápagos Islands: native and introduced species of *Solanum* section *Lycopersicon* (Solanaceae). *Systematics and Biodiversity* 2003, 29–53.
4. Lucatti AF, van Heusden AW, de Vos RCH, Visser, RGF, Vosman B: Differences in insect resistance between tomato species endemic to the Galapagos Islands. *BMC Evolutionary Biology* 2013, 13:175.
5. Ronen G, Cohen M, Zamir D, Hirschberg J: Regulation of carotenoid biosynthesis during tomato fruit development: expression of the gene for lycopene epsilon-cyclase is down-regulated during ripening and is elevated in the mutant Delta. *The Plant Journal* 2002, 17:341-351.
6. Darwin SC: The systematics and genetics of tomatoes on the Galápagos Islands (*Solanum*, Solanaceae). *PhD thesis* University College London, Department of Genetics, Evolution and Environment; 2009.
7. Ané C: Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biology and Evolution* 2011, 3:246–258.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

8. Ozminkowski R: Pedigree of variety Heinz 1706. *Report of the Tomato Genetics Cooperative* 2004, 54:26.

9. Labate JA, Robertson LD: Evidence of cryptic introgression in tomato (Solanum *lycopersicum* L.) based on wild tomato species alleles. *BMC Plant Biology* 2012, 12:133.

10. Nakazato T, Housworth HA: Spatial genetics of wild tomato speices reveals roles of the Andean geography on demographic history. *American Journal of Botany* 2010, 98:88-98.

11. Ranc N, Muños S, Santoni S, Causse M: A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (Solanaceae). *BMC Plant Biology* 2008, 8:130.

12. Goldman, A: The Heirloom Tomato: From Garden to Table: Recipes, Portraits, and History of the World's Most Beautiful Fruit. Bloomsbury USA; 1st edition (August 5, 2008).

13. Park SJ, Jiang K, Schatz MC, Lippman ZB: Rate of meristem maturation determines inflorescence architecture in tomato. *PNAS* 2012, 109:639-644.

14. Potato Genome Sequencing Consortium: Genome sequence and analysis of the tuber crop potato. *Nature* 2011, 475:189-195.

15. Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA: The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Research* 2011, 39:D1149–55.

16. Tanksley SD, Loaiza-Figueroa F: Gametophytic self-incompatibility is controlled by a single major locus on chromosome 1 in *Lycopersicon peruvianum. PNAS* 1985, 82:5093-5096.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

PeerJ PrePrints

17.Nesbitt, T, Tanksley, SD: Comparative sequencing in the genus Lycopersicon: Implications for the Evolution of Fruit Size in the domestication of cultivated tomato. *Genetics* 2002, 162:365-379.

18.Kawchuk LM, Hachey J, Lynch DR, Kulcsar F, van Rooijen G, Waterer DR, Robertson A, Kokko E, Byers R, Howard RJ, Fischer R, Prüfer D: Tomato Ve disease resistance genes encode cell surface-like receptors. *PNAS* 2001, 98:6511-6515.

19.Scott JW, Agrama HA, Jones JP: RFLP-based analysis of recombination among resistance genes to Fusarium wilt races 1, 2, and 3 in tomato. *J. Amer. Soc. Hort. Sci.* 2004, 129:394-400.

20.Arumuganathan K, Earle ED: Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* 1991, 9:208–218.

21.Wolf BW, Künstner A, Nam K, Jakobsson M, Ellegren H: Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection. *Genome Biology and Evolution* 2009, 1:308–319.

22.Koenig D, Jiménez-Gómez JM, Kimura S, Fulop D, Chitwoods DH, Leadlands LR, Kumar R, Covington MF, Devisetty UK, Tat AV, Tohge T, Bolger A, Schneeberger K, Ossowski S, Lanz C, Xiong G, Taylor-Teeples M, Brady SM, Pauly M, Weigel D, Usadel B, Fernie AR, Peng J, Sinha NR, Maloof JN: Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *PNAS* 2013, 110:E2655-E2662.

23.Yang Z: PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 2007, 24:1586-1591.

24.McDonald J H, Kreitman M: Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 1991 351:652-654.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

PeerJ PrePrints

25.Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: Ancient Admixture in Human History. *Genetics* 2012, 192:1064-1093.

26.Menda N, Strickler SR, Edwards JD, Bombarely A, Dunham DM, Martin GB, Mejia L, Hutton SF, Havey MJ, Maxwell DP, Mueller LA. *The Plant Journal* 2013, submitted.

27.Wang Y, Wang X, Skirpan AL, Kao T: S-RNase-mediated self-incompatibility. *Journal of Experimental Botany* 2003, 54:115-122.

28.Zhang HB, Zhao X, Ding X, Paterson AH, Wing RA: Preparation of megabase-size DNA from plant nuclei. *The Plant Journal* 1995, 7:175-184.

29.Li H, Durbin R: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009, 25:1754-1760.

30.Novoalign [http://novocraft.com/].

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

32.Li R, Li Y, Kristiansen K, Wang J: SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008, 24:713-715.

33.McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 2010, 20:1297-1303.

34.Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, L X Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}*; iso-2; iso-3. Fly* 2012, 6:80-92.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

PeerJ PrePrints

35.Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, 26:841-842.

36.Kent WJ: BLAT - The BLAST-like alignment tool. *Genome Research* 2002, 12:656-664.

37.Seqclean [https://sourceforge.net/projects/seqclean/].

38.Hamilton JP, Neeno-Eckwall EC, Adhikari BN, Perna NT, Tisserat N, Leach JE, Lévesque A, Buell CR: The Comprehensive Phytopathogen Genomics Resource: a web-based resource for data-mining plant pathogen genomes. *Database* 2011.

39.UniVec [NCBI: http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html].

40.Chen, K., Wallis, JW, McLellan, MD, Larson, DE, Kalicki, JM, Pohl, CS, McGrath, SD, Wendl, MC, Zhang, Q, Locke, DP, Shi, X, Fulton, RS, Ley, TJ, Wilson, RK, Ding, L, Mardis, ER: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* 2009, 6:677-681.

41.Thompson JD, Gibson TJ, Higgins DG: Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics* 2002, 215:403-410.

42.Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ: Basic local alignment search tool. *Journal of Molecular Biology* 1990, 215:403-410.

43.Bombarely A, Edwards KD, Sanchez-Tamburrino J, Mueller LA: Deciphering the complex leaf transcriptome of the allotetraploid species *Nicotiana tabacum*: A phylogenomic perspective. *BMC Genomics* 2012, 13:406.

44.Drummond, AJ, Rambaut, A: BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 2007, 7:214.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

45. Bouckaert, RB: DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 2010, 26:1372-1373.

46.Smit AFA, Hubley R, Green P: Repeat Masker [http://www.repeatmasker.org].

47.Dewey CN: Aligning Multiple Whole Genomes with Mercator and MAVID. *Comparative Genomics* 2008, 395:221-235.

48.Swofford DL: From PAUP*: phylogenetic analysis using parsimony. Version 4. Massachusetts: Sinauer Associates; 2003.

49.Ronquist F, Huelsenbeck JP: MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003, 19:1572-1574.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Figure legends

Figure 1. Feature density of Yellow Pear, *S. galapagense,* and *S.pimpinellifolium* in comparison to H1706. (A) SNP density on chromosome 4 of sequenced accessions. (B) SNP density on chromosome 5 of sequenced accessions. (C) Read depth on chromosome 4 of sequenced accessions. (D) Read depth on chromosome 5 of sequenced accessions (E) Gene density on chromosome 4 based on *H1706* annotations (F) Gene density on chromosome 5 based on H1706 annotations.

Figure 2. Putative deletion size distribution in combined assemblies.

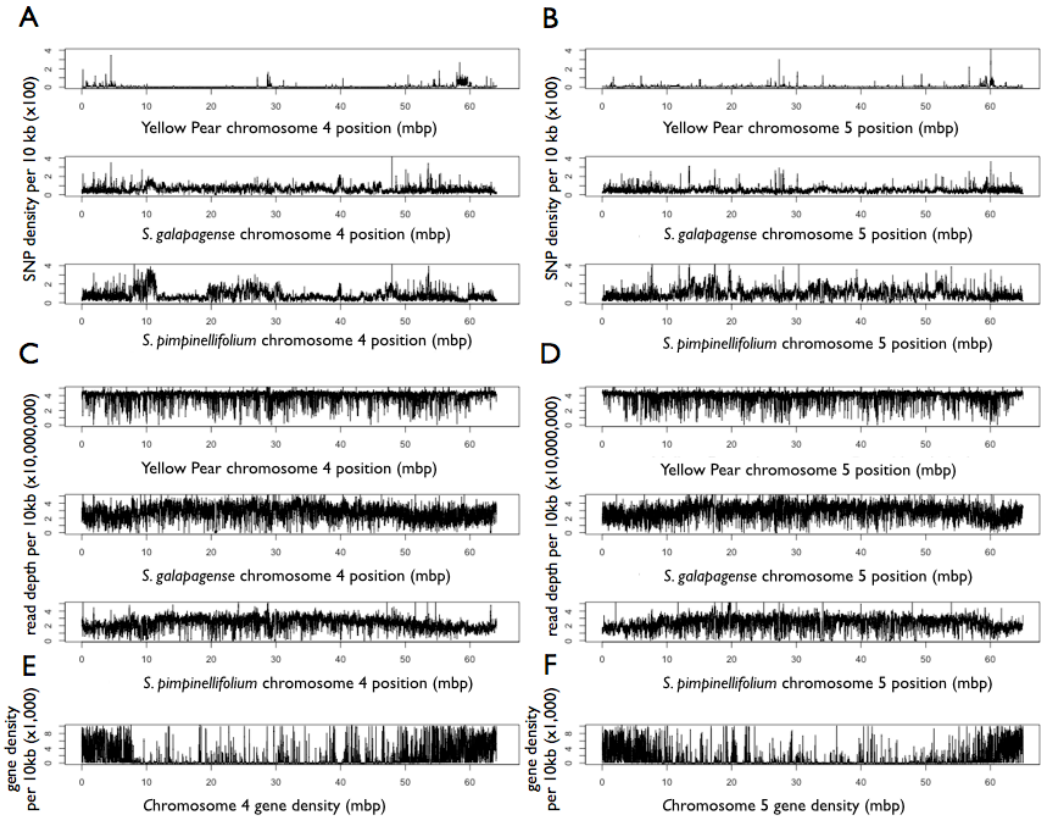
Figure 3. Gene trees inferred from coding sequence of 8,796 Solanum species genes. Phylogenetic trees were derived using maximum likelihood and were supported in at least 75 of 100 bootstrap replicates.

Figure 4. Tree topologies across selected chromosomes of *Solanum spp*. Coordinates are based on the H1706 reference genome. Posterior probabilities are shown for each tree. (A) Chromosome 4. (B) Chromosome 5. (C) Predominant tree topologies. 1=YP; 2=H1706; 3=*S. galapagense*; 4= *S. pimpinellifolium*; 5=*S. tuberosum.*

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

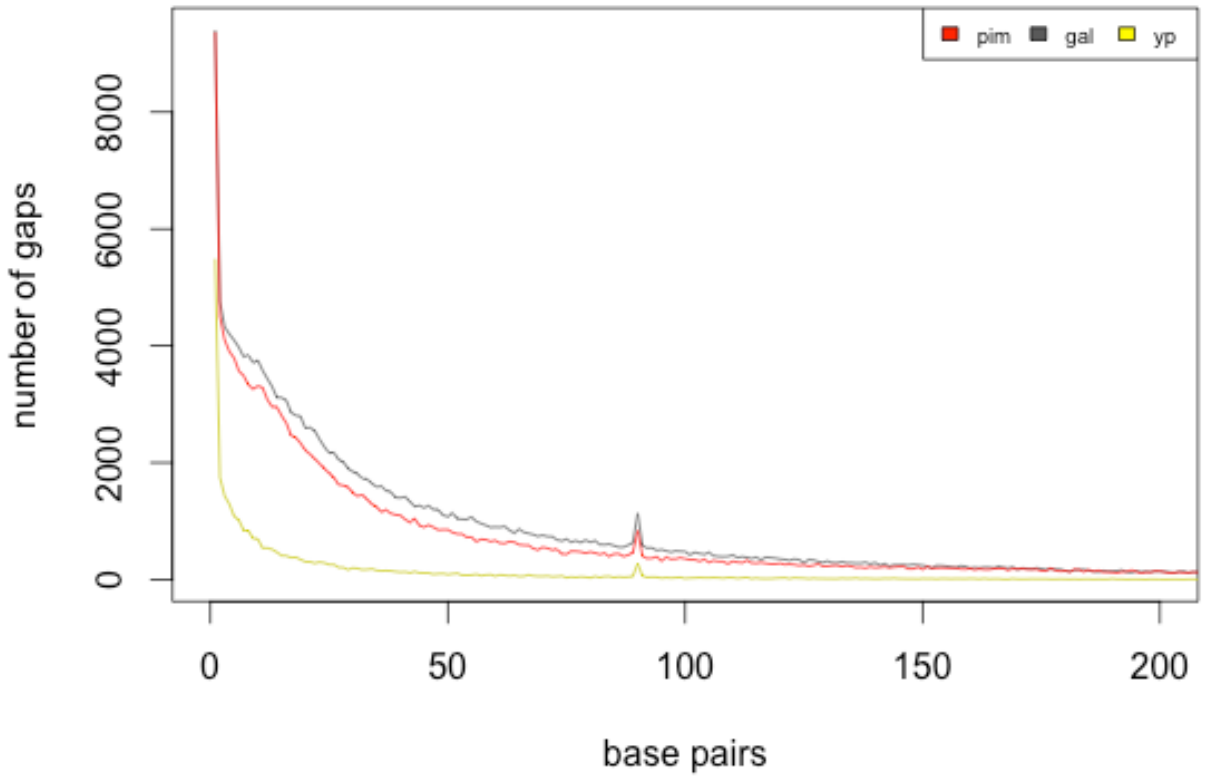
Figure 1.



PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

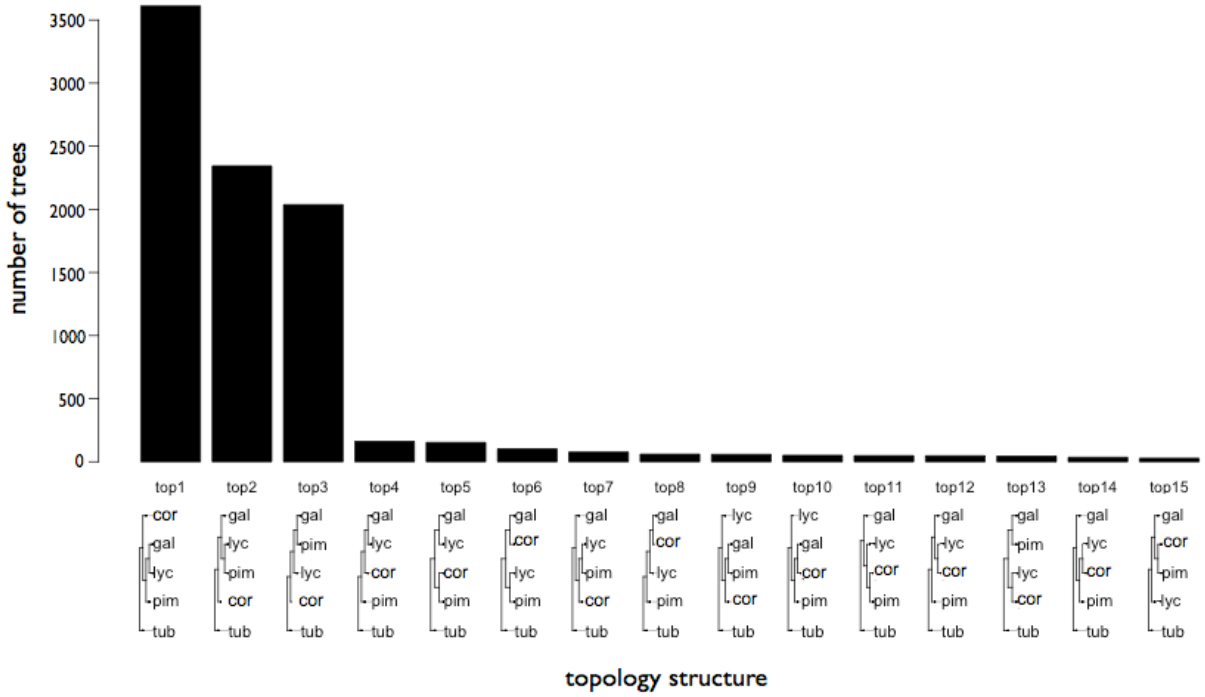
Figure 2.



PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

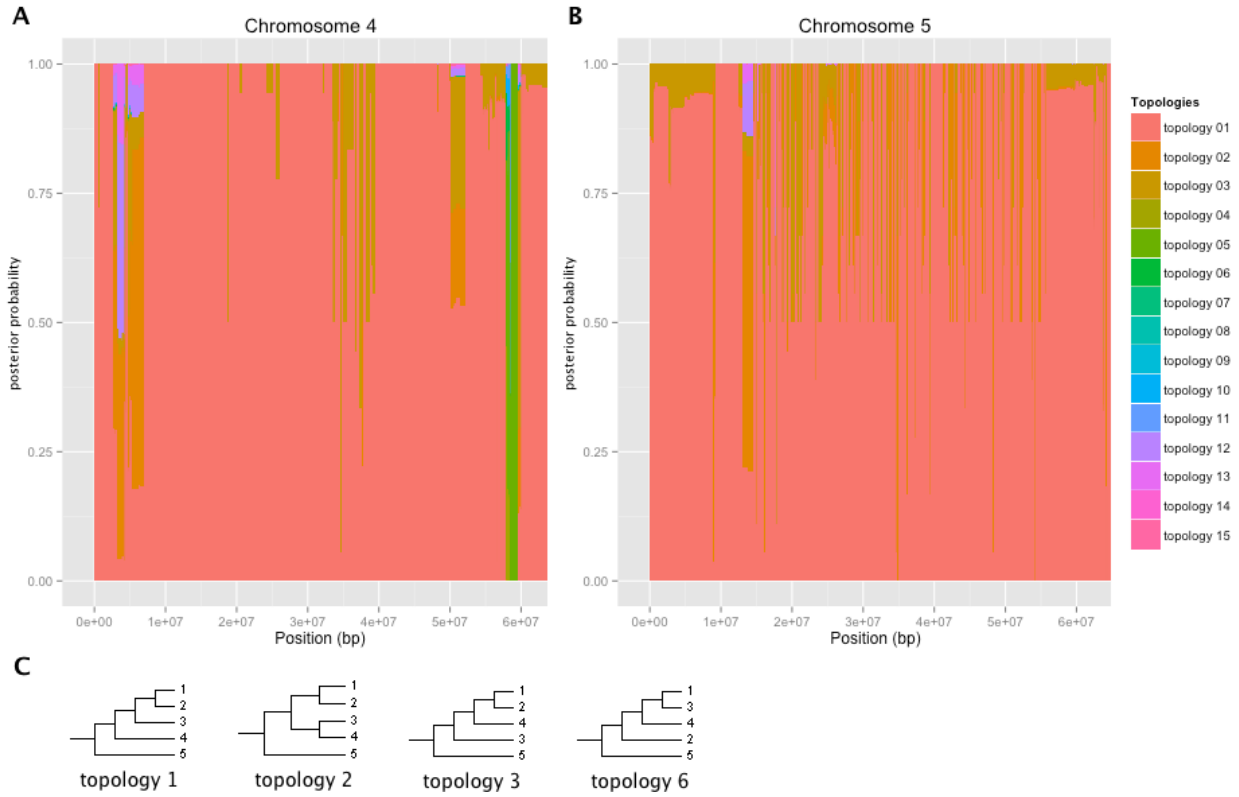
Figure 3.



PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Figure 4.



PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Tables

Table 1 Pairwise estimates of nonsynonymous (dN), synonymous (dS) mean substitution rate. Calculations are in comparison with *S. lycopersicum* 'Yellow Pear' and are based on 8,578 orthologous coding sequences for numbers not in parenthesis. Numbers in parenthesis are based on all usable coding sequences.

Species	dN	dS	*ω*
S. galapagense	0.0012 ± 0.0019 *(0.0029 ± 0.0058)*	0.0037 ± 0.0059 *(0.0052 ± 0.0117)*	0.3535 ± 2.0205 *(0.5191 ± 3.2039)*
S.pimpinellifolium	0.0013 ± 0.0022 *(0.0033 ± 0.0062)*	0.0043 ± 0.0065 *(0.0064 ± 0.0126)*	0.4305 ± 2.9802 *(0.5300 ± 3.3742)*
S. corneliomuelleri	0.0037 ± 0.0041	0.0151 ± 0.0123	0.3219 ± 1.0219
S. tuberosum	0.0332 ± 0.4361	0.1306 ± 1.3060	0.2386 ± 0.3127

[1] maximum likelihood estimate, values > 99 removed.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014

Table 2 Divergence time estimates of selected accessions. Calculations for Global Clock Method based on pairwise silent site substitutions for 3,611 genes. Calculations for coalescence method were performed with 8 genes. All genes used in calculations fit gene tree topology 1. Divergence date estimates are in reference to H1706. MYA=million years ago.

Species	dS	Divergence Date (MYA)	Divergence Date (MYA)
S. galapagense	0.0024 ± 0.0038	0.19	0.25
S. pimpinellifolium	0.0053 ± 0.0066	0.44	0.45
S. corneliomuelleri	0.0166 ± 0.0126	1.38	1.54
S. tuberosum	0.1335 ± 1.2383	11.07	NA

[1] based on global clock method.

[2] based on coalescence method.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.377v1 | CC-BY 4.0 Open Access | received: 30 Apr 2014, published: 30 Apr 2014