Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato

Daniel Koenig[a,b,1], José M. Jiménez-Gómez[a,c,1], Seisuke Kimura[a,d,2], Daniel Fulop[a,2], Daniel H. Chitwood[a], Lauren R. Headland[a], Ravi Kumar[a], Michael F. Covington[a], Upendra Kumar Devisetty[a], An V. Tat[a], Takayuki Tohge[e], Anthony Bolger[f], Korbinian Schneeberger[b,g], Stephan Ossowski[b,h], Christa Lanz[b], Guangyan Xiong[i], Mallorie Taylor-Teeples[a,j], Siobhan M. Brady[a,j], Markus Pauly[i], Detlef Weigel[b,3], Björn Usadel[f,k,l], Alisdair R. Fernie[e], Jie Peng[m], Neelima R. Sinha[a], and Julin N. Maloof[a,3]

[a]Department of Plant Biology and [m]Department of Statistics, University of California, Davis, CA 95616; [b]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; [c]Department of Plant Breeding and Genetics and [g]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany; [d]Department of Bioresource and Environmental Sciences, Kyoto Sangyo University, Kyoto 603-8555, Japan; [e]Department of Molecular Physiology and [f]Department of Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany; [h]Genes and Disease Program, Centre for Genomic Regulation, Barcelona 08003, Spain; [i]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720; [j]Genome Center, University of California, Davis, CA 95616; [k]Institute of Biology 1, Rheinisch-Westfaelische Technische Hochschule Aachen, 52056 Aachen, Germany; and [l]Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich, 52425 Jülich, Germany

Contributed by Detlef Weigel, May 30, 2013 (sent for review February 14, 2013)

Although applied over extremely short timescales, artificial selection has dramatically altered the form, physiology, and life history of cultivated plants. We have used RNAseq to define both gene sequence and expression divergence between cultivated tomato and five related wild species. Based on sequence differences, we detect footprints of positive selection in over 50 genes. We also document thousands of shifts in gene-expression level, many of which resulted from changes in selection pressure. These rapidly evolving genes are commonly associated with environmental response and stress tolerance. The importance of environmental inputs during evolution of gene expression is further highlighted by large-scale alteration of the light response coexpression network between wild and cultivated accessions. Human manipulation of the genome has heavily impacted the tomato transcriptome through directed admixture and by indirectly favoring nonsynonymous over synonymous substitutions. Taken together, our results shed light on the pervasive effects artificial and natural selection have had on the transcriptomes of tomato and its wild relatives.

domestication | biotic stress | abiotic stress

Domestication has long served as an important example of severe phenotypic divergence in response to selection. Darwin recognized the parallel between the processes of domestication and adaptation in the wild and used this analogy to emphasize the power of selection in generating phenotypic diversity (1). The genetic basis of domestication-associated phenotypes has been examined in several instances, most notably in maize, rice, tomato, and dogs (reviewed in refs. 2–5). The clear conclusion from these studies is that the rapid phenotypic divergence associated with domestication is often attributable to very few genetic loci (6). Improvements to DNA sequence technologies have allowed studies of the effect of domestication at the whole-genome level. Early population genetic analyses in maize found that very few genes (~5%) show evidence of positive selection during domestication of maize (7), and recent work using whole-genome resequencing has found a similar proportion of the genome was under positive selection (8). Evidence for strong selective sweeps at a limited number of loci has also been found in rice and dog genomes (9). Together with the previous genetic mapping work, these studies support the model that relatively few mutations experienced extremely strong selection by humans during domestication.

Although not the target of direct positive selection, the rest of the genome still experiences a dramatic shift in evolutionary pressures during domestication. Most characterized domestication events are associated with an extreme genetic bottleneck and alleviation of selective constraints in the original niche (10). These factors are predicted to increase the relative rate of nonsynonymous to synonymous (dN/dS) substitution, potentially resulting in the fixation of deleterious alleles (11). Previous studies comparing the distribution of polymorphisms between rice and dogs and their closest wild relatives have suggested that this may be the case (12, 13). However, the lack of genome-wide polymorphism data in multiple wild accessions has limited these comparisons because of ambiguous assignment of ancestral state. Evidence for changes at the transcriptional level during domestication have also been examined; for example, a recent study in maize has suggested widespread alteration of transcriptional networks during domestication (14). Although some of these changes are associated with genes that also show evidence of positive selection, changes in the topology of the gene-expression network might also result from fixation of mutations during the domestication bottleneck. Regardless, although absolute changes in gene expression or changes in network topology are thought to be important

Significance

One of the most important technological advances by humans is the domestication of plant species for the production of food. We have used high-throughput sequencing to identify changes in DNA sequence and gene expression that differentiate cultivated tomato and its wild relatives. We also identify hundreds of candidate genes that have evolved new protein sequences or have changed expression levels in response to natural selection in wild tomato relatives. Taken together, our analyses provide a snapshot of genome evolution under artificial and natural conditions.

Author contributions: D.K., J.M.J.-G., S.K., N.R.S., and J.N.M. designed research; D.K., J.M.J.-G., S.K., L.R.H., R.K., U.K.D., T.T., G.X., M.T.-T., S.M.B., M.P., B.U., and J.N.M. performed research; A.B., K.S., S.O., C.L., D.W., B.U., and A.R.F. contributed new reagents/analytic tools; D.K., J.M.J.-G., D.F., D.H.C., M.F.C., A.V.T., J.P., and J.N.M. analyzed data; and D.K., J.M.J.-G., D.F., N.R.S., and J.N.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE45774 and SRP019504).

[1]D.K. and J.M.J.-G. contributed equally to this work.

[2]S.K. and D.F. contributed equally to this work.

[3]To whom correspondence may be addressed. E-mail: weigel@tue.mpg.de or jnmaloof@ucdavis.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309606110/-/DCSupplemental.

during domestication, genome-wide comparison of expression between domesticated and multiple wild species is lacking.

One of the most heavily studied domestication events is that of tomato. Tomato is a member of a complex of 13 interfertile species that occupy a wide range of habitats in South America (15). The exact date of tomato domestication is debated, but it is clear that domesticated lines existed in Mexico at the time of the arrival of Europeans, and were brought back to Europe as a novelty, only to be used for food there in the 17th or 18th century. Tomato cultivars were subsequently reintroduced to the Americas. Thus, cultivated tomato has undergone a series of sequential bottlenecks, resulting in extremely low intraspecific genetic diversity (15). The most obvious domestication associated trait in tomato is a dramatic increase in fruit size. This trait has been the subject of extensive genetic analysis, and is controlled by a relatively small number of loci (16) making it typical of most domestication-associated traits. The high phenotypic diversity among wild tomato relatives and the relatively recent domestication of tomato itself makes it an excellent system to compare the effects of artificial and natural selection.

We deeply sequenced the transcriptomes of six species to ascertain the effects of natural and artificial selection on gene expression and sequence diversity. Our panel included one accession of domesticated tomato (*Solanum lycopersicum* M82), two related red-fruited wild species (*Solanum pimpinellifolium* and *Solanum galapagense*) and three green-fruited wild accessions from vastly differing habitats (*Solanum habrochaites*, a high altitude-adapted, chilling-tolerant accession; a high altitude drought-tolerant accession, *Solanum chmielewskii*; and *Solanum pennellii*, a desert-adapted accession) (Fig. 1*A*). These five wild species were chosen because of their dramatic phenotypic variability, but also because of their widespread use as genetic donors during cultivated tomato improvement, allowing us to define sequence and expression-level polymorphisms relevant to breeding and natural variation (17). Our analysis provides ample evidence for evolution in response to environmental cues in tomato relatives, and suggests interesting differences between artificial and natural selection.

Results

Characterization of Sequence Diversity in Wild and Cultivated Tomato. We conducted a series of experiments to define transcripts and identify sequence polymorphisms in our tomato panel. Two experiments (*Materials and Methods* and *SI Appendix*, Table S1) were conducted to ascertain interspecific variation in gene-expression levels. The first experiment compared gene expression in aerial seedling tissues of the species *S. pennellii*, *S. lycopersicum*, *S. habrochaites*, and *S. pimpinellifolium*. A second experiment compared six tissues collected from *S. lycopersicum* and *S. pennellii*. The remaining samples from either additional tissues or species were collected at separate times and used only for polymorphism discovery.

After alignment to the tomato reference genomic sequence (var. Heinz), our sequences covered an average of 67.4% of the annotated exonic gene space and allowed us to identify 1.5 million polymorphic sites among the 23.9 Mb covered in all samples (*SI Appendix*, Figs. S1–S6 and Tables S2 and S3, and Dataset S1). De novo contigs assembled from our reads covered 54% of the annotated genes and identified 34 transcripts not present in the current release of the Heinz reference genome (*SI Appendix*, Figs. S7 and S8, and Tables S4 and S5). Fewer than 20% (6 of 34) of these putative unique transcripts show homology with functionally annotated genes. Comparison of global patterns of nucleotide diversity across all accessions revealed a reduction in neutral divergence (dS) near the centromeres, but a relative increase in nonsynonymous substitution in the same regions (Fig. 1 *B* and *C*).

To initiate our evolutionary analysis, we used Bayesian inference methods to construct a phylogeny of the six species rooted with potato sequences (18) (*Solanum phureja*) (Fig. 1*A*).

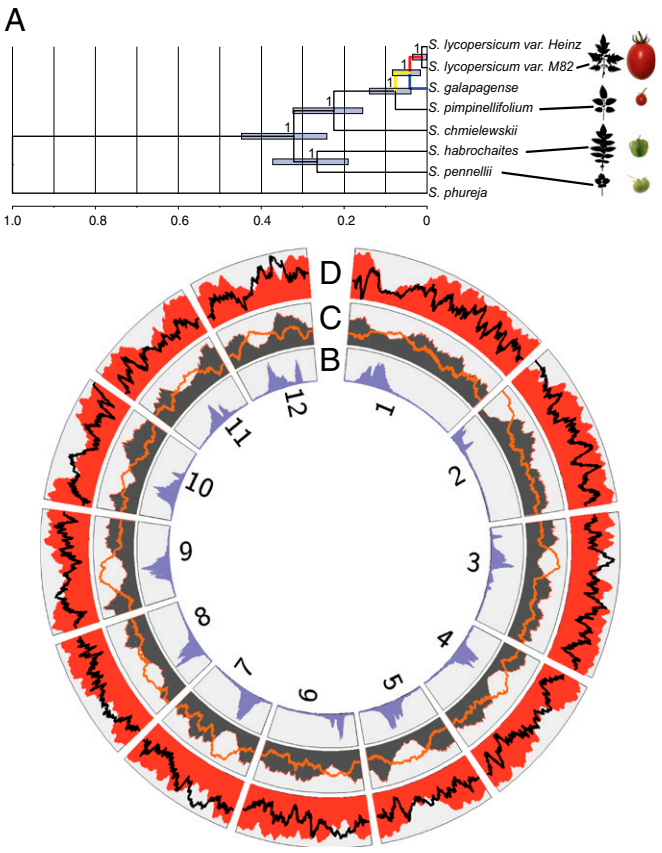

Fig. 1. Diversity in cultivated and wild tomatoes. (*A*) Bayesian relaxed-clock consensus chronogram, and examples of fruit and leaf divergence among tomato and wild relatives; nodes on the tree correspond to median branch lengths and blue bars represent 95% Bayesian confidence interval. (*B*) Distribution of mean distance to adjacent gene, larger distances are associated with centromeric sequences. (*C*) Single rate dS (gray) and single rate dN/dS (orange). (*D*) Frequency of expressed genes (red) and genes differentially expressed between tomato relatives (black). All plots reflect sliding windows (mean of 100 gene windows).

The resulting phylogeny is consistent with published tomato trees (19) resolving a monophyletic red/orange fruited clade and placing the green fruited *S. pennellii* and *S. habrochaites* in a sister clade. Like some previous studies but unlike others (19), the phylogeny places *S. galapagense* as the closest outgroup to domesticated samples. *S. pimpinellifolium*, *S. lycopersicum*, and *S. galapagense* have been shown to hybridize in the wild or through directed introgression (20) (in the case of *S. lycopersicum*) and it is possible that the difference in topologies results from incomplete lineage sorting in the three species and is specific to the particular accessions used in each study.

Consistent with previous studies, cultivated accessions are very similar to each other (< 1 SNP/kb), and a modest number of mutations separate cultivated tomato from its most closely related wild ancestors (< 5 SNP/kb) (*SI Appendix*, Fig. S3). By polarizing our data against the potato genome reference, we found that the spectrum of mutated sites varies between the lines. Mutations shared only by the cultivated tomato lines or unique to *S. galapagense* showed an increased ratio of nonsynonymous to synonymous substitutions (Fig. 2*A*). We directly tested whether the rate of nonsynonymous to synonymous substitution was elevated in cultivated tomato and *S. galapagense* by comparing the estimated tree-wide dN/dS to estimates for the terminal branches for each species and the branch leading to their most recent common ancestor (Fig. 2*B*). Each of the ter-

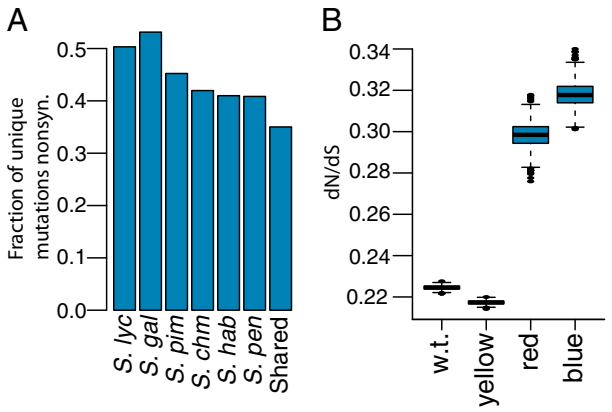

Fig. 2. Evidence for increased nonsynonymous substitution rate in *S. lycopersicum* and *S. galapagense*. (*A*) Fraction of species-specific derived mutations in the coding regions that are nonsynonymous. (*B*) Distributions of dN/dS estimates from 1,000 bootstraps of the transcriptome-wide alignment for the whole tree (w.t.) and the branches labeled with red, blue, and yellow in Fig. 1. *S. lyc*, *S. lycopersicum*; *S. gal*, *S. galapagense*; *S. pim*, *S. pimpinellifolium*; *S. chm*, *S. chmielewskii*; *S. hab*, *S. habrochaites*; *S. pen*, *S. pennellii*.

minal branches, but not the connecting branch, showed significant increases in dN/dS. Both *S. lycopersicum* and *S. galapagense* are thought to have experienced strong genetic bottlenecks (15, 21) (during domestication, and island colonization and recent adaptation, respectively). Our result is consistent with separate bottlenecks in these two species and increased accumulation of potentially deleterious mutations during cultivation and colonization. This change in mutation spectrum may be a result of relaxed purifying selection, fixation of mutations during the genetic bottleneck because of drift, or both.

Evidence for Positive Selection in Wild and Cultivated Tomato. Although relaxed purifying selection is expected to elevate dN/dS by random substitution throughout the genome, positive selection is expected to increase dN/dS within specific loci. From comparison of gene-level estimates of dN/dS in all species (22, 23), we identified 51 genes that show statistically significant ($P < 0.05$) evidence of evolution under positive selection across the phylogeny (Dataset S2). Many of these genes have not been characterized in tomato, but annotated genes included the tomato homolog of the *Arabidopsis thaliana ARGONAUTE 2* and the known tomato-resistance gene *immunity to fusarium wilt-2C4* (24, 25), consistent with rapid evolution of protein sequences in response to pathogen pressure. Homologs of the aluminum transporter *ALUMINUM SENSITIVE 1* and the calcium uptake transporter *MID1-COMPLEMENTING ACTIVITY 1* also showed significantly elevated dN/dS pointing to positive selection in response to abiotic factors, such as soil chemistry (26, 27). This second set of genes is particularly interesting considering the high salt tolerance observed in wild tomato relatives (28).

Divergence in Gene Expression in Wild and Cultivated Tomato. We next searched for evidence for differential expression between aerial seedling tissues of *S. lycopersicum*, *S. habrochaites*, *S. pimpinellifolium*, and *S. pennellii*. For this process, seedling tissues were chosen to minimize the effects of developmental and environmental variation on gene expression. We detected expression of 25,012 transcripts in at least a single accession, and 20,389 in all surveyed accessions. Consistent with previous observations (29), gene expression was low in centromere proximal regions and higher in gene-dense chromosomal arms (Fig. 1*D* and *SI Appendix*, Fig. S9). We fit a generalized linear model (*SI Appendix*, *SI Materials and Methods* and Fig. S10) to our expression data to identify 7,903 genes showing evidence of differential expression among species (*SI Appendix*, Fig. S10). Gene ontology (GO) enrichment analysis of these revealed overrepresentation of genes involved in stress response, defense response, photosynthesis, response to high light, and redox pathways (*SI Appendix*, Table S6). Enrichment for these categories indicates that abiotic and biotic stresses have played a major role driving transcriptional variation among these species.

Interspecific comparisons based on nucleotide alignments and pairwise gene-expression differences revealed a general concordance in tree topology but a striking increase in the *S. pennellii* gene-expression branch length (Fig. 3 *A* and *B*). The number of expression changes specific to the *S. pennellii* lineage was much higher than any other lineage (Fig. 3*C* and *SI Appendix*, Tables S7 and S8), indicating that the transcriptional landscape of *S. pennellii* is highly diverged relative to the other three species. A small but significant increase in unique expression changes was also found in *S. lycopersicum* compared with *S. pimpinellifolium*, suggesting the possibility of accelerated divergence in expression in the domesticated lineage (141 and 91 genes, respectively, χ^2 P value = 0.0007). GO term enrichment analysis identified genes involved in salt stress in all comparisons with *S. pennellii* and modification to sucrose metabolism and starch metabolic process in all comparisons with *S. lycopersicum*; in addition, redox pathways were enriched in many comparisons (*SI Appendix*, Table S9

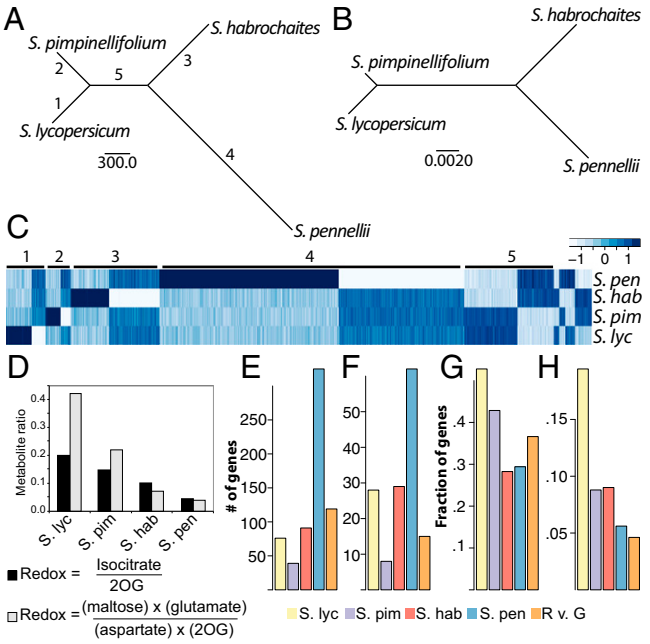

Fig. 3. Interspecific variation in expression. (*A*) Neighbor-joining tree built from the number of pairwise differentially expressed genes compared with (*B*) the unrooted genetic tree from Fig. 1*A*. The scale bar in *A* is for the number of differentially expressed genes and the scale bar in *B* is the expected number of substitutions per site. (*C*) Heatmap depicting scaled expression values of genes separated into two groups by significant contrasts (*SI Appendix, SI Materials and Methods*). The numbers correspond to the branch of the tree on which the changes are assumed to have occurred. (*D*) Product/substrate redox ratio of NAD(P)-linked reactions, calculated as described by refs. 74 and 75. Black and gray indicate redox value of isocitrate dehydrogenase and malate dehydrogenase reactions, respectively. 2OG: 2-oxoglutarate. (*E* and *F*) Number of differentially expressed genes showing evidence of accelerated expression divergence (two-rate Brownian motion fit better than one-rate Brownian motion and OU) at $\Delta AIC > 4$ or $\Delta AIC > 10$, respectively. (*G* and *H*) Proportion of differentially expressed genes unique to each lineage (as determined by pairwise contrasts) showing evidence of accelerated expression divergence at $\Delta AIC > 4$ or $\Delta AIC > 10$, respectively. RvG indicates genes that show contrasting rates of evolution in the red and green fruited lineages.

and S10). We validated that several of these expression changes are reflected in the metabolic state of the plants. Fructose levels were six- to ninefold lower in all wild species compared with *S. lycopersicum* (30). Analysis of existing GC-MS data (30, 31) revealed that the product-to-substrate ratio of redox-coupled NAD(P) reactions in *S. habrochaites* and *S. pimpinellifolium* were twofold and in *S. pennellii* more than 10-fold lower than in *S. lycopersicum*, indicating that the NAD(P) pool is in a more oxidized state in the three wild species (Fig. 3*D*). These metabolic changes combined with enrichment in transcriptional changes provide strong evidence that redox pathways are rapidly evolving among these species. Furthermore, the substantial shift in *S. pennellii* is consistent with adaptation to high light conditions. In summary, the pathways identified by these analyses are consistent with the expected selective pressures on each of these lineages, with strong natural selection for life in a desert environment for *S. pennellii* and artificial selection for palatable fruits during breeding of domesticated tomato.

Analysis of Selective Pressures on Gene Expression. Gene-expression variation can result from random genetic drift or changes in selective pressure. To identify genes that have potentially undergone a shift in selection regime, we compared the fit of three evolutionary models to the gene-expression levels in our dataset: a model of evolution under random drift (Brownian motion single rate), stabilizing selection (Ornstein-Uhlenbeck, OU), or a change in evolutionary rate along a particular lineage (Brownian motion two rate) (32–34) (*SI Appendix*, Table S11 and Dataset S3). Genes whose expression values showed a substantially better fit to the two-rate model and that had accelerated evolutionary rates in a particular lineage were considered candidates for alteration in selective regime in that lineage. Fit was assessed for each of the three models and then compared between the accelerated rate model and the other models using the change in Akaike information criteria (ΔAIC). Increased ΔAIC indicates stronger fit for the accelerated rate model compared with both of the other models (see *SI Appendix* for additional information). Among differentially expressed genes ($P < 0.01$), there was evidence for differing rates of evolution across the tree in 1,764 genes (22.3% of differentially expressed genes, ΔAIC > 4) and strong evidence in 428 genes (5.4% of differentially expressed genes; ΔAIC > 10) (*SI Appendix*, Table S11). The largest group of genes was evolving at a faster rate along the *S. pennellii* branch, but increasing the ΔAIC threshold increased the relative number of genes found in the other branches (Fig. 3 *E* and *F*). Furthermore, the proportion of differentially expressed genes with evidence of accelerated evolution of expression levels was higher in *S. lycopersicum* than in *S. pennellii* (or any of the other branches) (Fig. 3 *G* and *H*). These results indicate that much of the rapid divergence in gene expression that has occurred in *S. pennellii* can be explained by neutral processes. In contrast, relatively few genes have changed in *S. lycopersicum*, but these genes are more likely to show evidence for a *S. lycopersicum*-specific change in evolutionary rate.

Genes accelerated in the green- and red-fruited lineages included *yellow-flesh*, a major locus controlling fruit color (35, 36). We also found many genes accelerated along the *S. pennellii* branch that are involved in responses to environmental stresses, such as salt, drought, heat, and oxidative damage, as well as genes in the abscisic acid pathway (Dataset S3). This finding is consistent with the results from differential expression and codon substitution models, and combined indicate that alteration in the pathways regulating stress responses has been important in the evolutionary history of this organism.

Evolution of the Tissue-Specific Expression in *S. pennellii* and *S. lycopersicum*. Natural variation has frequently been shown to involve tissue-specific gene expression alterations. We therefore examined whether gene-expression patterns might have been altered during domestication or in response to natural selection by contrasting *S. lycopersicum* var. M82 and the desert adapted *S. pennellii* (37–40). Gene-expression values between *S. lycopersicum* and *S. pennellii* were compared across a panel of six tissue types, including root, vegetative, and floral tissues.

We used principle components analysis (PCA) to identify major sources of variance in our transcriptome dataset (Fig. 4 *A* and *B*). Variation in expression across tissues explained the two largest principle components, but species-driven differences were also evident. Despite this substantial interspecific variation in gene expression, the pattern of gene expression across tissue types was positively correlated between species for the vast majority of genes (Fig. 4*B* and *SI Appendix*, Fig. S11). To examine the tissue specificity of expression differences between species, we applied PCA to between-species log fold-change values calculated for each tissue and found the majority (56%) of the variance was explained by global shifts in gene expression (Fig. 4 *C* and *D* and *SI Appendix*, Fig. S12). By fitting a statistical model accounting for species and tissue effects we identified 3,474 transcripts [false-discovery rate (FDR) < 0.01; 1,718 with log fold-change > 1] differentially expressed between species and 7,844 across tissues (FDR < 0.01) (Dataset S4). Only 166 transcripts were identified where the pattern of expression across tissues was significantly different between species, consistent with the general conservation in tissue-specific expression. We con-

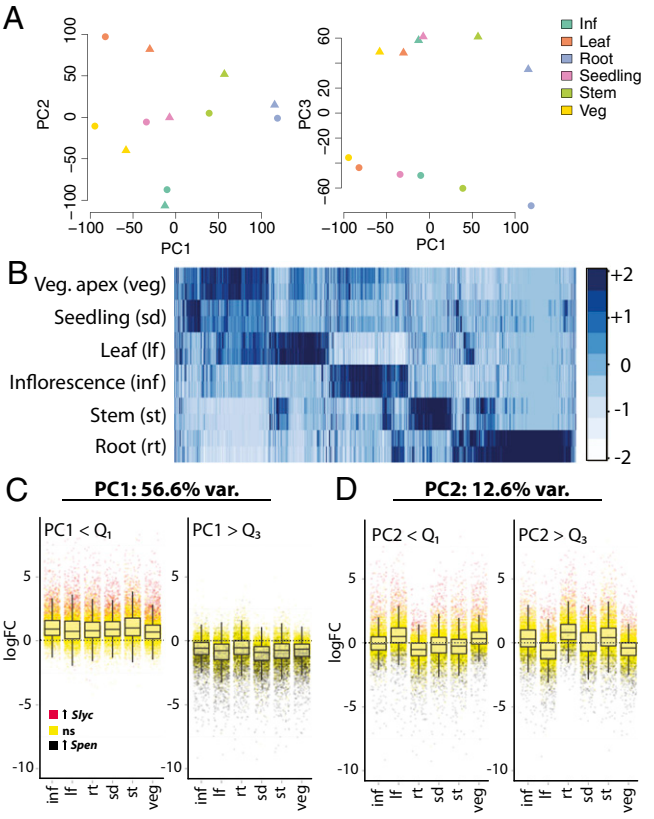

Fig. 4. Differential expression in *S. lycopersicum* and *S. pennellii*. (A) PCA factorial maps showing the largest components of variance, which separate samples by tissue (PC1 and PC2) and species (PC3). Triangles represent *S. pennellii* samples and circles *S. lycopersicum*. (*B*) Heat map comparing scaled expression values for *S. lycopersicum* (*Top*) and *S. pennellii* (*Bottom*) across tissues. Darker blue indicates higher expression. (*C*) PCA on log fold-changes by tissue. PC1 explains variance relating to global shifts in gene expression. (*D*) The remaining PCs describe tissue-specific shifts. Q3 and Q1 indicate the upper and lower quantile of the data, respectively (inf, inflorescence meristem; lf, leaf; rt, root; sd, aerial seedling; st, stem; veg, vegetative meristem).

firmed our relative expression estimates using quantitative RT-PCR and found strong correlation with our RNAseq data (ρ = 0.91) (*SI Appendix*, Fig. S13) validating our methodology.

Evolution of the Gene Coexpression Networks of *S. pennellii* and *S. lycopersicum*. To gain additional insight into the pattern of gene-expression changes between *S. pennellii* and *S. lycopersicum*, we built weighted gene coexpression networks for each species using genes significantly differentially expressed across tissues from our previous analysis (Fig. 5 and *SI Appendix*, Fig. S14). This approach allows us to compare the pattern of gene-expression correlations in both species, rather than the absolute level of gene expression, and has been shown to provide additional evolutionary insight (41). For both species, three major modules of highly coexpressed genes were identified (Fig. 5 *A* and *B*, *SI Appendix*, Fig. S14 and Table S12, and Dataset S5) (a fourth small module was also identified in *S. lycopersicum* but was not considered for the remainder of the analysis). The largest module (green; 852 genes found in both species) contained genes highly induced or repressed in photosynthetic tissues (leaf, vegetative shoot, and aerial seedling tissues) and was enriched for GO terms related to photosynthesis, carbon metabolism, and response to light (*SI Appendix*, Tables S13 and S14). A second module (purple; 272 genes found in both species) separated root tissues from all other tissues (*SI Appendix*, Tables S15 and S16, and Datasets S6–S9). The final large module (yellow; 144 genes in both species) differentiated vegetative and inflorescence shoot tissues from others and was enriched for GO terms related to cell division (*SI Appendix*, Tables S17 and S18). The overlap between these modules indicates extensive conservation of coexpression networks between the two species.

Although modules often overlapped between the two species, we noticed that characteristics of the two networks were not equivalent. In particular, the connectivity (as measured by the sum of the absolute values of the correlation coefficients of a focal gene with all other genes, see *SI Appendix, SI Materials and Methods*) of the *S. pennellii* network was on average higher than that of *S. lycopersicum* (Fig. 5*C* and *SI Appendix*, Fig. S14). This signal was primarily because of genes highly connected in both species but more highly connected in *S. pennellii*. Calculating connectivity for genes within each module gave similar values in both species for the purple and yellow modules, but connectivity was strongly reduced in the green module in *S. lycopersicum* (Wilcoxon test *P* value < 2e-16) (Fig. 5*D*). To further explore this finding, we identified species-specific connections (edges) between genes in each module and between genes not assigned to a module (Fig. 5*E*). If the networks had changed similarly since the two species diverged, one would expect equivalent numbers of gain/loss of edges in each network. In agreement, about the same number of species-specific edges were found between genes either not assigned to a module or in the yellow and purple modules. In contrast, a much higher number of *S. pennellii*-specific edges were identified in the green module. Taken together, these data demonstrate that photosynthetic tissue-specific gene expression is more tightly correlated in *S. pennellii* than in *S. lycopersicum*.

Effect of Introgression on the Transcriptome of Domesticated Tomato. An important strategy in tomato improvement is the extensive use of wild germplasm during breeding. Previous work suggested the possibility of large introgressions in the tomato reference sequence var. Heinz (29). Such introgressions combine previously independently evolving alleles that may result in novel changes in expression. We searched for evidence of introgressions and found that SNPs differentiating the Heinz and M82 cultivars were nonrandomly distributed (Fig. 6*E*). These regions of high diversity showed increased allele sharing with *S. pimpinellifolium*, indicating recent introgression from this or a closely related species. Using this pattern of diversity (*Materials and Methods*), we defined 550 candidate introgressed genes in Heinz and 2,479 in M82. The large number of candidate loci introgressed in M82

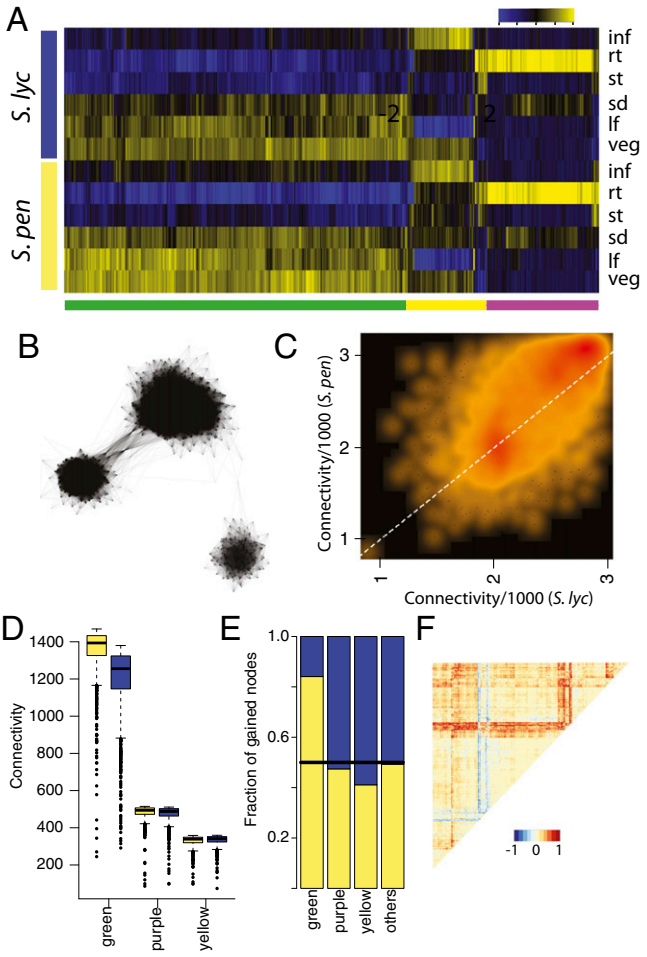

Fig. 5. Evolution of coexpression networks in *S. lycopersicum* and *S. pennellii*. (*A*) Heatmap depicting expression of genes assigned to the three modules in both species. Scaled $\log_2$ expression values are shown with yellow and blue indicating high and low expression respectively. Green, yellow and purple bars indicate membership in the three identified transcription modules. (*B*) Global depiction of conserved coexpression network components. Three clear clusters that correspond to the three major modules are evident. (*C*) Comparison of connectivity (sum of the absolute correlation of expression with all other genes) for genes in the two species. Black indicates a low density of points and red indicates a high density. Connectivity is positively correlated, but the highest values are increased in *S. pennellii*. White dashed line indicates a slope of 1. (*D*) Intramodule connectivity for each module in each species. Yellow boxes are *S. pennellii* values and blue *S. lycopersicum*. (*E*) Fraction of differential edges specific to *S. pennellii* (yellow) and *S. lycopersicum* (blue) for each module. (*F*) Heatmap showing the change in connectivity for all gene pairs in the green module. Red indicates correlations that are found only in *S. pennellii*, blue indicates correlations found only in *S. lycopersicum*, and yellow indicates correlations of approximately the same strength in both networks.

highlights the challenge of linkage drag during breeding using wild accessions, and may contribute to reduced genome-wide divergence in nucleotide sequence and divergence in gene expression between cultivated accessions and wild accessions.

Transgressive or nonparental expression phenotypes are a well-described characteristic of expression in hybrid lines (42), and thus introgression in tomato might result in new expression phenotypes. We examined whether introgressions of this size might contribute to expression divergence by comparing gene expression in *S. lycopersicum*, *S. pennellii*, and an introgression line (IL) where a portion of chromosome 4 from *S. pennellii* was introgressed into *S. lycopersicum* (Fig. 6 *A*–*D*)

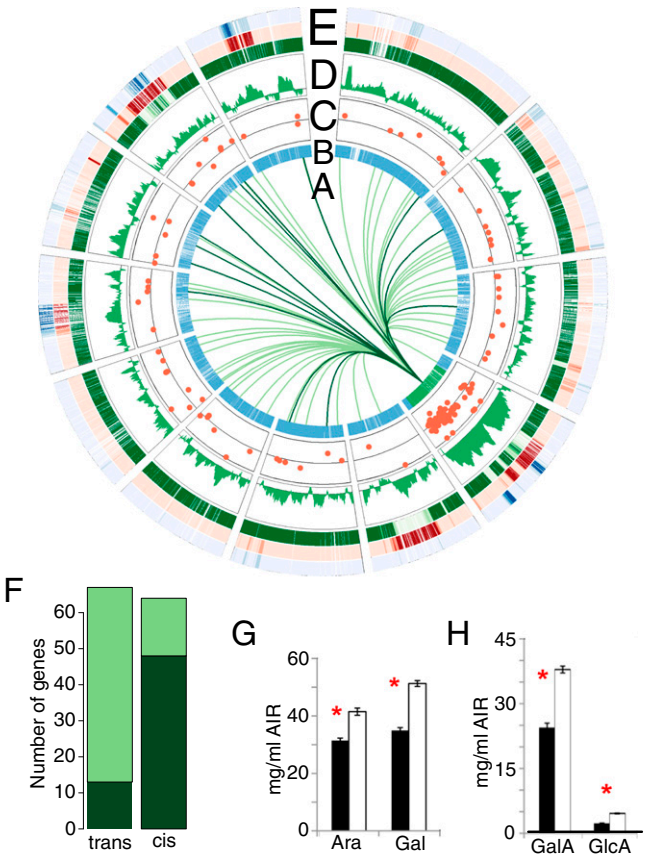

Fig. 6. *cis* and *trans* expression divergence. (*A*) Transgressive (light green) and *S. pennellii*-like (dark green) *trans* regulation relationships with genes on IL4-3. (*B*) Genotype by gene in IL4-3 (M82, blue and *S. pennellii*, green). (*C*) Log fold-change for differentially expressed genes between IL4-3 and M82. The black line indicates 0. (*D*) Correlation of log fold-change in expression between M82 and either the IL or PEN in sliding windows. A strong increase is seen on chromosome 4 indicating higher correspondence in gene expression between *S. pennellii* and the introgression line. (*E*) Heatmap showing median dissimilarity (polymorphisms per 1,000 bp) between M82 and Heinz (green), M82 and *S. pimpinellifolium* (red), and Heinz and *S. pimpinellifolium* (blue). Increasing dissimilarity is indicated by lighter color; increasing similarity is indicated by darker color. Introgressions into one of the two cultivated lines show increased polymorphism rate between Heinz and M82 and decreased polymorphism between *S. pimpinellifolium* and the acceptor cultivated line (see chromosome 5). Shared introgressed segments show decreased polymorphism between *S. pimpinellifolium* and both cultivated lines. Many of these introgressions are larger in M82 (see chromosome 11). Large introgressions were frequently associated with centromeres, likely resulting from the increased linkage drag in these regions during breeding. Sliding window size for *B*, *D*, and *E* is 100 genes. (*F*) Number of expression changes in *cis* or *trans* to the IL4-3 introgression. Light green indicated transgressive changes and dark green indicates *S. pennellii* like expression. (*G* and *H*) Abundance of pectic monosaccharides galacturonic acid (GalA), glucoronic acid (GlcA), arabinose (Ara), and galactose (Gal) present in the walls (AIR, alcohol insoluble residue) of roots from M82 (black) and *S. pennellii* (white). Red asterisk: $P < 0.001$.

(43). Of the 131 genes differentially expressed between the IL and M82, 61 (47%) exhibited nonparental expression levels and 70 showed expression similar to *S. pennellii*. Genes exhibiting *S. pennellii*-like expression were enriched in the introgressed fragment but the majority of genes showing nonparental expression patterns were found outside of the fragment (Fig. 6*F*). The enrichment for nonparental expression *in trans* to the introgression provides evidence of the existence of epistatically interacting mutations within each lineage that, when combined, result in unique expression phenotypes. An additional possible contributing factor is the recent discovery of transgressive siRNA expression patterns in tomato hybrids (44). These results point to introgression as a possible source of unique expression phenotypes in cultivated tomato.

Expression Divergence Correlates with Phenotypic Differences Among Wild and Cultivated Accessions. Our combined comparison of sequence and transcriptional diversity in cultivated and wild tomato relatives identified distinct footprints of selection under artificial and natural conditions. Adaptation to an extreme desert climate manifests as dramatic phenotypic shifts seen in *S. pennellii* compared with cultivated tomato. These phenotypes include up to 20-fold higher levels of epicuticular lipid deposition in *S. pennellii* leaves (45), amphistomic leaves with reduced stomatal pore size (~13% smaller, $P = 5.17 \times 10^{-6}$), and alterations in cell wall composition in the roots (*SI Appendix*, Fig. S15 and S16). Each of these phenotypes can be correlated with gene-expression profiles in our data.

A thick cuticle with an increased accumulation of epicuticular waxes is known to limit water loss and increase water-deficit tolerance (46–49). In the tomato relatives, epicuticular waxes account for up to 20% dry weight of *S. pennellii* leaves, whereas they make up only 0.9% of *S. lycopersicum* leaf dry weight (45). This striking increase in accumulation of epicuticular waxes is accompanied with marked differences in the expression of genes associated with wax deposition between *S. lycopersicum* and *S. pennellii* in our datasets (*SI Appendix*, Table S19). For example, the genes that encode the tomato orthologs of two enzymes involved in the production of aliphatic wax component precursors, *ECERIFERUM6* (CER6) and CER10 (50–52) are significantly higher in *S. pennellii* in comparison with *S. lycopersicum*. *FIDDLEHEAD*, which encodes a condensing enzyme involved in synthesis of cuticular lipids (53) and the genes encoding orthologs of CER1, CER2, and CER8, which are involved in conversion of very long-chain fatty acids to alkanes in *Arabidopsis* (54–57) are also higher in *S. pennellii.* Furthermore, *S. pennellii* has higher expression of a *CER5*-like gene, which might be involved in wax secretion (58) and the genes encoding the drought responsive nonspecific lipid transfer proteins (LTP) LTP1 and LTP2 (59, 60). Together, these results demonstrate a concerted up-regulation of candidate genes for wax accumulation in desert adapted *S. pennellii.*

S. pennellii leaves have several developmental features consistent with drought adaptation including reduced surface:volume ratio and changes in stomatal density (61). One developmental regulator that might be involved is *SCREAM1*, a positive regulator of stomatal index (the ratio of stomata to epidermal cells) (62, 63). *S. lycopersicum* shows almost twofold lower levels of *SCREAM1* ($P = 0.00018$). Consistent with these changes, *S. pennellii* has an increased adaxial stomatal index relative to *S. lycopersicum*, yielding a roughly even stomatal index on both leaf surfaces (*SI Appendix*, Fig. S16) (19, 64). Typical for a desert plant, *S. pennellii* has thick succulent leaves (1.46× the thickness of *S. lycopersicum*), (*SI Appendix*, Fig. S16), thus the relative increase in adaxial stomata may be required for efficient CO_2 diffusion in these thicker leaves (64, 65).

Previous studies have reported that root growth under drought conditions can be promoted by cell-wall modulation of glucoronoxylan and rhamnogalacturonan side chains in cell-wall components (66, 67). Correlating with these observations, genes expressed nearly exclusively in the root include many genes involved in cell-wall metabolism, such as multiple pectinesterases and polygalacturonases, several β-galactosidases, and a reversibly glycosylated protein involved in UDP-arabinofuranose production (68), the precursor for arabinan biosynthesis. To validate the relevance of the expression differences we examined root primary cell-wall composition in *S. lycopersicum* and *S. pennellii* and found that *S. pennellii* had higher levels of the abundant galactan and arabinan side-chains of rhamnogalacturonan I (Fig. 6 *G* and *H*) (69), consistent with its desert habitat.

Discussion

Tomato is one of our most important vegetable crops and its improvement is largely dependent on introgression of beneficial alleles from wild germplasm (15). Here we have identified hundreds of thousands of polymorphic positions that distinguish cultivated tomato from its wild relatives. All of these species have individual attributes that could be potentially valuable for tomato crop improvement, and our study provides the raw material necessary for marker-assisted introgression of such traits.

We have shown that domestication was associated with the fixation of many potentially deleterious protein and expression-level changes. The consequences of such changes are unknown, but it is possible that some have decreased vigor in domesticated lines. Adaptation to extreme environments among tomato relatives appears to have caused a broad alteration of transcriptional networks in parallel with positive selection at the sequence level for a number of genes related to environmental adaptation. This is particularly the case for the desert-adapted *S. pennellii*. Our finding that gene-expression changes in *S. pennellii* were highly accelerated relative to nucleotide divergence suggests that the previously noted importance of regulatory changes in morphological evolution (70, 71) is likely a genome-scale phenomenon. The signal of adaptation to extreme environments in the *S. pennellii* transcriptome is on par with that seen for biological processes classically thought to evolve at an accelerated rate, such as defense response and reproductive biology. Previous work in maize has suggested extensive transcriptional rewiring in response to domestication (14). The most extensive network rewiring that we discovered in *S. lycopersicum* relates to light responsiveness. Loss of connectivity in this network may reflect selection for reduced light response in *S. lycopersicum*, or may reflect a more robust response in the desert-adapted *S. pennellii*; this hypothesis is amenable to future genetic experimentation. In contrast to adaptation to pressures emanating from the natural environment, as deduced from differences between wild tomato species, we have found artificial selection and domestication to be associated with a relatively small number of changes at both at the sequence and transcriptional level. Taken together, our studies highlight both parallels and contrasts between natural and artificial selection and their effects on genome evolution.

Materials and Methods

Plant Materials. *S. lycopersicum* var. M82 (LA3475), *S. pennellii* (LA0716), and the *S. pennellii* introgression line IL4-3 (LA4051) were donated by Dani Zamir, The Hebrew University of Jerusalem, Jerusalem, Israel (43). *S. habrochaites* (LA1777) and *S. pimpinellifolium* (LA1589) were obtained from the C. M. Rick Tomato Genetics Resource Center, University of California at Davis. *S. chmielewskii* (LA1840) was donated by Keygene, Wageningen, The Netherlands. *S. galapagense* (LA0530) was donated by Maria Asins at the Instituto Valenciano de Investigaciones Agrarias, Valencia, Spain. The obligate outcrossing *S. habrochaites* line was maintained by growth of 10 or more plants and cross-pollinated by hand. All other accessions were maintained by selfing.

RNA Isolation. In the transcriptome experiment, total RNA from all tissues except fruits were extracted using TRIzol (Invitrogen) according to the manufacturer's standard protocol. The following modifications were included in the protocol to extract RNA from fruits: Total RNA from the aqueous phase in the chloroform extraction step was precipitated with 0.25 volume isopropanol and 0.25 volume of 1.2 M sodium chloride/0.8 M sodium citrate buffer, washed with 70% (vol/vol) ethanol, and resuspended in water. Another precipitation step with 0.8 volume lithium chloride and 3 volumes 100% (vol/vol) ethanol, was done if the 260/230 absorbance ratio of the total RNA was less than 1.5.

The RNeasy plant mini kit (Qiagen) was used to extract total RNA for the seedling experiment.

Sequencing and Read Filtering. A total of 57 libraries from *S. lycopersicum* var. M82, *S. pimpinellifolium*, *S. pennellii*, *S. habrochaites* were sequenced in 14 lanes from seven different 84 cycle runs of the Illumina GA II, returning 406,874,298 paired-end and 169,290,821 single-end reads. Additionally, single libraries from *S. galapagense* and *S. chmielewskii* were sequenced in a Hiseq2000 to obtain 67,504,782 and 53,873,978 100-bp paired end reads, respectively. After separating reads by barcode, removing Illumina adapter sequences, and trimming low-quality bases, we used in our analysis 547,612,718 reads with a minimum length of 50 bp (average of 81 bp).

RNAseq Read Alignment. Three different strategies were used for RNAseq read analysis. For polymorphism detection and total coverage calculations we aligned the reads against the *S. lycopersicum* genomic reference. For quantification of gene expression we created a matched set of contigs that were used as a reference. De novo assembly was performed on reads obtained from *S. lycopersicum* var. M82 and *S. pennellii* to identify unannotated transcripts and transcripts not synthesized by *S. lycopersicum* var. Heinz.

SNP Calling. A custom Bioperl script was used to detect SNPs and indels between each sequenced species and the reference sequence (72). Homozygous SNPs/indels were called in positions with a minimum coverage of four reads and an allele frequency higher than 0.66 for SNPs and 0.33 for indels. Heterozygous SNPs were called in positions with at least four reads per allele and a frequency of at least 25% in both alleles. To avoid calling polymorphisms from the ends of the reads that span exon-intron junctions, we divided the reads into five equal regions and discarded SNPs and indels covered only by a single region of the reads. All polymorphisms from all species were merged in a matrix and their positions genotyped in all species where the polymorphism was not present. These genotypes were called using the same allele frequency thresholds as above but no coverage threshold.

Statistical Analysis. All statistical analysis was done using the R statistical programming environment (73).

ACKNOWLEDGMENTS. We thank Stacey Harmer and Danelle Seymour for helpful comments on the analysis and manuscript; the Tomato Genetics Resource Center, Keygene, Dr. Maria Asins, and Dr. Dani Zamir for providing seed; the Tomato Sequencing Consortium for prepublication access to the tomato genome sequence; Niels Müller and Kerstin H Richau for technical assistance with *Solanum galapagense* and *Solanum chmielewskii* sequencing; Heike Keller (Max Planck Institute for Developmental Biology), Fernando Carrari, and Gabriel Lichtenstein (Instituto Nacional de Tecnología Agropecuaria, Argentina) for helping with the *Solanum pennellii* genome build; Brian Moore for access to the AutoParts software prior to publication; Liam Revell, Carl Boettiger, and Luke Harmon for discussions on the phylogenetic modeling of gene expression; Sergei Kosakovsky Pond for assistance with HyPhy analyses; and the iPlant Collaborative (http://www.iplantcollaborative.org) for providing computational resources and data storage. RNA-seq reads and SNPs from this study can be viewed in a genome browser at http://phytonetworks.ucdavis.edu/tomato and have been submitted to the sol genomics network (http://solgenomics.net) for distribution. Promoter sequences, de novo contigs with no match to the tomato 2.40 references chromosomes, and the matched *S. lyc*/*S. pen.* CDS set used in this study can be downloaded from http://phytonetworks.ucdavis.edu/Download. This work was funded through a National Science Foundation Grant (IOS-0820854) to N.R.S., J.N.M., and J.P., and a Human Frontier Science Program Fellowship (LT000783) to D.K.

1. Darwin C (1868) *The Variation of Animals and Plants Under Domestication* (J. Murray, London).
2. Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127(7):1309–1321.
3. Parker HG, Shearin AL, Ostrander EA (2010) Man's best friend becomes biology's best in show: genome analyses in the domestic dog. *Annu Rev Genet* 44:309–336.
4. Izawa T, Konishi S, Shomura A, Yano M (2009) DNA changes tell us about rice domestication. *Curr Opin Plant Biol* 12(2):185–192.
5. Paran I, van der Knaap E (2007) Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. *J Exp Bot* 58(14):3841–3852.
6. Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. *Trends Plant Sci* 15(9):529–537.
7. Wright SI, et al. (2005) The effects of artificial selection on the maize genome. *Science* 308(5726):1310–1314.
8. Hufford MB, et al. (2012) Comparative population genomics of maize domestication and improvement. *Nat Genet* 44(7):808–811.
9. Huang X, et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490(7421):497–501.
10. Gepts P (2004) Crop domestication as a long-term selection experiment. *Plant Breed Rev* 24(2):1–44.

11. Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics* (Roberts and Co, Greenwood Village, CO), pp xxvii, 734 pp.
12. Lu J, et al. (2006) The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends Genet* 22(3):126–131.
13. Cruz F, Vilà C, Webster MT (2008) The legacy of domestication: Accumulation of deleterious mutations in the dog genome. *Mol Biol Evol* 25(11):2331–2336.
14. Swanson-Wagner R, et al. (2012) Reshaping of the maize transcriptome by domestication. *Proc Natl Acad Sci USA* 109(29):11878–11883.
15. Atherton JG, Rudich J (1986) *The Tomato Crop: A Scientific Basis for Improvement* (Chapman and Hall, London, New York), pp xv, 661 pp.
16. Grandillo S, Tanksley SD (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. *Theor Appl Genet* 92(8):935–951.
17. Mattoo A, Razdan MK, eds (2007) *Genetic Improvement of Solanaceous Crops Volume 2. Tomato.* (Science, Enfield, NH), pp xx, 637.
18. Xu X, et al.; Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189–195.
19. Spooner DM, Peralta IE, Knapp S (2005) Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes *Solanum* L. section Lycopersicon (Mill.) Wettst. *Taxon* 54(1):43–61.
20. Darwin SC, Knapp S, Peralta IE (2003) Taxonomy of tomatoes in the Galápagos Islands: Native and introduced species of *Solanum* section Lycopersicon (Solanaceae). *Syst Biodivers* 1(1):29–53.
21. Nuez F, Prohens J, Blanca JM (2004) Relationships, origin, and diversity of Galapagos tomatoes: Implications for the conservation of natural populations. *Am J Bot* 91(1): 86–99.
22. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725–736.
23. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11(5):715–724.
24. Zhang X, et al. (2011) *Arabidopsis* Argonaute 2 regulates innate immunity via miRNA393(*)-mediated silencing of a Golgi-localized SNARE gene, MEMB12. *Mol Cell* 42(3):356–366.
25. Ori N, et al. (1997) The I2C family from the wilt disease resistance locus I2 belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell* 9(4):521–532.
26. Larsen PB, Cancel J, Rounds M, Ochoa V (2007) *Arabidopsis* ALS1 encodes a root tip and stele localized half type ABC transporter required for root growth in an aluminum toxic environment. *Planta* 225(6):1447–1458.
27. Yamanaka T, et al. (2010) MCA1 and MCA2 that mediate Ca2+ uptake have distinct and overlapping roles in *Arabidopsis*. *Plant Physiol* 152(3):1284–1296.
28. Frary A, et al. (2010) Salt tolerance in *Solanum pennellii*: Antioxidant response and related QTL. *BMC Plant Biol* 10:58.
29. Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641.
30. Schauer N, Zamir D, Fernie AR (2005) Metabolic profiling of leaves and fruit of wild species tomato: A survey of the *Solanum lycopersicum* complex. *J Exp Bot* 56(410): 297–307.
31. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc* 1(1):387–396.
32. Bedford T, Hartl DL (2009) Optimization of gene expression by natural selection. *Proc Natl Acad Sci USA* 106(4):1133–1138.
33. Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* 50(4):1404–1417.
34. O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60(5):922–933.
35. Fray RG, 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(4):589–602.
36. Ronen G, Cohen M, Zamir D, Hirschberg J (1999) 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. *Plant J* 17(4):341–351.
37. Frary A, et al. (2000) fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289(5476):85–88.
38. Chen KY, Cong B, Wing R, Vrebalov J, Tanksley SD (2007) Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes. *Science* 318(5850): 643–645.
39. Fridman E, Carrari F, Liu YS, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* 305(5691):1786–1789.
40. Martin B, Thorstenson YR (1988) Stable carbon isotope composition (deltaC), water use efficiency, and biomass productivity of *Lycopersicon esculentum*, *Lycopersicon pennellii*, and the F(1) hybrid. *Plant Physiol* 88(1):213–217.
41. Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci USA* 103 (47):17973–17978.
42. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* 102(5):1572–1577.
43. Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141(3):1147–1162.
44. Shivaprasad PV, Dunn RM, Santos BA, Bassett A, Baulcombe DC (2012) Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. *EMBO J* 31(2):257–266.
45. Fobes JF, Mudd JB, Marsden MP (1985) Epicuticular lipid accumulation on the leaves of *Lycopersicon pennellii* (Corr.) D'Arcy and *Lycopersicon esculentum* Mill. *Plant Physiol* 77(3):567–570.
46. Kosma DK, et al. (2009) The impact of water deficiency on leaf cuticle lipids of *Arabidopsis*. *Plant Physiol* 151(4):1918–1929.
47. Zhang JY, et al. (2005) Overexpression of WXP1, a putative *Medicago truncatula* AP2 domain-containing transcription factor gene, increases cuticular wax accumulation and enhances drought tolerance in transgenic alfalfa (*Medicago sativa*). *Plant J* 42(5): 689–707.
48. Zhang JY, Broeckling CD, Sumner LW, Wang ZY (2007) Heterologous expression of two *Medicago truncatula* putative ERF transcription factor genes, WXP1 and WXP2, in *Arabidopsis* led to increased leaf wax accumulation and improved drought tolerance, but differential response in freezing tolerance. *Plant Mol Biol* 64(3):265–278.
49. Aharoni A, et al. (2004) The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in *Arabidopsis*. *Plant Cell* 16(9):2463–2480.
50. Millar AA, et al. (1999) CUT1, an *Arabidopsis* gene required for cuticular wax biosynthesis and pollen fertility, encodes a very-long-chain fatty acid condensing enzyme. *Plant Cell* 11(5):825–838.
51. Fiebig A, et al. (2000) Alterations in CER6, a gene identical to CUT1, differentially affect long-chain lipid content on the surface of pollen and stems. *Plant Cell* 12(10): 2001–2008.
52. Zheng H, Rowland O, Kunst L (2005) Disruptions of the *Arabidopsis* Enoyl-CoA reductase gene reveal an essential role for very-long-chain fatty acid synthesis in cell expansion during plant morphogenesis. *Plant Cell* 17(5):1467–1481.
53. Yephremov A, et al. (1999) Characterization of the FIDDLEHEAD gene of *Arabidopsis* reveals a link between adhesion response and cell differentiation in the epidermis. *Plant Cell* 11(11):2187–2201.
54. Aarts MG, Keijzer CJ, Stiekema WJ, Pereira A (1995) Molecular characterization of the CER1 gene of *Arabidopsis* involved in epicuticular wax biosynthesis and pollen fertility. *Plant Cell* 7(12):2115–2127.
55. Negruk V, Yang P, Subramanian M, McNevin JP, Lemieux B (1996) Molecular cloning and characterization of the CER2 gene of *Arabidopsis* thaliana. *Plant J* 9(2):137–145.
56. Xia Y, Nikolau BJ, Schnable PS (1996) Cloning and characterization of CER2, an *Arabidopsis* gene that affects cuticular wax accumulation. *Plant Cell* 8(8):1291–1304.
57. Lü S, et al. (2009) *Arabidopsis* CER8 encodes LONG-CHAIN ACYL-COA SYNTHETASE 1 (LACS1) that has overlapping functions with LACS2 in plant wax and cutin synthesis. *Plant J* 59(4):553–564.
58. Pighin JA, et al. (2004) Plant cuticular lipid export requires an ABC transporter. *Science* 306(5696):702–704.
59. Thoma S, et al. (1994) Tissue-specific expression of a gene encoding a cell wall-localized lipid transfer protein from *Arabidopsis*. *Plant Physiol* 105(1):35–45.
60. Trevino MB, OConnell MA (1998) Three drought-responsive members of the nonspecific lipid-transfer protein gene family in *Lycopersicon pennellii* show different developmental patterns of expression. *Plant Physiol* 116(4):1461–1468.
61. Kebede H, Martin B, Nienhuis J, King G (1994) Leaf anatomy of two *Lycopersicon* species with contrasting gas exchange properties. *Crop Sci* 34(1):108–113.
62. Kanaoka MM, et al. (2008) SCREAM/ICE1 and SCREAM2 specify three cell-state transitional steps leading to *Arabidopsis* stomatal differentiation. *Plant Cell* 20(7): 1775–1785.
63. Abrash EB, Bergmann DC (2010) Regional specification of stomatal production by the putative ligand CHALLAH. *Development* 137(3):447–455.
64. Nakazato T, Warren DL, Moyle LC (2010) Ecological and geographic modes of species divergence in wild tomatoes. *Am J Bot* 97(4):680–693.
65. Mott KA, Gibson AC, O'Leary JW (1982) The adaptive significance of amphistomatic leaves. *Plant Cell Environ* 5(6):455–460.
66. Leucci MR, Lenucci MS, Piro G, Dalessandro G (2008) Water stress and cell wall polysaccharides in the apical root zone of wheat cultivars varying in drought tolerance. *J Plant Physiol* 165(11):1168–1180.
67. Keppler BD, Showalter AM (2010) IRX14 and IRX14-LIKE, two glycosyl transferases involved in glucuronoxylan biosynthesis and drought tolerance in *Arabidopsis*. *Mol Plant* 3(5):834–841.
68. Konishi T, et al. (2007) A plant mutase that interconverts UDP-arabinofuranose and UDP-arabinopyranose. *Glycobiology* 17(3):345–354.
69. Mohnen D (2008) Pectin structure and biosynthesis. *Curr Opin Plant Biol* 11(3): 266–277.
70. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134(1):25–36.
71. Zhao Q, et al. (2008) The role of regulatory genes during maize domestication: Evidence from nucleotide polymorphism and gene expression. *Genetics* 178(4): 2133–2143.
72. Stajich JE, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10):1611–1618.
73. R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
74. Heineke D, et al. (1991) Redox transfer across the inner chloroplast envelope membrane. *Plant Physiol* 95(4):1131–1137.
75. Sies H (1982) Nicotinamide nucleotide compartmentation. *Metabolic Compartmentation*, ed Sies H (Academic, London), pp 235–257.

Supplementary Information Appendix

Comparative transcriptomics in wild and domesticated tomato

Daniel Koenig, José M. Jiménez-Gómez, Seisuke Kimura, Daniel Fulop, Daniel H. Chitwood, Lauren R. Headland, Ravi Kumar, Michael F. Covington, Upendra Kumar Devisetty, An V. Tat, Takayuki Tohge, Anthony Bolger, Korbinian Schneeberger, Stephan Ossowski, Christa Lanz, Guangyan Xiong, Mallorie Taylor-Teeples, Siobhan M. Brady, Markus Pauly, Detlef Weigel, Björn Usadel, Alisdair R. Fernie, Jie Peng, Neelima R. Sinha, and Julin N. Maloof[1]

Materials & Methods

Tissue collection

Libraries from mRNAs of seedlings of *S. lycopersicum*, *S. pimpinellifolium, S. habrochaites,* and *S. pennellii* were used in the "Seedling experiment" and libraries from six different tissues in *S. lycopersicum* and *S. pennellii,* were used for the "Transcriptome analysis" or "Transcriptome experiment" (see Table S1).

For the seedling experiment, tomato seeds (*Solanum lycopersicum*, *S. pennellii*, *S. pimpinellifolium* and *S. habrochaites*) were germinated on MS plates kept in dark for 3 days with 5 seeds per plate. Then, plates were exposed to light and grown upright on plates at 22 ºC in a Conviron controlled environment chamber under a mixture of cool-white and far-red fluorescent lights in a complete randomized design. Light intensity averaged 95 μE with a red to far-red ratio of 0.48 or 3.1. The shoot tissue was collected from these seedlings 10 days after sowing on plate.

In the transcriptome experiment seeds from *S. lycopersicum* and *S. pennellii* were germinated and grown as indicated above. After 10 days, seedlings were transferred to soil and kept in the same conditions until anthesis. All samples were grown in simulated shade and sun conditions to expand the number of transcripts covered, but for the purpose of expression modeling these treatments were treated as replicates. After flowering the plants were transferred to the greenhouse. The following tissues were used in the "Transcriptome experiment": Roots and aerial tissues were collected from seedlings 10 days after germination. Vegetative meristems were collected from plants when the 3rd leaf reached 1 mm (around 30 to 37 days after germination). The stem between the 4th and 5th leaves and inflorescence meristems were collected when these meristems were fully formed (50 days after germination for M82 and 56 for *S. pennellii*). Young green fruits and mature fruits were collected from plants in the greenhouse. For the introgression line IL4-3, growth conditions and tissue collection were the same as in the vegetative meristem sample described above.

Library preparation

mRNA-seq libraries for the transcriptome experiment were prepared using the Illumina mRNA-seq sample prep kit (Illumina, RS-100-0801) according to the

manufacturer's protocol. Custom paired-end adapters (PE adapter) were used to multiplex libraries. Eight PE adapters with a unique 3-bp barcode sequence (AAA, AGG, CAC, CGT, GCT, GTC, TCA and TTG) at the end of the adapter were used for the library preparation. Barcodes were chosen so any one sequencing error in the barcode cannot transform one barcode into another, dramatically reducing the chance of contamination between libraries due to sequencing errors. The primer sequences used for making the barcoded PE adapters are follows. PE1-AAA, P-TTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG: PE2-AAA, ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAAT; PE1-AGG, P-CCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG, PE2-AGG, ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGT; PE1-CAC, P-GTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-CAC, ACACTCTTTCCCTACACGACGCTCTTCCGATCTCACT, PE1-CGT, P-ACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-CGT, ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTT; PE1-GCT, P-AGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-GCT, ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTT; PE1-GTC, P-GACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-GTC, ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCT; PE1-TCA, P-TGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-TCA, ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAT; PE1-TTG, P-CAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG; PE2-TTG, ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGT.

PE1 and PE2 primers were mixed in annealing buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 50 mM NaCl) and annealed by heating to 95 ºC and gradually cooling down to 4 ºC. The cDNA libraries were quantified using Bioanalyzer (Agilent), then pooled in random subsets of 8 samples and sequenced (paired-end, 85 bp each) in the Illumina genome analyzer (GAII).

Mapping to the genomic reference

Single nucleotide polymorphisms and indels up to 15 nt long were obtained from the alignments of all available short sequences from each species separately. *S. lycopersicum* var. M82 reads were aligned to the *S. lycopersicum* var. Heinz v2.4 chromosome sequence using BWA version 0.5.7 with parameters -k 1 -l 25 -n 0.02 -e 15 -i 10(1). The fraction of edits per read (parameter –n) was raised to 0.05 for all wild species to account for the divergence between these species and cultivated tomato. This parameter was calculated from binomial distributions using the SNP rates estimated from(2). We developed a custom Perl script to extract reads mapping to multiple locations. To test if reads align to putative splice junctions we mapped them using TopHat version 1.2.0 and Bowtie version 0.12.7 (parameters -g 1 --segment-mismatches 1 -F 0 -a 8 -m 1 -i 14 -I 10000)(3, 4). The resulting alignment files from TopHat and BWA were merged, and SAMtools and Picard were used to filter uniquely mapped reads and remove duplicated reads(1) (http://picard.sourceforge.net). Insert sizes were estimated with Picard for each library and range from 81 to 229 bp, with an average of 178 bp and a median of 198 bp. We then used a more sensitive algorithm available in the GATK toolkit to realign the reads overlapping SNPs and indels(5, 6). A genome browser has been set up containing the results our alignments with respect to the *S. lycopersicum* var. Heinz genome (http://phytonetworks.ucdavis.edu/tomato/). Statistics for sequencing depth and transcript coverage can be found in Tables S1-3.

Generating and mapping to a matched set of reference cDNAs

To facilitate mapping and accurate expression analysis of RNAseq reads to different species across the tomato complex we took advantage of a draft *S. pennellii* genomic sequence (v0.6.1; *S. pennellii* consortium) to build a matched set of reference cDNAs for *S. lycopersicum* and *S. pennellii.* The goal was to obtain a matched set of references of equal length containing sequences known to exist in both species and retaining species-specific polymorphisms. The following steps were used for each coding sequence (CDS) defined in the ITAG2.3 set. 1) *S. lycopersicum* CDSs were used to BLAST against *S. pennellii* scaffolds using MegaBLAST(7), (settings -e 1e-50 -m 7 -N 2 -t 18 -W 11 -A 50) to identify the appropriate scaffold and region. 2) *S. pennellii* scaffold sequence

encompassing the BLAST hit region and an additional 2kb on either side was retrieved and GMAP(8) was used to thread the *S. lycopersicum* CDS onto the *S. pennellii* scaffold (settings: -n 1 -f 1). 3) GMAP output was parsed to create matching *S. lycopersicum* and *S. pennellii* CDSs. Only the matching regions were retained. 4) The matched sets were then filtered to only retain good hits. To accomplish this, the predicted *S. pennellii* CDSs were BLASTed against the full *S. lycopersicum* CDS set (a reciprocal blast) using MegaBLAST (settings as above). To retained matched pairs we required that the best reciprocal BLAST hit was to the original ITAG CDS, that the best BLAST hit had an E-value at least 10^3 more significant than the second best hit, and (because we were also interested in obtaining upstream promoter regions—see below) that the 5' HSP be at least 50 bp, have a 90% identity, and be within 300 bp of the query start. In this way, from the original 34,727 annotated ITAG CDS (median length 834) we created 28,801 matched CDS pairs (median length 849). While a number of gene models are lost using this technique, it is justified for differential expression analysis by the increased short-read mapping accuracy allowed by the matched set. These sequences are available at http://phytonetworks.ucdavis.edu/Download/ .

We used BWA and SAMtools to map RNAseq reads to the matched reference cDNA set. *S. lycopersicum* and *S. pimpinellifolium* were mapped to the *S. lycopersicum* set, whereas *S. habrochaites* and *S. pennellii* were mapped to the *S. pennellii* set. The parameters for BWA were "-n 0.1 -e 12 -k 1 -l 25"; SAMtools was used with -n 1 to select reads that mapped unambiguously to the reference. Read counts from this alignment were used in analysis of differential expression.

***de novo* Assembly**

For *de novo* assembly we first used derep_tree.pl from Rnnotator(9) to remove duplicated reads. This yielded 60,371,072/58,806,408 unique paired-end reads from the 117,300,560/116,317,314 original reads from *S. lycopersicum* and *S. pennellii* respectively. *de novo* contigs were then assembled using ABySS(10), with the call: "abyss-pe n=10" and "-k" set as (23,33,43,53 and 63). We then used CAP3(11) to further assemble overlapping regions from a pool of all contigs from each of the 5-kmer assemblies. The resulting cap3 contigs and singlets were combined together and those

that are smaller than 200 nucleotides were eliminated. We were able to assemble 79.3%/85.4% of the reads to obtain 37,778/38,039 contigs longer than 200 nucleotides for *S. lycopersicum* var. M82 and *S. pennellii* respectively. To determine quality and coverage of assembled contigs, we used BLAST (blastall -E 2 -G 2 -F "m D" -e 10^-10) to align them to the annotated transcriptome and genome sequence. Maximum identity scores of 95% and 85% were used as cutoff values for intra- and interspecies comparisons, respectively. Sequence comparisons to reference sequences revealed that 79.5%/82.1% of the contigs longer than 200bp matched the tomato *S. lycopersicum* var. Heinz transcriptome (ITAG2.3), while 99.4%/99.0% had hits to the *S. lycopersicum* var. Heinz genome (CHR2.40). As expected, the majority of *de novo*-assembled contigs (79.4%/82.0%) are represented in both the reference transcriptome and genome (Figure S8). The few contigs matching only ITAG2.3 are mostly (72%) plant transcripts whose splicing pattern likely disrupts genomic alignment. The other contigs in this category are non-nuclear-encoded plant genes and five apparent alignment artifacts that correspond to bacterial genes or have no significant similarity to any sequence in NCBI's non-redundant nucleotide database. Of particular interest in the analysis of the *de novo* assembly are the contigs that did not have significant alignments to either of the reference sequences. These 'novel contigs' may represent genes that were not identified when building the reference transcriptome and genome. We visually inspected the contigs and corresponding BLAST results from ITAG2.3, CHR2.40 and NCBI's non-redundant nucleotide and protein databases. As expected, many contigs corresponded to non-nuclear-encoded plant genes and plant-associated bacterial, viral, or fungal genes. We did, however, identify 44 novel contigs that represented 34 unique genes not found in the reference sequences (the sequences of these contigs are available for download at http://phytonetworks.ucdavis.edu).

Coding sequence alignments for phylogenetic analyses

We generated fasta files containing alignments with the predicted coding sequences from the tomato reference genome and all species in our experiments. This analysis included transcripts with acceptable coverage in at least 80% of their length in all six studied species. Each allele was created with a custom R/Bioconductor script that

substituted the SNPs and deletions identified in our study in the *S. lycopersicum* var. Heinz annotated coding sequences (ITAG v2.3). Insertions were added in the corresponding species while padding the sequences from the remaining species to maintain their open reading frame. Positions in the alleles with low or absent coverage were substituted with dashes.

By virtue of the way they were generated, these putative transcripts were already aligned to the reference tomato gene coding sequences. Alignments were further screened for frameshifts caused by indels by translating each alignment and ensuring that the amino acid version had a similar percent identity to the nucleotide one. The resulting alignments numbered 11,751. These alignments contain only high quality SNPs.

Potato allele mining

For each tomato coding sequence we determined the putative potato homolog as the reciprocal best hits in BLAST searches comparing the protein sequences from both transcriptomes (Tomato ITAG v2.3 and potato PGSC DM v3.4)(12). Homologous proteins were considered only if they aligned in the correct orientation, presented a single homolog in the other species with a bit score higher than 150, the second hit was not above a bit score of 1000, and the difference between the bit score of the first and second hit was above 150. We required these thresholds to occur in the BLAST searches performed in both directions. This yielded 10,709 putative potato-tomato homolog pairs. Custom Bioperl scripts were used to translate to protein the inter-specific tomato DNA alignments described above, realign them together with the corresponding potato homolog using MUSCLE(13) and convert them back to cDNA(14). We discarded alignments with less than 90% identity at the protein level. We used the remaining 9,405 alignments to infer the ancestral potato genotypes of 1,395,553 coding SNPs. These alignments are available at http://phytonetworks.ucdavis.edu/Download.

Polymorphism effect calculation

Polymorphism effects on the protein sequence were calculated with a custom Perl script that integrated SNP into codons based on the *S. lycopersicum* var. Heinz ITAG annotation v2.3 coding sequences. SNPs in the same codon were considered as a single

polymorphism. For SNPs located in the UTRs, we developed a custom script that extended the open reading frame from the start or stop codons to the five prime or three prime UTR respectively. We then calculated the effect of SNPs located in these hypothetical proteins as before.

Interspecific SNP effect comparisons and identification of introgressions

To compare of the ratio of non-synonymous to synonymous SNPs private to each lineage we polarized our SNP data against potato gene sequences and considered only biallelic sites with complete information in all lines. We then calculated the ratio non-synonymous to synonymous derived mutations specific to each species (excluding polymorphic sites within *S. lycopersicum*).

Introgression from *S. pimpinellifolium* into cultivated lines was identified by first calculating local SNP rates between *S. pimpinellifolium* and each cultivated accession individually. We considered genes covered over 80% in each compared sample, and calculated the mean SNP rate (SNP/covered bp) in 30 gene-sliding windows. To set a threshold for significant windows we then permuted the gene order (across chromosomes) 10,000 times and re-performed the sliding window analysis selecting the lowest 0.005 percentile from the maximum values found across all permutations as our final threshold. Genes that fell into windows significant were then called as introgressed. This approach is conservative with respect to introgression size, only the largest introgressions will be identified, and it is likely that many small introgressions exist.

General pipeline for the inference of substitution rates and tests

Custom HyPhy Batch Language (HBL) wrappers were written for several methods implemented in HyPhy v2 (15). The wrappers were typically organized in a trio, with a top-wrapper, a mid-wrapper, and at the bottom of the hierarchy one or several method files modified to return pertinent parameter estimates. The top-wrapper handles Message Passing Interface (MPI) messaging to iterate through all the alignments in the dataset in parallel on a specified number of processors; it calls the mid-wrapper in each processor and retrieves mid-wrapper results when a given locus' analysis is done in a given processor. The mid-wrapper calls one (or more) modified method files in series on a

single processor, and retrieves the results from these method files. The method files called by the mid-wrapper are modified to return parameters values to the mid-wrapper. The top-wrapper collates the results and at the end of iterating through all the alignments saves a summary table containing the results for all the loci out to a file.

Typically, we ran a nucleotide model comparison step first, and then used the nucleotide substitution matrix chosen in step 1 to fit substitution models, which were all codon models in this paper except for the models used in the search for fast-evolving introns that could be used to infer the species phylogeny, and test hypotheses. For all codon analyses we used the MG94 model of codon substitution crossed with the chosen nucleotide substitution matrix.

The nucleotide model comparison was performed assuming gamma distributed among-site rate variation (fit with four discrete bins), so in effect it is limited to comparing the fit of the substitution matrix. The HyPhy method file used was NucModelCompare.bf, which utilizes the Akaike Information Criterion (AIC) and hierarchical likelihood ratio tests to compare the fits of the 203 substitution matrices that are simplifications of the general time reversible (GTR) matrix. Two substitution matrices are returned by this method. One is the best-fit matrix according to AIC, and the other is the matrix closest to the model-averaged nucleotide substitution rates. We chose to use the latter for all our analyses, as for cases where there is little to no uncertainty as in the best-fit matrix the model-averaged matrix will be the same as the AIC-chosen matrix, and when there is uncertainty it is taken into account in choosing the matrix. Using the model-averaged matrix is also advantageous because our datasets have a relatively low number of species.

Our HyPhy scripts are available upon request, and will be submitted as a user contribution to the HyPhy community forum.

Phylogeny methods

In order to reconstruct the phylogeny of the species used in this study, 27 fast evolving loci were mined from our transcriptome datasets (6 species, including the M82 variety of tomato) and the tomato and potato reference genomes; the studied species are all very close relatives, thus the need for fast evolving loci. This phylogeny, rooted with

potato, was used for all downstream analyses of codon sequence evolution (topology only) and gene expression evolution (topology and chronogram branch lengths). Ten genes (coding sequence only) and 17 introns comprise the 27 loci. All chromosomes are represented among the chosen loci. Regions of the genomes of the tomato varieties Heinz and M82 that from our analysis we believe to be introgressed from *S. pimpinellifolium* were excluded from the locus search.

332 introns and 11,221 genes were screened for phylogenetic informativeness. The 332 introns were mined from our data with minimum length and coverage cutoffs of 200 bp and 80%, respectively, and assembled as detailed above for the gene coding sequences. Loci were chosen to maximize information. Thus, we chose genes with the fastest codon substitution rates among those in each of several locus length ranges above 1000 bp. HyPhy v2 (15) was used to estimate mean alignment-wide codon substitution rates for the genes (see further methods below). We used the sum of the synonymous and non-synonymous rates to screen the genes. We used the PhyDesign webserver for profiling phylogenetic informativeness to screen the 332 introns (16, 17). The final selection of the 27 loci was made after careful quality control to ensure that none of our chosen loci appeared to be chimeras caused by the mis-mapping of reads from a paralogous locus. The 27 loci were concatenated prior to phylogeny reconstruction.

Two preliminary trees were reconstructed using MrBayes v3.2.1 (18), a tree with unconstrained branch lengths (i.e., non-clock) and a strict clock tree. The branch lengths of these trees were subsequently used to choose a relaxed clock model and to set its priors. Default priors were used for the non-clock and relaxed clock trees. For all tree inferences the data were partitioned by locus, and the GTR + Γ model was used. Nucleotide frequencies were unlinked across partitions. The substitution rate matrices were linked for sets of loci found to have highly correlated rates in a prior non-clock analysis, in which we unlinked both nucleotide frequencies and substitution matrices. The gamma shape parameter for modeling among-site rate variation was linked across all partitions and the rate distribution discretized with 12 bins. No locus multipliers were used.

The Independent Gamma Rates (IGR) Bayesian relaxed clock model, as implemented in MrBayes v3.2.1, was used to simultaneously infer the topology and the

time-calibrated branch lengths of the species tree (19). The branch lengths were only calibrated in a relative sense, since no fossil data were available for our taxa and our calibration thus consisted of constraining the tree height equal to 1. IGR is an uncorrelated continuous clock model that estimates a common gamma distribution from which the branch rates are drawn independently. The IGR model was chosen because there appears to be no autocorrelation of branch rates in our dataset, as evidenced by a non-significant one-sided two-sample Kolmogorov-Smirnov test ($p = 0.6183$) that the log rate ratios of parent-offspring pairs are lower than those of random branch pairs. The branch rates used for this test are rough rate estimates obtained by dividing strict clock by non-clock branch lengths. Plotting the log rate ratios of parent-offspring pairs against 1000 resamples (n = number of parent-offspring pairs = 12, no replacement) of those of random pairs also reveals the lack of difference in the two distributions of log rate ratios. Moreover, using the resampling of random branch pairs to calculate 1000 Kolmogorov-Smirnov tests yields only three tests out of 1000 (0.3%) with a p-value below 0.05, which further suggests a lack of branch rate autocorrelation.

Following Ronquist et al. (18, 20),we set the median of the exponential prior for the IGR variance increase parameter equal to the slope of the linear regression of non-clock branch length variance as a function of strict clock branch length. Likewise, we used a lognormal prior for the clock base rate, with a mean equal to the strict clock rate and a logarithm of the standard deviation equal to 0.3, so as not to make the prior too informative. A uniform branch length prior was used for the IGR analysis. The strict and relaxed clock analyses were rooted with potato reference sequences (*Solanum phureja*).

Six independent metropolis-coupled MCMC chains were run in parallel, each with one cold and three heated chains and using the default value of the heating parameter. The chains were sampled every 1000 generations and run for 20 million generations. A burn-in of 10 million generations was used, after diagnosing convergence with Tracer v1.5. Data files and MrBayes commands and log files are available upon request.

Methods for the bottleneck analysis

Custom HBL wrappers for the TestBranchDNDS.bf HyPhy method file were used to test for evidence of a bottleneck effect on genome-wide dN/dS (i.e., the non-

synonymous to synonymous substitution rate ratio); the GTR nucleotide substitution matrix crossed with the MG94 model of codon substitution was used for this inference. The expectation is that lineages that have recently gone through a population bottleneck will have fixed a higher proportion of slightly deleterious alleles than other lineages (due to the effect of genetic drift on species with smaller effective population sizes), which will be reflected in higher average genome-wide dN/dS. We were specifically interested in testing for bottlenecks in the domesticated tomato lineage and in the Galapagos Islands' endemic species *S. galapagense*, as domestication and island dispersal are expected to involve bottlenecks.

The TestBranchDNDS.bf test uses a null hypothesis of a single omega (i.e., ω = dN/dS) for the whole tree and an alternative hypothesis that divides up the tree into a background ω portion and a set of focal branches whose omegas are allowed to vary independently. The focal branches used were the three branches of the domesticated tomato lineage (the stem branch and the terminal branches of the M82 and Heinz cultivars) and the *S. galapagense* branch. An additional likelihood ratio test was performed using the above alternative hypothesis as the null, and an alternative hypothesis that adds the parent branch of the tomato stem and *S. galapagense* branches to the set of focal branches. This was done to test whether or not, despite the inferred species tree topology, the bottlenecks due to domestication and island dispersal occurred independently.

The 11,221 gene coding sequence alignments (without potato sequences) that exclude genomic regions of tomato cultivars M82 and Heinz apparently introgressed from *S. pimpinellifolium* were concatenated for this analysis and treated as a single locus. Excluding potentially introgressed regions was done to avoid an artifactual upward bias in the estimated omegas of M82 and Heinz due to the introgressions causing incongruence between the gene tree topologies and the species tree topology assumed to calculate the codon rates.

Given that p-values depend on both effect and sample sizes and that we have a very large sample size of 11,221 genes, we decided to not rely exclusively on p-values and explore the uncertainty in our ω estimates by means of 1000 bootstrap resamples (by codon) of the concatenated alignment.

Methods for the estimation of mean alignment-wide codon substitution rates

Custom HBL scripts were written to calculate for 11,751 genes (see above for alignment construction methods) the mean alignment-wide dN/dS estimates using maximum likelihood (ML), and mean alignment-wide estimates of dN and dS using a distance method. Both the ML and distance inferences were performed using the model-averaged nucleotide substitution matrix (see above) crossed with the MG94 model of codon substitution. For genomic regions that we believe to be introgressed from *S. pimpinellifolium* into either or both M82 and Heinz tomato cultivars, we respectively dropped either or both terminal taxa from the alignments and the fixed species tree topology used to infer the rates. These mean alignment-wide estimates of ω, dN, and dS were used for chromosomal sliding window plots, as well as for screening genes for phylogenetic informativeness once the introgressed regions were filtered out.

Methods for the positive selection scan of coding sequence

A scan for positive selection using the dN/dS ratio was conducted on the 11,221 genes without *S. pimpinellifolium* introgression. The non-synonymous to synonymous substitution rate ratio (ω = dN/dS) is commonly used as a measure of the magnitude and sign of selection on coding sequence(21, 22). Custom HBL wrappers for the PARRIS.bf HyPhy method script (23) were used to fit models in the M3 framework (24, 25) to our data, yet using two synonymous rate classes rather than a single one to allow for some heterogeneity in synonymous rates among sites. Thus, our null model had two synonymous rate classes and two omega ratio classes, of which the highest is constrained to not exceed 1. Likewise, our alternative model had two synonymous rate classes and three omega ratio classes. These models are so-called sites models and were thus used to test for evidence of pervasive positive selection anywhere in the alignment by means of a likelihood ratio test (LRT).

In order to account for nucleotide substitution bias, the model-averaged nucleotide substitution matrix crossed to the MG94 codon model (22) was used to infer the codon rates. The LRT p-values from this positive selection scan were corrected for multiple testing using the FDR method of Benjamini and Hochberg after filtering out genes with

less than four total expected substitutions (i.e., (dN + dS) * gene length). At an adjusted p-value cutoff of 0.2 we found 103 genes with evidence of pervasive positive selection. Altogether 381 genes cross the nominal 0.05 unadjusted *p*-value significance threshold.

Visualization and plotting of genomic data

Circular displays of SNP and gene expression data were generated using Circos software(26).

GO annotation

Blast2go(27), Interproscan(28) and Annex(29) were used to augment ITAG GO annotations. A blast cutoff of 1e-6 was used for Blast2go analysis. GO enrichment was analyzed with the GOseq package (30) in Bioconductor. GO annotations are available in Dataset S6 and S7.

Arabidopsis best-hit annotation

To help with biological interpretation we annotated each ITAG2.3 predicted protein with its best hit from *Arabidopsis*. First, blastp was used to query an *Arabidopsis* TAIR10 peptide database with the ITAG2.3 predicted proteins. We retained the best hit from each query required that the e-value was < e-20, that more than 75% of the query was covered, and that there was greater than 50% identity. These annotations are available in Dataset S8.

Analysis of S. pennellii and S. lycopersicum transcriptome differential gene expression

Sequences were filtered, trimmed and mapped to the appropriate matched cDNA reference as described above. To reduce loss of counts due to inefficient mapping of paired ends to the shortened matched cDNA sequences, we mapped only the first paired end for our expression analysis. Matched cDNA references were screened out of the analysis if reads from both species showed biased mapping to one of the species specific matched cDNAs (LF > 1 for both *S. lycopersicum* and *S. pennellii* reads). Poor samples were identified and removed using a combination of replicate correlation coefficient,

correlation plots, and MA-plots. The raw count data was then normalized using a modified Trimmed Mean of M values method(31). Low expressed genes were filtered on a minimum sum of twenty counts over all samples for further analysis. Genes that did not pass this threshold were considered not expressed. Differential expression was calculated by fitting a quasi-Poisson generalized linear model at the gene level using tissue, species, and tissue by species interaction as factors and extracting significance using a F-test. P-values were adjusted using the Benjamini and Yekutieli method(32) and genes were considered statistically significant if they passed a threshold of $p < 0.01$ and an absolute log-fold change threshold of one for species effect. For tissue and interaction effects we used only the p-value threshold to call significant genes. Visualization of data was accomplished using the R packages vennerable (https://r-forge.r-project.org/projects/vennerable/), ggplot2(33), and gplots (http://cran.r-project.org/web/packages/gplots/index.html).

Analysis for S. lycopersicum, S. pennellii, S. habrochaites, and S. pimpinellifolium seedling gene expression

Analysis was performed as with the transcriptome data with the following modifications. *S. pimpinellifolium* and *S. habrochaites* were aligned to the *S. lycopersicum* or *S. pennellii* matched cDNA references respectively. To determine differential expression, we first identified genes significant for species effect using the above method. Then, each pairwise comparison was individually reiterated through the model using only genes significant for the overall species effect. Multiple testing correction was done as above. A p-value threshold of 0.01 was used to select genes significantly differentially expressed in pairwise comparisons. The total number of significant differences in pairwise comparison was used as a distance matrix for construction of the neighbor-joining tree using the R package ape. Genes classified as specifically differentially expressed in a particular lineage were called only if all species were unambiguously divided into two groups by significant contrasts.

Analysis of S. lycopersicum var. M82, S. pennellii, and IL4-3 differential gene expression

Sequences were mapped and filtered as in the transcriptome experiment. Count data were generated from the average of counts found for the two matched cDNA references. The previously applied statistical model is not appropriate for the smaller sample size of this experiment, so differential expression for the M82 vs. *S. pennellii* and M82 vs. IL4-3 comparisons was called using the R package DESeq. The resulting significance values were used to identify genes differentially expressed at a p-value of 0.05 and an absolute log2 fold change greater than one.

Characterizing patterns of selection and drift on gene expression

A four taxon tree made by pruning the outgroup (*S. phureja*), the two ingroup species without gene expression data (*S. galapagense* and *S. chmielewskii*), and the Heinz cultivar from the species tree was used for phylogenetic continuous trait modeling of gene expression. The median chronogram from the Bayesian tree reconstruction was used for this analysis. The model-fitted gene expression values were log2-transformed prior to analysis.

We followed a model comparison approach with Ornstein-Uhlenbeck (OU) and Brownian Motion (BM) models in order to examine the influence of drift and different modes of selection on gene expression(34, 35). The R package geiger was used to fit single optima OU models to the 20,438 genes in the seedling gene expression dataset(36). Estimates of the rate and attraction parameters were used to calculate the equilibrium variance, which is a measure of stabilizing selection with smaller values indicating higher constraint on expression values(37). The OU equilibrium variance is equal to the rate of diversification (σ^2) divided by two times the attraction (2λ); lower values of this measure indicate more constrained patterns of expression divergence and higher values indicate less constrained divergence. Because we did not scale the gene expression values, there is no *a priori* neutral expectation for the value of the equilibrium variance. However, this statistic is informative of the strength of stabilizing selection, and its distribution can reveal broad patterns in the transcriptome. The number of sampled species precluded the fitting of multiple optima OU models to the data.

One and two-rate BM models (BM1 and BM2, respectively) were also fit to the data using the brownie.lite() function in the R package phytools(38). Following the convention, we used a ΔAIC greater than 4 to indicate moderate support for the BM2 model, and a ΔAIC greater than 7, greater than 10, and greater 20 to indicate progressively stronger support. We also used ΔAIC to compare the fit of the BM2 model to that of a single optimum OU model (OU) for each gene. Better fit by BM2 over *both* OU and BM1 models was taken as evidence that the gene in question had experienced a pattern of accelerated evolution along the focal branch(es), rather than a pattern characterized by either stabilizing (OU) or neutral selection (BM1); a pattern of stochastically changing trait optima is also well fit by a BM1 model(38). To avoid issues with singular matrices, the least variable 1% of genes was discarded prior to analysis.

For the BM2 model, the phylogenetic tree branch(es) along which the second BM rate operates was treated as an additional model parameter; the seven rate regimes assayed were each of the four terminal branches, the internal branch separating red- and green-fruited species, and both red- or green-fruited species with the same alternative rate. Hence, for each gene, the best-fit BM2 model (i.e., among the seven regimes) was compared to the null model, BM1. Because of the low number of species sampled the X^2 distribution may not be an appropriate probability distribution for the likelihood ratio statistic, so we opted to use the difference in Akaike Information Criterion (ΔAIC) between models as a relative measure of goodness of fit.

Plant material for RT-PCR and qRT-PCR validation

For RT-PCR and qRT-PCR validation experiments the following tissues were collected. Around 15 seeds (per plate) of both *S. lycopersicum* and *S. pennellii* were placed on a ½ MS media in petri dishes and incubated for 3 days in the dark wrapped in aluminum foil under appropriate conditions. The plates were grown vertically using the same chamber conditions used for the RNAseq experiments. After 10 days, samples of seedlings (shoots only) were collected. On day 50, tissues were collected from *S. lycopersicum* floral apices that include the inflorescence meristem. At this time mature stems (4th developed internode) and leaves were also collected. The same tissues were collected a few days later from *S. pennellii*. For vegetative meristems, tissue was

collected from both *S. lycopersicum* and *S. pennellii* plants. The appropriate stage was the 3rd leaf stage (meristem + 2 leaves). Overall the following tissues were collected from both *S. lycopersicum* and *S. pennellii:* roots, seedling, inflorescence meristem, mature stem, mature leaves, and vegetative meristem (at 3rd leaf stage).

RNA extraction and cDNA synthesis

Total RNA was extracted from tomato tissues with Trizol Reagent (Invitrogen) using manufacturer's protocol and cleaned by sodium acetate/ethanol precipitation. All RNA samples were first treated with DNaseI (QIAGEN) according to manufacturer's instructions. First strand cDNA synthesis was performed using 1 μg of total RNA with oligo (dT) using Superscript III Reverse Transcriptase (Invitrogen) according to manufacturer's instructions. The cDNA was diluted to 1:100 and used as a template for RT-PCR amplifications.

Primers for differential expression validation

All primer sets for RT-PCR and qRT-PCR validation were designed using either GeneScript (https://www.genscript.com/ssl-bin/app/primer) or BatchPrimer3 (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi) primer design tools. Primers used are listed in Dataset S9.

Molecular marker design

For each SNP between *S. lycopersicum* var. M82 and *S. pennellii* we used custom Bioperl scripts to generate 16 bp long alleles for each species containing the SNP in the middle(14). These alleles were queried for the modification of a restriction site. We used Primer3(39) to design primers on the 500 bp region surrounding the SNPs that produced a restriction site deletion or insertion between the species. For indels, we used Primer3 to design primers using in silico-generated alleles on the 500 bp region surrounding the polymorphism. We selected primer combinations based in the restriction enzyme revealing the polymorphism or in the estimated size of the resulting fragments to ensure proper detection in an agarose gel. Primers used are listed in Dataset S9.

Reverse Transcription PCR (RT-PCR)

The PCR amplification conditions involved a 98°C hold for 2 min, followed by a 30 cycles at 98°C for 30 s, 55°C for 30 s, 72°C for 30 s and a final extension at 72°C for 7 min. The tomato Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene was used as housekeeping control. The RT-PCR products were separated on a 1% agarose gel.

Quantitative Real-Time PCR (qRT-PCR)

Quantitative Real-time PCR was done using the SYBR Green PCR master mix (Biorad) and an iQ5 Real-Time PCR detection system (Biorad, Hercules, CA). Gene specific primers (Dataset S9**)** were designed for all the genes using the Primer3 program v. 0.4.0(39). A standard curve was constructed using serial dilutions of RT product prior to qRT-PCR validations and the efficiency of each primer set was determined using the equation [(10 (-1/-slope)-1)·100]. Amplification efficiency was performed on the software and only those PCR primers with amplification efficiency between 95-110% were used. A melt curve analysis was performed following amplification to confirm specificity of products over primer dimers, and a no reverse transcriptase control was used to ensure products were amplified cDNA rather than genomic DNA. The PCR involved a 95°C denaturation step for 10 min followed by 40 cycles of a two-step PCR protocol as follows: denaturation at 95°C for 15 s and annealing/extension at 60°C for 1 min. A further 71 cycles at 55°C for 30 s was followed later on. A sample volume of 20 µl was used for the analysis, which contained a 1X final concentration of SYBR green PCR master mix, 200 nM gene specific primers and 1 µl of template. All reactions were carried out in triplicate using 96 well plates and the data were analyzed with iQ5 optical system software. Relative expression levels between *S. lycopersicum* and *S. pennellii* were then calculated as fold changes and then converted into log2 FC (Fold Change) values. For each qRT-PCR, a specific housekeeping gene (GAPDH) was chosen for normalization that did not exhibit any significant change in expression. Each qRT-PCR was carried out three times.

Genomic PCRs

Genomic DNA was extracted from germinated seedlings (2–3 days old) of *S. lycopersicum* and *S. pennellii* using a fast and pure modified Dellaporta plant DNA extraction method(40). Genomic DNA was diluted to 50 ng/µl and used as a template in a PCR final volume of 10 µl containing 1X standard PCR buffer, 200 µM of each dNTP, 0.25 µM of each primer, 0.1 µl of Taq DNA polymerase and 1µl of template DNA. The reaction included a 2 min denaturation at 94°C followed by 35 cycles of PCR (94°C for 30 s; 55°C for 30 s; 72°C for 30 s) with a final extension time of 7 min at 72°C. The PCR products were separated on a 2% agarose gel to detect the INDELs.

Resequencing

For resequencing, genomic DNA extractions and genomic PCRs were performed as above. PCR products were then purified using AMPure XP magnetic beads (AGENCOURT) according to the manufacturer's protocol for preparation for sequencing. The amplified PCR products were sequenced with an ABI 3730 Capillary Electrophoresis Genetic Analyzer. Sequence traces for each putative SNP identified were inspected visually to verify sequence polymorphisms.

PCA of global vs. tissue-specific changes in gene expression

Principle component analysis to detect relationships between samples in the transcriptome experiment was performed using mean scaled fitted expression values and the prcomp function in the R statistical computing environment.

To further examine how much variance in gene expression values across tissues between the two species was due to global vs. tissue-specific changes in gene expression, a PCA was undertaken on logFC (*S. pennellii* minus *S. lycopersicum*) values on a per tissue basis. Data was variance scaled using arguments within the prcomp() function, but unlike other analyses undertaken in this work, not mean-centered, as to preserve magnitude and direction of gene expression changes. PCs were examined by plotting log2FC values on a per tissue basis of genes possessing PC values in the lower or upper quartiles of the distribution. PC1 was observed to represent variance representing overall, global changes in gene expression across tissues. This was verified by a high correlation

between PC1 values and overall logFC (r = 0.999). Other PCs explained variance in logFC values with respect to tissue. Variance explained by each PC is as follows: PC1, 56.4%; PC2, 12.6%; PC3, 10.7%; PC4, 8.3%; PC5, 7.1%; PC6, 5.9%.

Weighted gene coexpression network construction

We constructed the networks independently for each species using the fitted expression values for genes significant for the tissue or tissue by species interaction from the model described above. Genes were further required to be expressed in both species and to have a coefficient of variation greater than 0.1 across all fitted values resulting in a final set of 5,097 genes for analysis. Unsigned network construction and module detection was performed using the WGCNA package for R (41). Optimal power parameters were calculated for each species independently, and then the larger of the two (24, for *S. lycopersicum*) was used for module detection. Module detection was performed as per default excepting a modified tree cut height (0.75). The resulting modules were then combined across species based on overlap. The network diagram was constructed in cytoscape (42) using genes assigned to a module in both species and edges that fell into the top 10% of correlations within this gene set in either species. Species-specific edges were identified as in (43) using untransformed unsigned correlations. In this method, the edge weight is normalized to the mean edge weight in each network before calculating network specificity. We calculated the specificity of the edge using normalized and unnormalized data with similar results. The data shown is unnormalized. Connectivity was calculated for the whole network and for each module as the sum of the untransformed correlations for each gene.

Root metabolites and phenotypes

S. lycopersicum var. M82 and *S. pennellii* seeds were stratified at -80°C for 24 hours prior to sterilization with 70% ethanol, followed by 50% bleach, and rinsed several times in sterile water. Plants were grown on MS medium (4.33 g/L MS, 0.5 g/L MES, pH 5.7) with 1.25 g/L phytagel. Germination was scored and roots were harvested at seven days after germination and frozen in liquid nitrogen before lyophilization.

A 1 cm section was cut from the middle of each tomato root seven days after germination. Sections were embedded in 3% agarose in bundles of three to four roots. These agarose plugs were then stored in FAA at 4°C for at least 16 hours. After rehydration in an ethanol series, the plugs were sectioned using a vibratome into 200 μm sections. These sections were then stained in toluidine blue for the same time period and rinsed in distilled water. Sections were imaged with an Olympus Vanox compound microscope at 20X magnification.

Cell wall analysis

For cell wall analysis, lyophilized 7-day-old root tissue was subjected to a monosaccharide compositional analysis(44). Briefly, cell wall material (alcohol insoluble residue, AIR) was prepared from the root tissue and subjected to trifluoroaceticacid hydrolysis, hydrolyzing the matrix polysaccharides. The resulting monosaccharides were converted into their corresponding alditol acetates and analyzed by GC-MS as described. The hydrolyzate was also subjected to HPAE-chromatography coupled to a pulsed amperometric detector to separate and quantify the uronic acids as described(45).

Leaf phenotypes

S. lycopersicum var. M82 and *S. pennellii* plants were grown on soil under 12 hr/ 12 hr light/dark cycles. Plants were grown for 38 days in a replication of ten. Dental polymer (Heraus Kulzer, Germany) was applied to the middle of both the abaxial and adaxial surfaces of terminal leaflets of the second leaf to produce epidermal surface molds. Clear nail varnish was applied to these molds and subsequently mounted onto slides. Three images were taken per impression on a Nikon Eclipse E600 microscope at a magnification of 20X and analyzed using ImageJ software.

Pavement cell area was determined by tracing three pavement cells per image, three images per plant, and averaging the results. Stomatal pore size was calculated by measuring the length of three closed stomatal pores per image, three images per plant. Again results were averaged to give mean stomatal pore length. Absolute stomatal density was determined by recording the number of stomata per 0.28 mm^2 field of view. Stomatal index was calculated as number of stomata per field of view divided by the total

number of pavement cells plus stomata. The number of pavement cells per image was determined by dividing the area of field of view by the average pavement cell size for each plant. Sample means were compared using a two-tailed Student's t-test.

For leaf sections, small tissue samples were taken from the middle of the distal primary leaflet of the second leaves of 38-day-old *S. lycopersicum* and *S. pennellii* plants. Ten samples were taken for each species and sectioned using a Vibratome® Series 1000 Sectioning System. Sectioned material was observed using a Nikon Eclipse E600 microscope. Three leaf sections were captured per plant and images were analyzed using ImageJ software. Leaf blade thicknesses for each plant were calculated as the average of the three images taken.

References

1. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
2. Jimenez-Gomez JM & Maloof JN (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. *BMC Plant Biology* 9(1):85.
3. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
4. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
5. Depristo MA*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
6. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
7. Zhang Z, Schwartz S, Wagner L, & Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7(1-2):203-214.
8. Wu TD & Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859-1875.
9. Martin J*, et al.* (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* 11(1):663.
10. Simpson JT*, et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6):1117-1123.
11. Huang X & Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9(9):868-877.
12. Consortium TPGS (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189-195.
13. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792.
14. Stajich JE*, et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10):1611-1618.
15. Pond SLK, Frost SDW, & Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5):676-679.
16. Lopez-Giraldez F & Townsend J (2011) PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology* 11(1):152.
17. Baudry E, Kerdelhue C, Innan H, & Stephan W (2001) Species and recombination effects on DNA variability in the tomato genus. *Genetics* 158(4):1725-1735.
18. Ronquist F*, et al.* (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Systematic Biology*
19. Lepage T, Bryant D, Philippe H, & Lartillot N (2007) A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution* 24(12):2669-2680.

20. Ronquist F*, et al.* (2012) A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Systematic Biology*
21. Goldman N & Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11(5):725.
22. Muse SV & Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* 11(5):715-724.
23. Scheffler K, Martin DP, & Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. *Bioinformatics* 22(20):2493-2499.
24. Vonholdt BM*, et al.* (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* 464(7290):898-902.
25. Yang Z, Nielsen R, Goldman N, & Pedersen A-MK (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics* 155(1):431-449.
26. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.
27. Gotz S*, et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36(10):3420-3435.
28. Quevillon E*, et al.* (2005) InterProScan: protein domains identifier. *Nucleic acids research* 33(suppl 2):W116-W120.
29. Myhre S, Tveit H, Mollestad T, & L√¶greid A (2006) Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics* 22(16):2020-2027.
30. Young MD, Wakefield MJ, Smyth GK, & Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 11(2):R14.
31. Robinson MD & Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11(3):R25.
32. Benjamini YY, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29:1165-1188.
33. Wickham H (2009) ggplot2: Elegant graphics for data analysis. *New York: Springer*
34. Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet* 25(5):471-492.
35. Hansen TF & Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50(4):1404-1417.
36. Harmon LJ, Weir JT, Brock CD, Glor RE, & Challenger W (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics* 24(1):129-131.
37. Bedford T & Hartl DL (2009) Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A* 106(4):1133-1138.
38. O'Meara BC, Ané C, Sanderson MJ, & Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60(5):922-933.

39. Rozen S & Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365-386.
40. Dellaporta S, Wood J, & Hicks J (1983) A plant DNA minipreparation: Version II. *Plant Molecular Biology Reporter* 1(4):19-21.
41. Langfelder P & Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.
42. Smoot ME, Ono K, Ruscheinski J, Wang PL, & Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431-432.
43. Oldham MC, Horvath S, & Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A* 103(47):17973-17978.
44. York WS, Darvill AG, McNeil M, Stevenson TT, & Albersheim P (1986) Isolation and characterization of plant cell walls and cell wall components. *Methods in Enzymology*, ed Arthur Weissbach HW (Academic Press), Vol Volume 118, pp 3-40.
45. De Ruiter GA, Schols HA, Voragen AG, & Rombouts FM (1992) Carbohydrate analysis of water-soluble uronic acid-containing polysaccharides with high-performance anion-exchange chromatography using methanolysis combined with TFA hydrolysis is superior to four other methods. *Anal Biochem* 207(1):176-185.

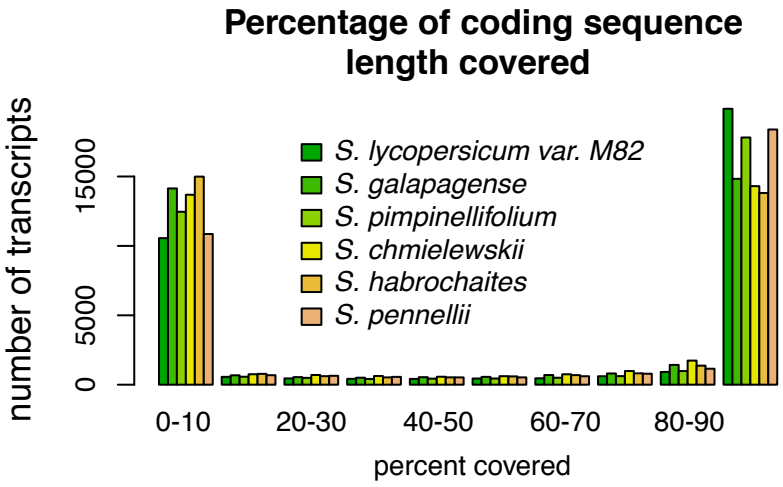

Fig. S1. Percent length of the annotated coding regions covered. For each annotated transcript we calculated the percentage of the coding region length that is covered by 4 or more reads. For each species, we plot the histogram of percent coverage. Most coding regions are either completely covered (90-100% of their length covered) or not covered at all (0-10% of their length covered), indicating that the majority of transcripts present in our samples were completely sequenced.

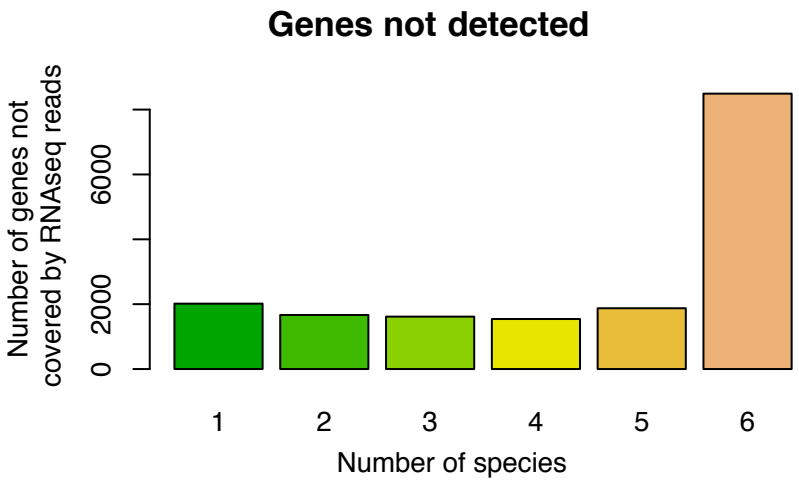

Fig. S2. Number of genes not covered by RNAseq reads.
Histogram of the genes not covered by RNAseq reads in each of the samples analyzed. The majority of the genes not covered in one sample are not covered in any of the other samples.

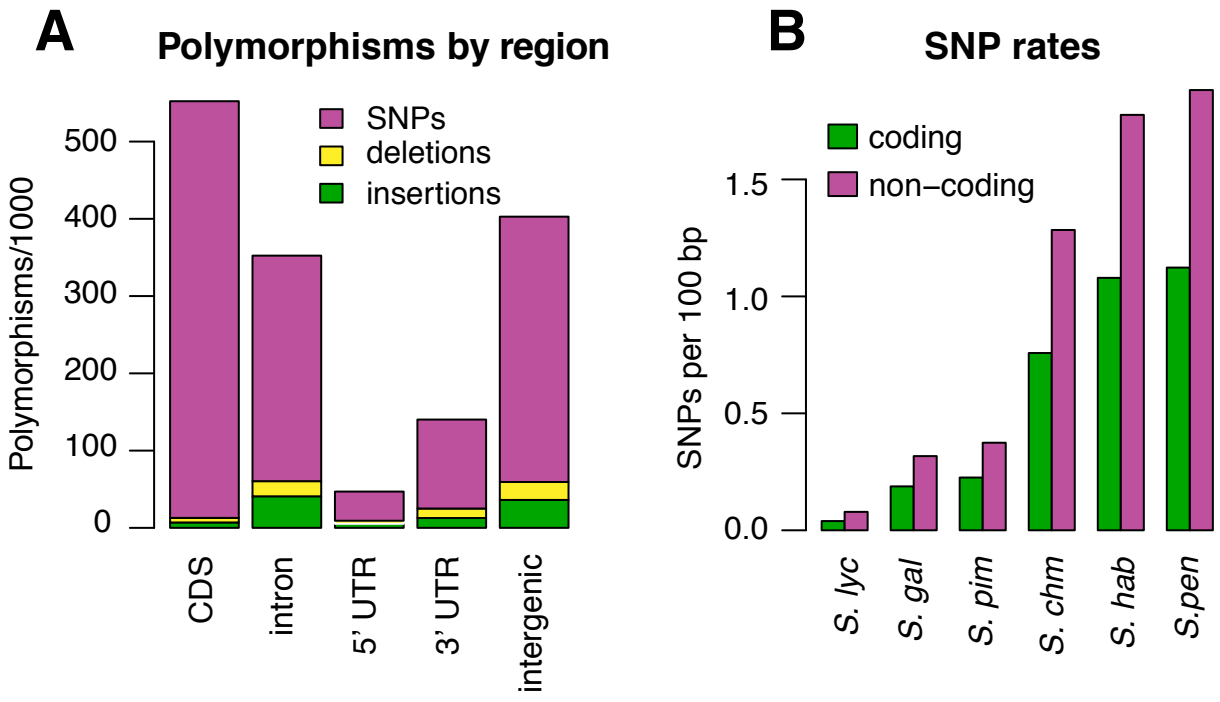

Fig. S3. Polymorphism numbers by region.
A. Barplot of the number of SNPs, insertions and deletions found in each annotated region. Most SNPs were found in coding regions. Insertions and deletions were more abundant in non-coding regions. **B**. SNP rates per species and region. Number of polymorphisms per 100 bp in each species analyzed compared with *S. lycopersicum* var. Heinz. SNP rates increase with phylogenetic distance. SNP rates in all species were higher in non-coding versus coding regions.

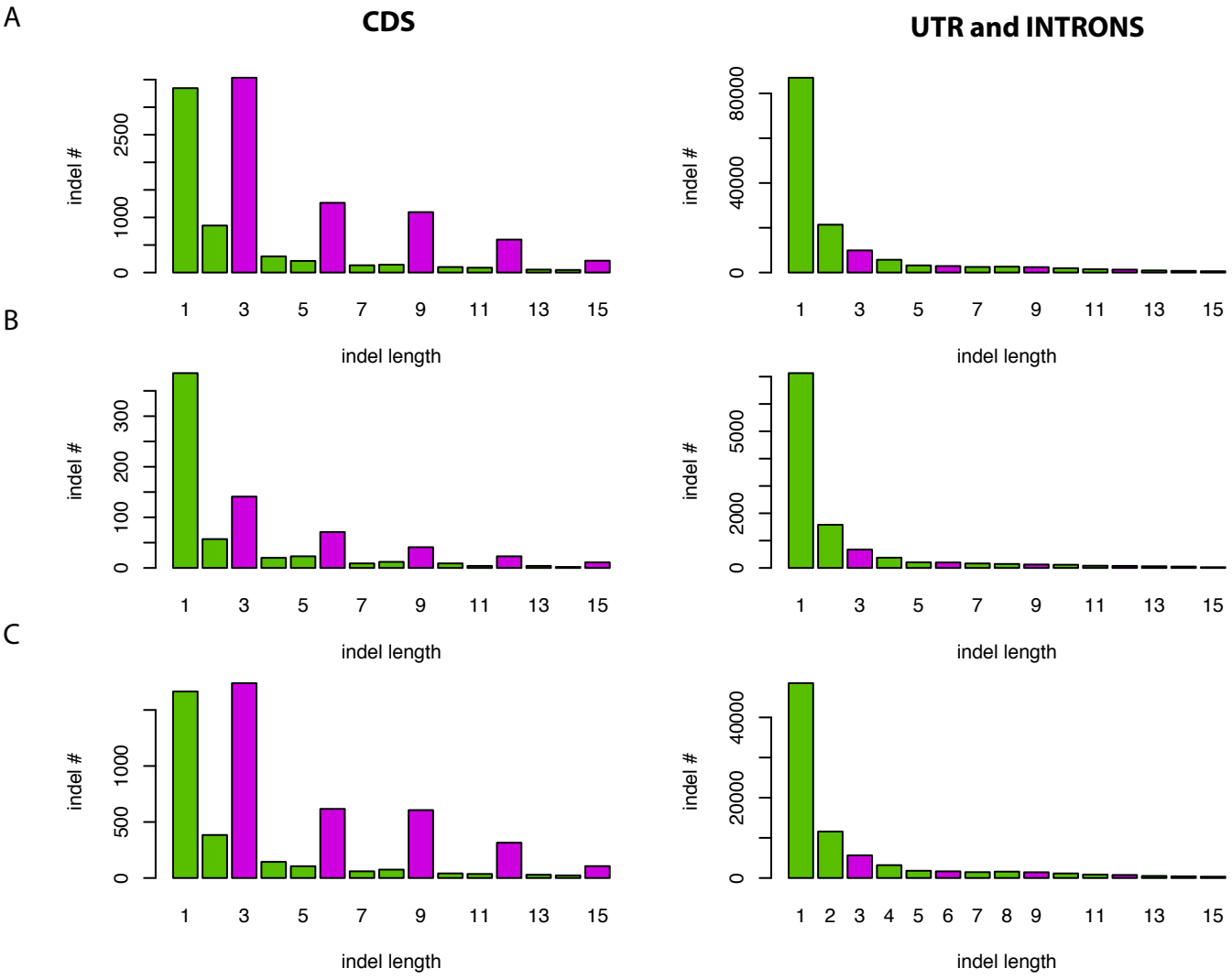

Fig. S4. Length of Insertions and Deletions
Insertion lengths in all samples A., unique to S. lycopersicum and S. galapagense B., or in S. lycopersicum and S. galapagense but not unique to those species C. In coding regions (column 1) there is a bias towards insertions and deletions with lengths in multiples of three, which conserve the open reading frame.
In UTRs (column 2), this bias disappears.

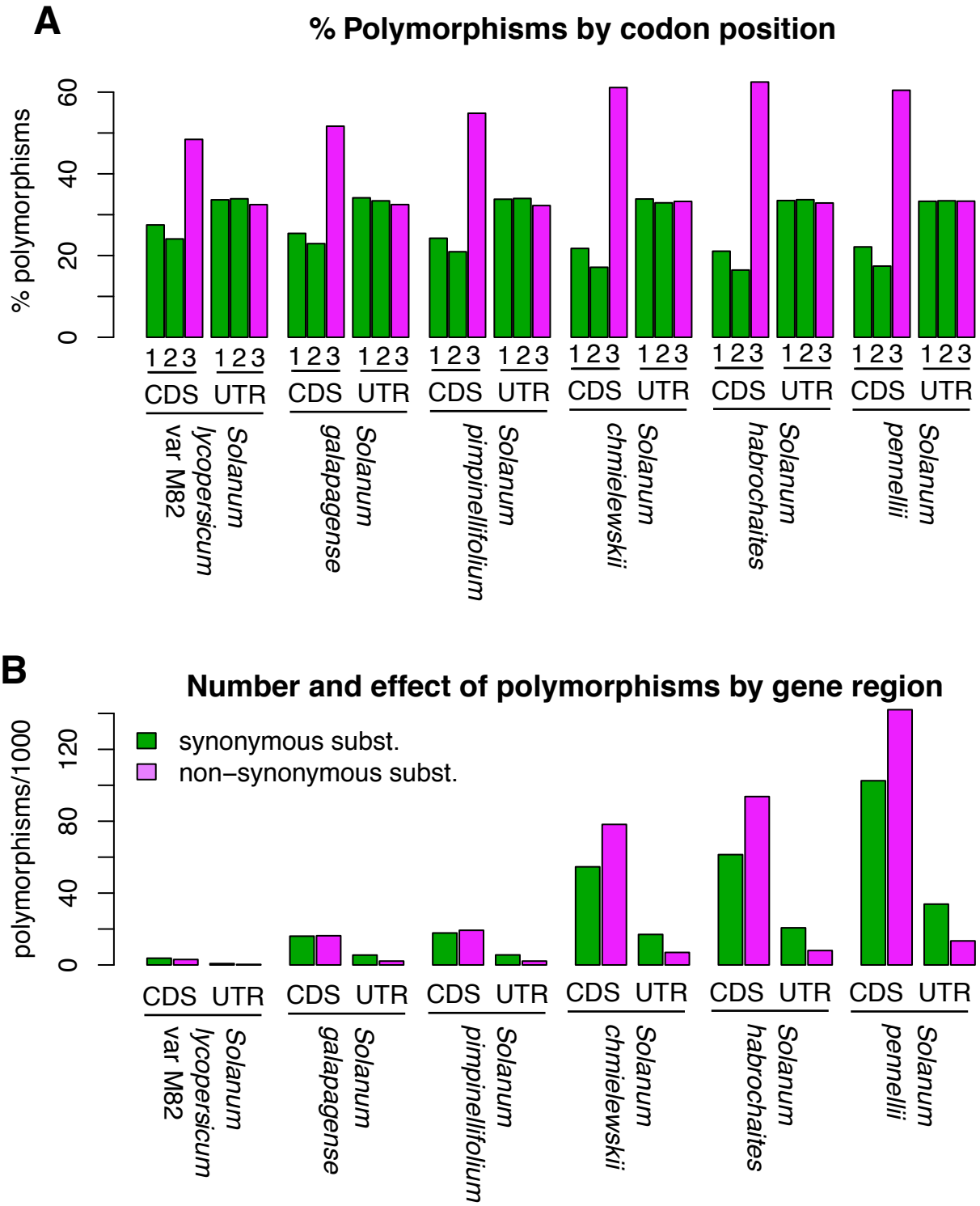

Fig. S5. Effect of polymorphisms.
For each gene, we extended the open reading frame to the UTRs and calculated the position and effect of the SNPs both in the coding region and in the UTRs. **A**. Percentage of SNPs by codon position. In coding regions most SNPs occur in the third position of the codon, the majority of which result in synonymous substitutions. Polymorphisms in UTRs are located equally in all positions of the putative codons. **B**. Number synonymous and non-synonymous SNPs in coding regions and in UTRs. For this analysis we used high quality (quality score>3) and high frequency (frequency>0.9) SNPs. Synonymous substitutions are more frequent than non-synonymous substitutions in coding regions, due to selective constrains. In UTRs non-synonymous substitutions are more frequent due to the higher probability of this type of mutation in regions with relaxed constraint.

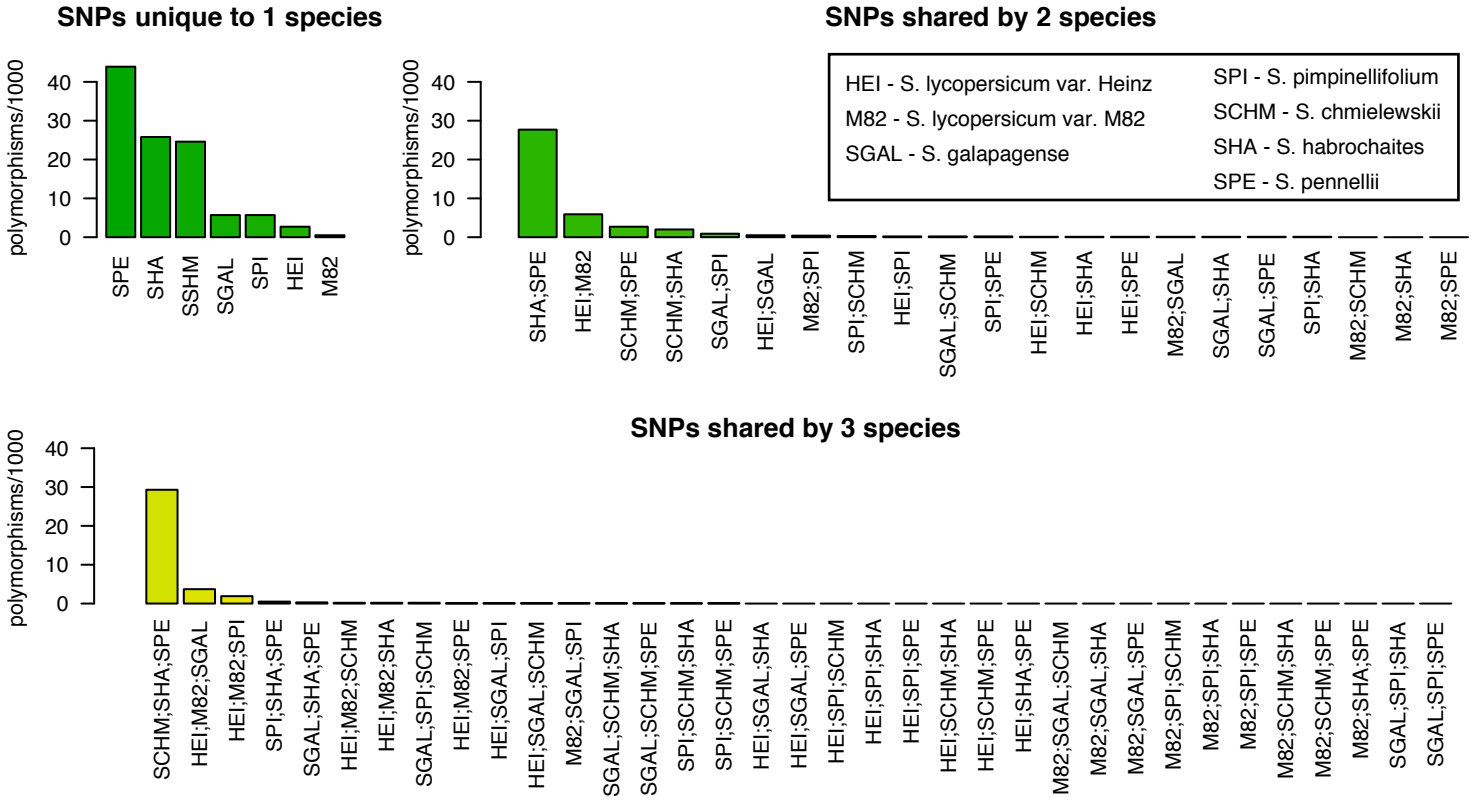

Fig. S6. Allele distribution per species.
Only SNPs in genes, biallelic, homozygous and covered by at least 4 reads in all species were used in this analysis. Y axis represent thousands of SNPs in each species combination. The majority of SNPs private to a single species distinguish *S. pennellii* from everything else. *S. habrochaites and S. chmielewskii* present more than 20,000 private SNPs. *S. pennellii* and *S. habrochaites* share more SNPs than any other two species, and these two species have more SNPs in common with *S. chmielewskii* than any other three species.

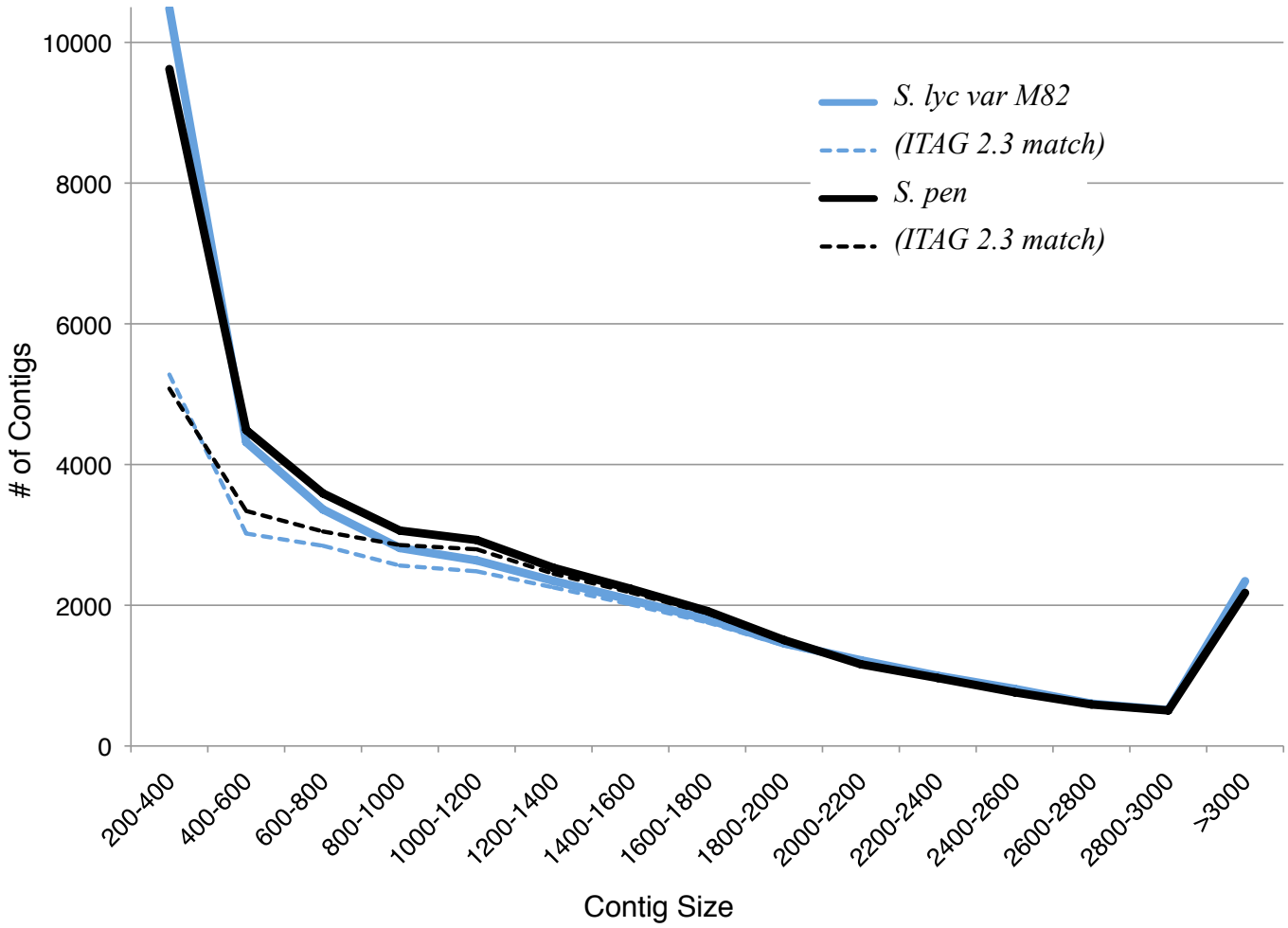

Fig. S7. Number of de novo-assembled contigs broken down by contig size. Solid lines represent total number of *S. lycopersicum* var. M82 (blue) and *S. pennellii* (black) contigs for a given contig size and dashed lines represent the number of contigs with a sequence match in the *S. lycopersicum* var. Heinz ITAG2.3 transcriptome reference sequence.

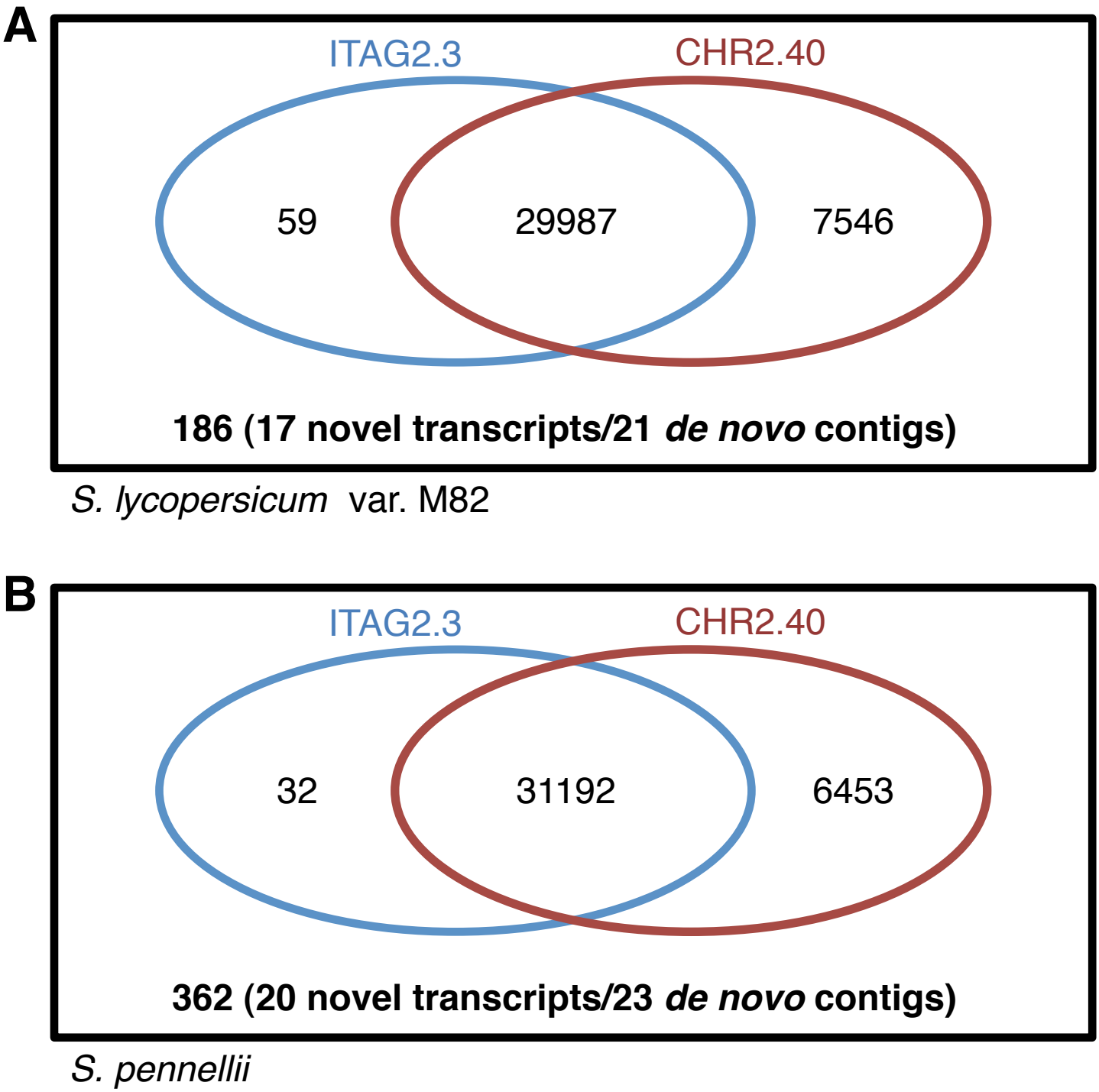

Fig. S8. Although most contigs from the de novo assemblies correspond to reference sequences, a few dozen represent nuclear-encoded plant genes missing from the reference sequences. Venn diagrams display the overlap between the ITAG2.3 transcriptome and the CHR2.40 genome reference sequences of *S. lycopersicum* var. Heinz with de novo assembled contigs of (**A**) *S. lycopersicum* var. M82 and (**B**) *S. pennellii*. Combined, the two de novo assemblies revealed 44 contigs representing 34 unique novel transcripts.

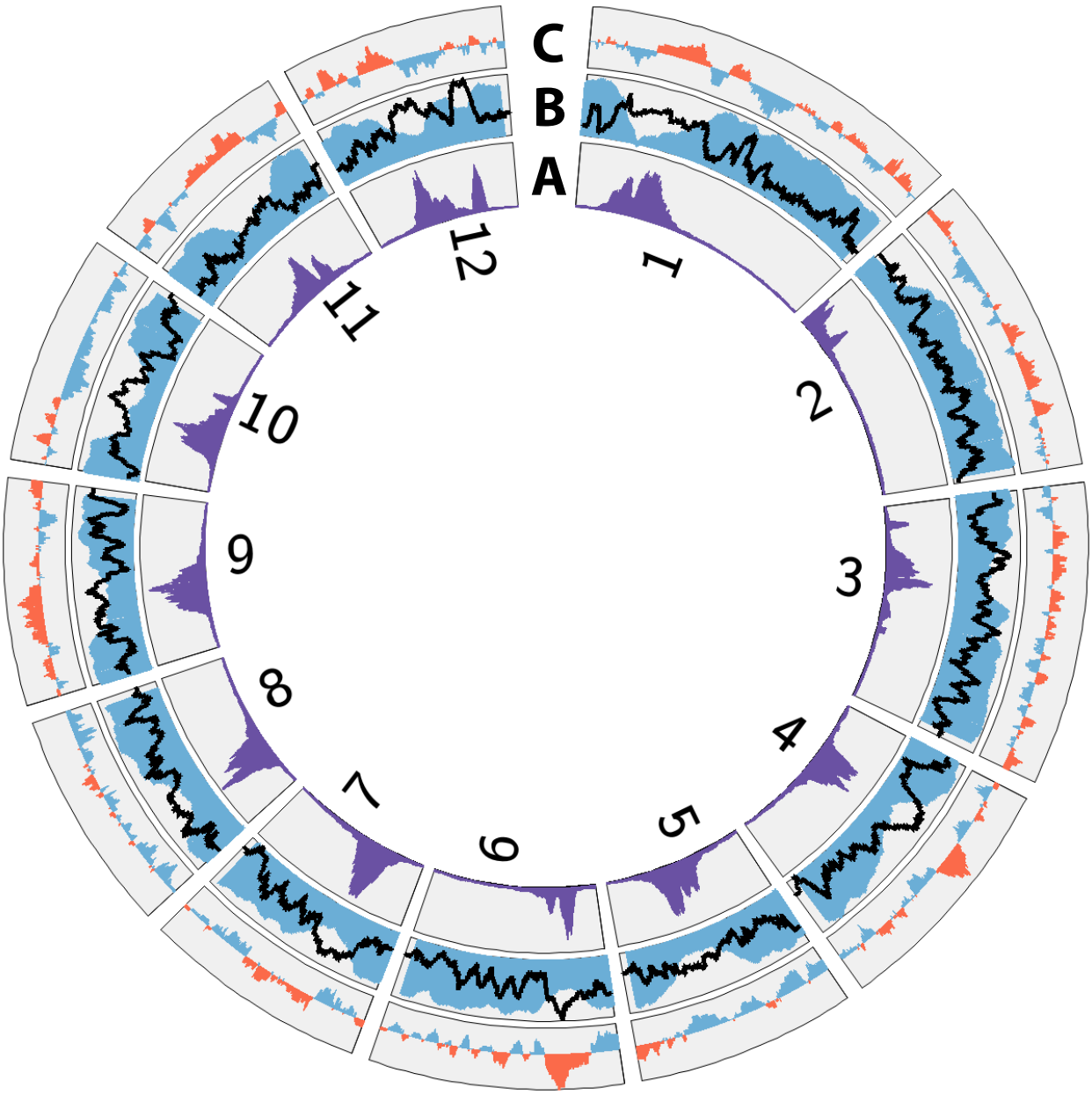

Fig. S9. Pairwise gene expression comparison of *S. lycopersicum* and *S. pennellii* **A**. Mean distance to adjacent gene. **B**. Frequency of detected (blue) and differentially expressed (black) genes in pairwise comparison. **C**. Frequency of differentially expressed genes polarized by direction of change (red: higher expression in *S. pennellii*; blue: higher in *S. lycopersicum*). All plots represent sliding windows of 100 genes. Numbers inside the circle indicate chromosome number. Some centromeres show a bias towards upregulation in *S. pennellii*.

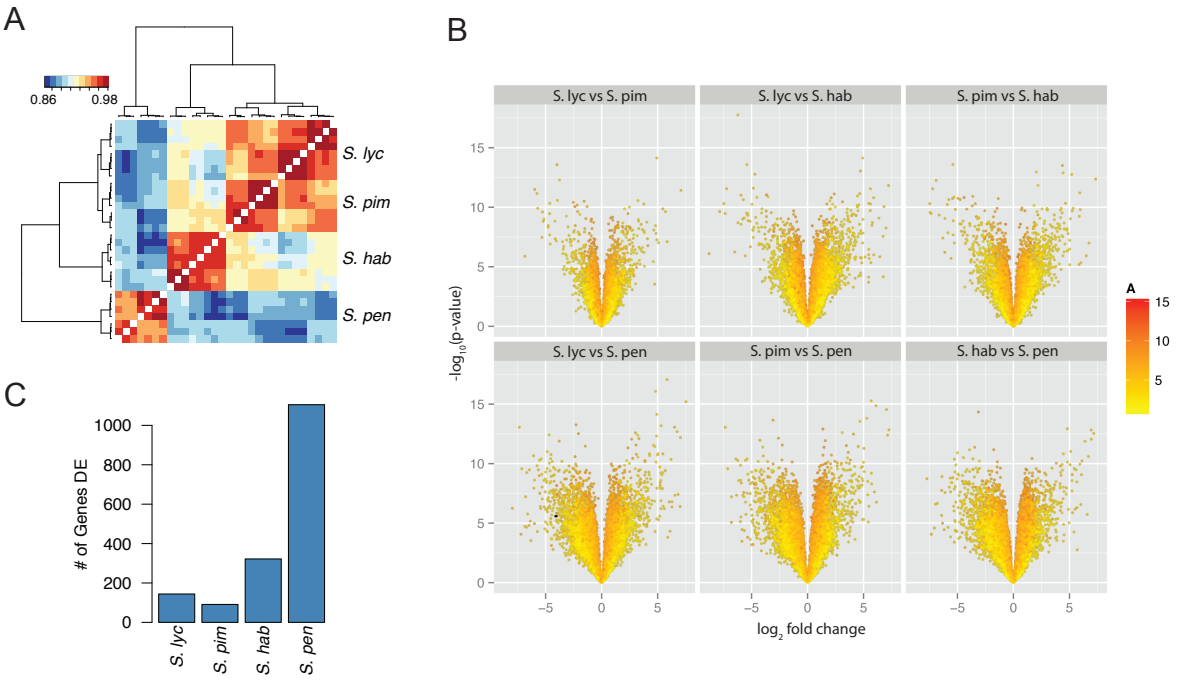

Fig. S10 A. Heatmap depicting correlations between replicate samples in the seedling experiment. Red indicates high correlation. **B.** Volcano plots showing the relationship between logFC and statistical significance. Orange and red indicate higher abundance (counts). **C.** Number of genes specifically diferentially expressed in each lineage.

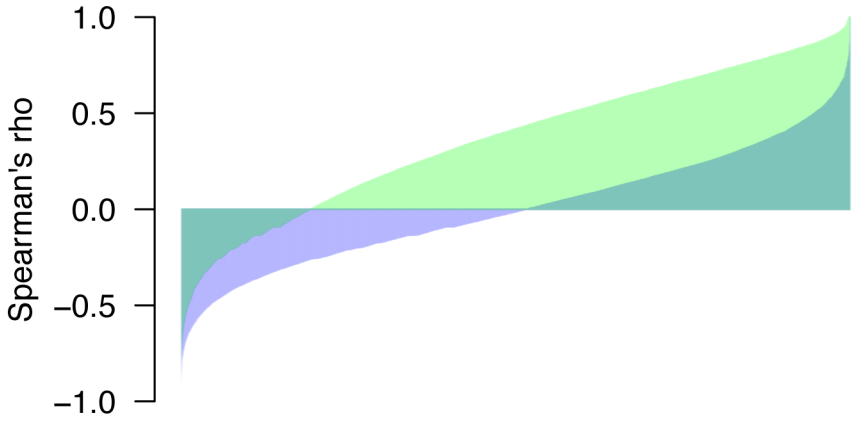

Fig. S11. Spearman correlation in expression pattern across tissues in S. pennellii and S. lycopersicum in green and an example of correlations calculated after permuting the tissue IDs in blue. The true correlation is enriched for positive correlations.

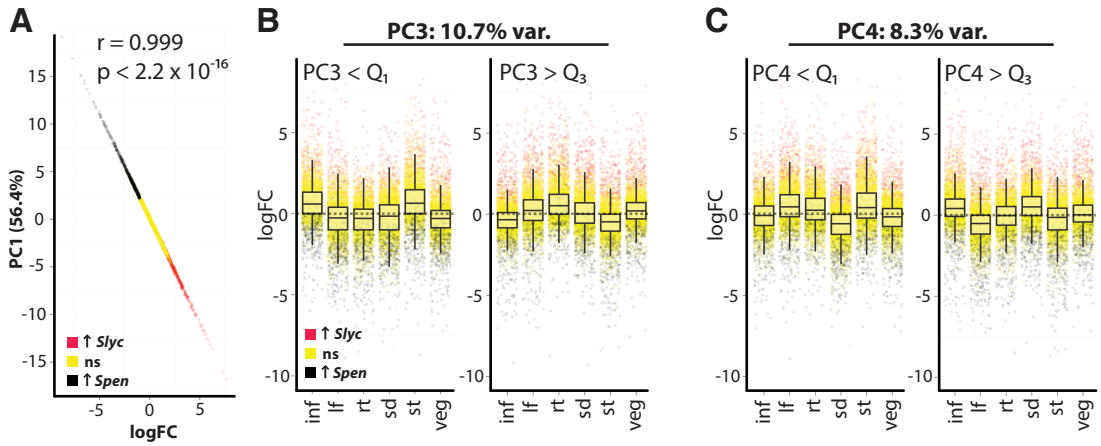

Fig. S12. Principal Component Analysis of log Fold-Change values across tissues between *S. lycopersicum* and *S. pennellii*. **A**. The first principal component, explaining 56.6% of all variance, reflects global changes in gene expression across all tissues. This is reflected by the fact that PC1 values are essentially equivalent to logFC. Shown is a scatterplot of PC1 values vs. logFC, which are highly correlated (r=0.999, p < 2.2 x 10^-16). **B-C**. Principal components besides PC1 (explaining 43.4% of all variance) reflect tissue specific changes in logFC between species. Shown as examples are PC3 (**B**) and PC4 (**C**); see fig. 4 for the first two principal components. logFC values for genes occupying the lowest (<Q1) and highest (>Q3) quartile values for the PCs are shown. Red, genes with significantly higher expression in *S. lycopersicum*; Black, genes with significantly higher expression in *S. pennellii*; yellow, genes with no significant difference in gene expression between species.

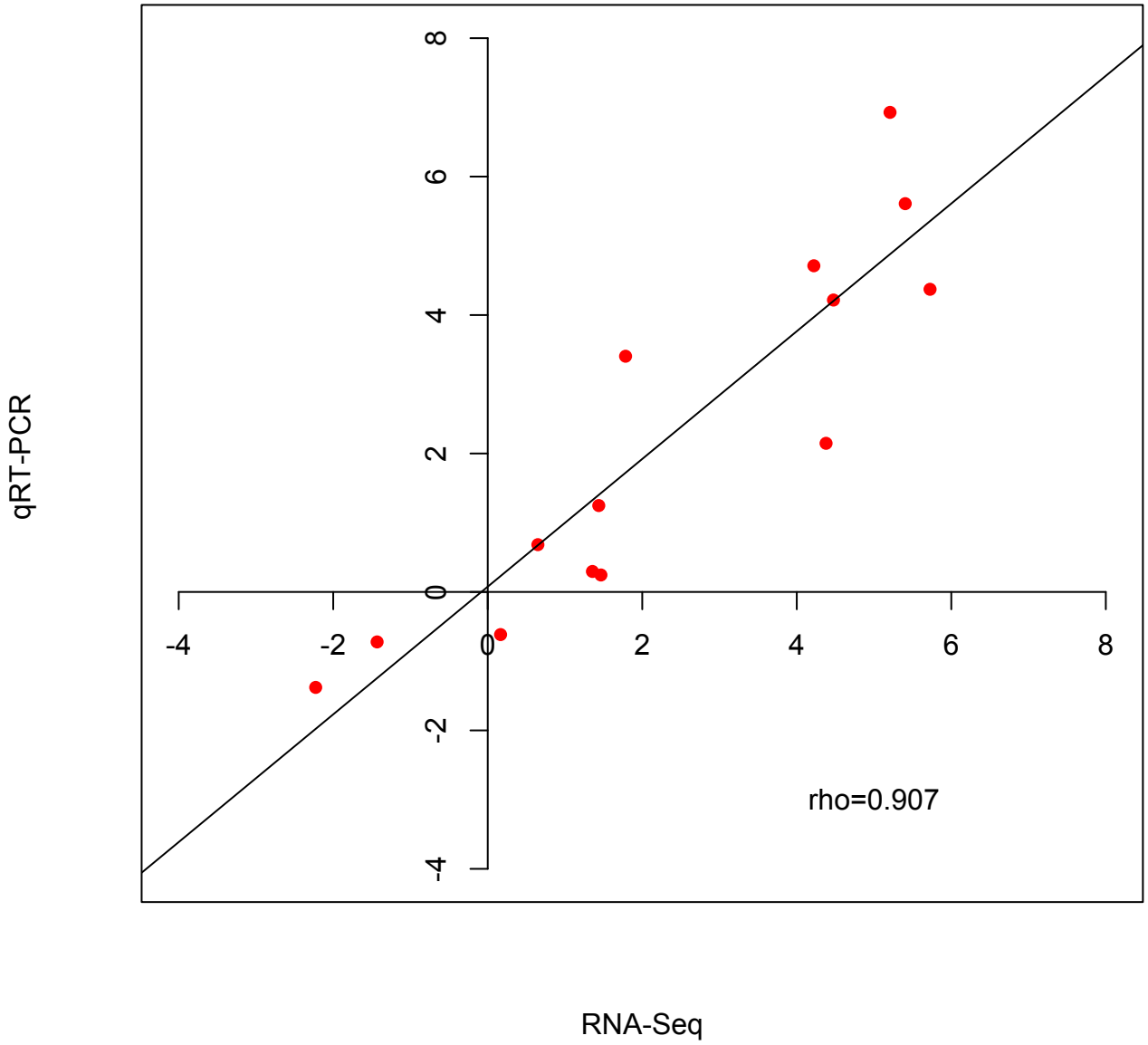

Fig. S13. Correlation between RNAseq reads and qRT-PCR analysis of gene expression. qRT-PCR was performed for 14 genes identified as differentially expressed between *S. lycopersicum* and *S. pennellii*. 'rho' indicates Spearman's correlation coefficient between RNASeq and qRT-PCR

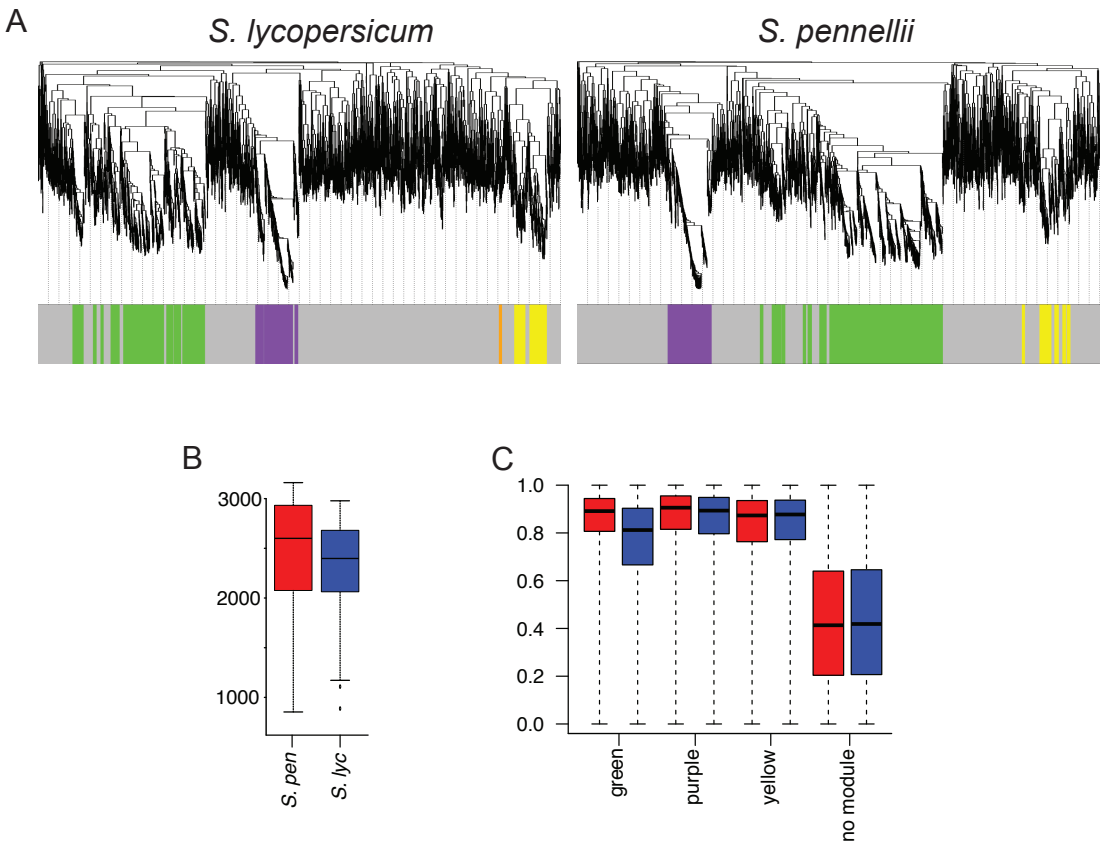

Fig. S14. **A.** Hierarchical clustering tree and module assignments for WGCNA analysis. **B.** Connectivity in the two species for all genes differentially expressed in tissues **C.** Distribution of correlation coefficients between genes for each module in both species (red is *S. pennellii* and blue is *S. lycopersicum*).

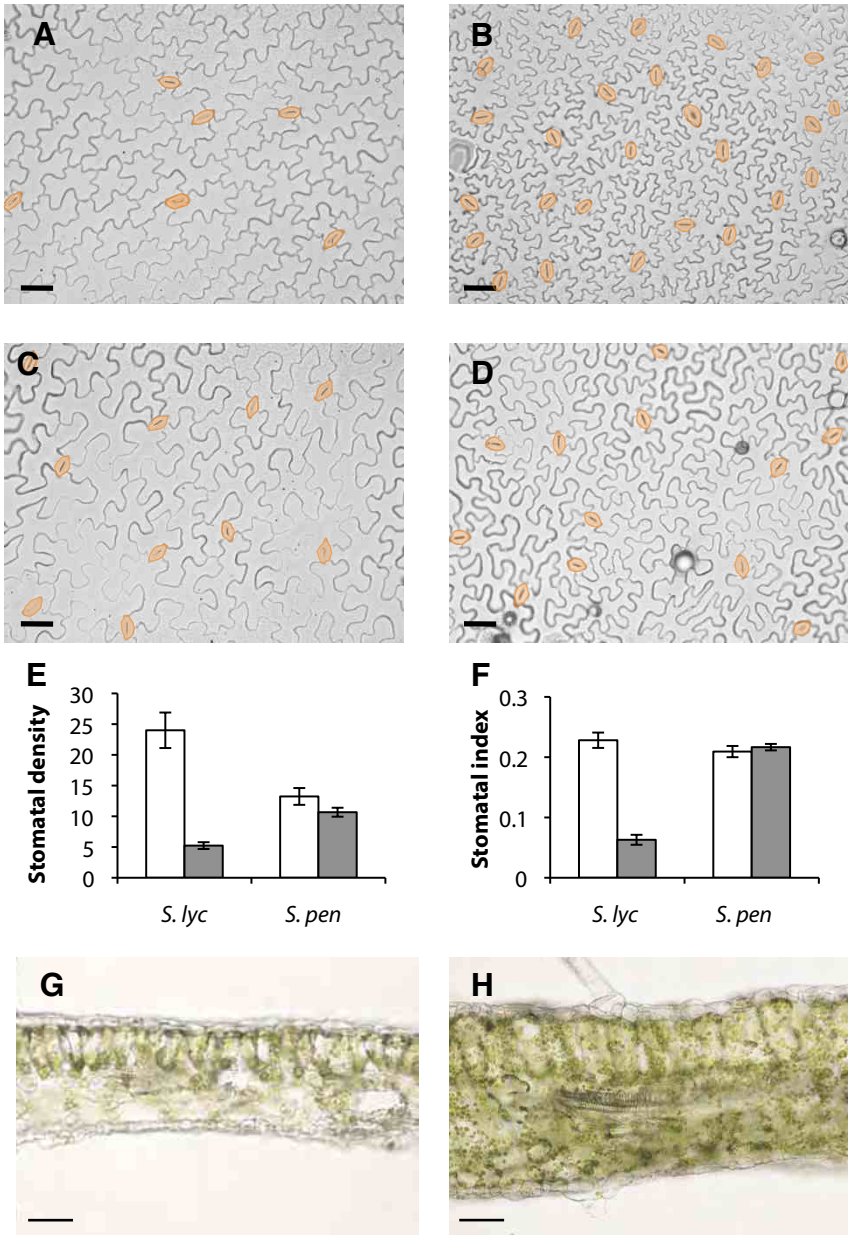

Fig. S15. Epidermal patterning and leaf thicknesses in *S. lycopersicum* var. M82 and *S. pennellii*. **A-D** Photographs of leaf epidermal peels of *S. lycopersicum* (**A, B**) and *S. pennellii* (**C, D**). Peels were taken for both the adaxial (**A** and **C**) and abaxial (**B** and **D**) surfaces. Stomata are highlighted in orange. e-f Epidermal layer traits are plotted as barcharts. **E**. Mean number of stomata per 0.28 mm^2. **F** Stomatal index. White bars = abaxial, grey bars = adaxial. Bars show standard error, n = 10. **G-H** Cross sections from *S. lycopersicum* (**G**) and *S. pennellii* leaves (**H**). Scale bar = 50µm.

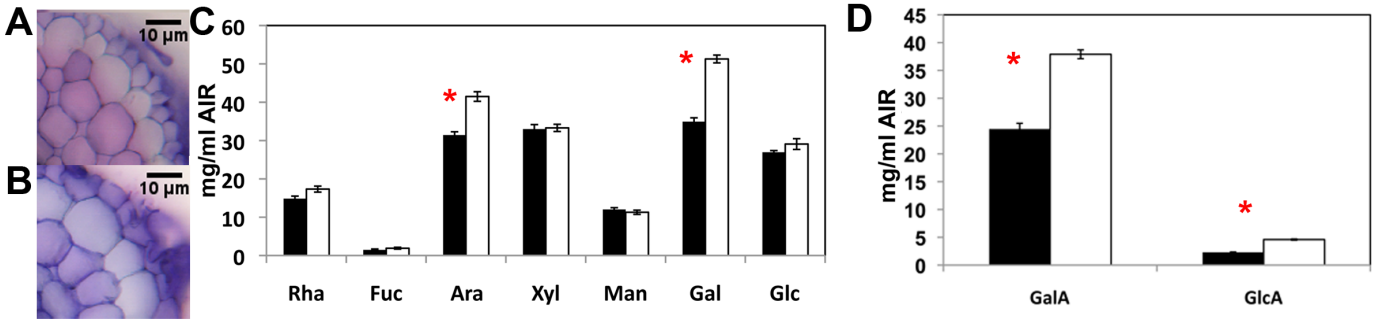

Fig. S16. Differences in root cell wall composition.
A-B. Toluidine Blue stained cross sections of *S. lycopersicum var.* M82 (**A**) and *S. pennellii* (**B**) roots show differences in staining, suggesting differences in cell wall structures between these two species.
C-D. Peptin structure is signficantly different between *S. lycopersicum var.* M82 (**A**) (black bars) and *S. pennellii* (white bars). These differences are signficant at $p < 0.01$ for Arabinose and Galactose (**C**) and for their uronic acids (**D**). **C**, and portions of **D**, are also presented in Fig 4; they are retained here for comparison with the full data set. Abbreviations: Rha-rhamnose, Fuc-fucose, Ara-arabinose, Xyl-xylose, Man-mannose, Gal-galactose, Glc-glucose, GalA-galactouronic acid, GlcA-glucouronic acid, AIR-alcohol insoluble residue.

Table S1. RNAseq libraries analyzed in this study.

Genotype	Tissue	Libraries	Reads
IL4-3	leaf	4	17,930,304
S.lyc-M82	floral	1	7,732,131
S.lyc-M82	floral	1	13,100,687
S.lyc-M82	fruit	1	4,715,863
S.lyc-M82	fruit	1	5,369,695
S.lyc-M82	leaf	1	5,919,262
S.lyc-M82	leaf	1	16,512,216
S.lyc-M82	root	1	12,608,946
S.lyc-M82	root	1	5,646,694
S.lyc-M82	seedling	1	5,688,116
S.lyc-M82	seedling	1	7,324,106
S.lyc-M82	stem	1	7,270,164
S.lyc-M82	stem	1	8,508,000
S.lyc-M82	vegetative meristem	1	7,302,549
S.lyc-M82	vegetative meristem	1	15,950,854
S. pen	floral	1	6,431,197
S. pen	floral	1	7,741,510
S. pen	fruit	1	5,169,418
S. pen	fruit	1	8,063,516
S. pen	leaf	1	14,590,777
S. pen	leaf	1	10,374,462
S. pen	root	1	7,154,686
S. pen	root	1	7,202,798
S. pen	seedling	1	8,131,555
S. pen	seedling	1	7,809,745
S. pen	stem	1	4,543,815
S. pen	stem	1	6,951,925
S. pen	vegetative meristem	1	10,675,266
S. pen	vegetative meristem	1	16,439,635
S. pim	leaf	1	69,790,669
S.lyc-M82	seedling	3	22,991,721
S.lyc-M82	seedling	3	20,939,901
S. pen	seedling	3	15,947,276
S. pen	seedling	3	21,171,540
S. pim	seedling	3	25,221,546
S. pim	seedling	3	15,974,475
S. hab	seedling	3	19,595,390
S. hab	seedling	3	24,604,834
S. gal	floral	1	42,444,250
S. chm	floral	1	37,070,820

Table S2. Number of reads sequenced and mapped.

	Filtered reads	Reads uniquely mapped	% Mapped
S. lyc	167,580,905	140,018,000	83.55
S. gal	42,444,250	30,104,377	70.93
S. pim	110,986,690	94,168,151	84.85
S. chm	37,070,820	25,830,627	69.68
S. hab	44,200,224	36,034,281	81.53
S. pen	158,399,121	126,218,010	79.68

Table S3. Number of base pairs covered by RNAseq reads.

	Total	CDS	Intron	Five prime UTR	Three prime UTR	Intergenic
Total bp in Heinz reference	781666411	35972459	67529494	1694148	3326212	673144098
S. lyc	55118989 (7.1%)	26934918 (74.9%)	13351827 (19.8%)	1277986 (75.4%)	2859796 (86%)	10694462 (1.6%)
S. gal	39290573 (5%)	22132932 (61.5%)	6620111 (9.8%)	899550 (53.1%)	2616879 (78.7%)	7021101 (1%)
S. pim	47208574 (6%)	25122993 (69.8%)	9946563 (14.7%)	1126426 (66.5%)	2697997 (81.1%)	8314595 (1.2%)
S. chm	32924140 (4.2%)	22651515 (63%)	3268624 (4.8%)	810507 (47.8%)	2084854 (62.7%)	4108640 (0.6%)
S. pen	49133872 (6.3%)	26212146 (72.9%)	9856323 (14.6%)	1118785 (66%)	2594706 (78%)	9351912 (1.4%)
S. hab	31874479 (4.1%)	21653109 (60.2%)	3284723 (4.9%)	789578 (46.6%)	2148825 (64.6%)	3998244 (0.6%)
Total covered by reads in all species	23897216 (3.1%)	18588006 (51.7%)	981343 (1.5%)	584423 (34.5%)	1783546 (53.6%)	1959898 (0.3%)
Total covered in any species	70693184 (9%)	28685468 (79.7%)	19877868 (29.4%)	1356475 (80.1%)	2998629 (90.2%)	17774744 (2.6%)

Table S4. Summary of *de novo*-assembled contigs.
(N50 is defined as the contig size above which 50% of the sequence information is contained)

	# of contigs	N50	# of contigs > N50	median length	mean length	max length	sum
S. lycopersicum cv. M82	37778	1814	7836	850	1175	15784	4.4E+07
S. pennellii	38039	1738	8196	879	1171	15778	4.5E+07

Table S5. Summary of novel *de novo* contig BLAST results against NCBI's non-redundant nt database and the *S. lycopersicum* var. Heinz reference transcriptome (ITAG2.3) and genome (CHR2.40). The results are for the best alignment of the best hit. Boxed and shaded regions indicate multiple contigs that correspond to single genes.

BLAST vs NCBI:

de novo contig ID	contig length	top non-redundant nt blast hit	max % identity	alignment length	e-value	bit score
SlycM82_Contig2987	567	dbj|AB061242.1| Solanum tuberosum StHRGP mRNA for hydroxyproline-rich glycoprotein, partial cds	91.6	322	3.0E-115	424
SlycM82_Contig26907	236	dbj|AK246855.1| Solanum lycopersicum cDNA, clone: FC26BF04, HTC in fruit	97.7	220	3.0E-107	396
SlycM82_Contig30271	604	dbj|AK319919.1| Solanum lycopersicum cDNA, clone: LEFL1003BD10, HTC in leaf	100.0	586	0.0	1162
SlycM82_k43_149561	663	dbj|AK319932.1| Solanum lycopersicum cDNA, clone: LEFL1003BH05, HTC in leaf	100.0	536	0.0	1063
SlycM82_k43_156096	793	dbj|AK319932.1| Solanum lycopersicum cDNA, clone: LEFL1003BH05, HTC in leaf	100.0	675	0.0	1338
SlycM82_Contig28637	849	dbj|AK319932.1| Solanum lycopersicum cDNA, clone: LEFL1003BH05, HTC in leaf	100.0	741	0.0	1469
SlycM82_Contig25020	2142	dbj|AK320195.1| Solanum lycopersicum cDNA, clone: LEFL1006BE10, HTC in leaf	100.0	2106	0.0	4175
SlycM82_Contig506	1239	dbj|AK321037.1| Solanum lycopersicum cDNA, clone: LEFL1017BB10, HTC in leaf	99.9	1239	0.0	2448
Spen_Contig2092	462	dbj|AK321128.1| Solanum lycopersicum cDNA, clone: LEFL1019DD06, HTC in leaf	82.6	432	5.0E-64	254
SlycM82_Contig5991	1244	dbj|AK321133.1| Solanum lycopersicum cDNA, clone: LEFL1019DH02, HTC in leaf	98.6	1150	0.0	2153
Spen_Contig24133	237	dbj|AK321258.1| Solanum lycopersicum cDNA, clone: LEFL1022BB07, HTC in leaf	92.6	54	1.0E-10	75.8
Spen_Contig4194	1414	dbj|AK322600.1| Solanum lycopersicum cDNA, clone: LEFL1039DB06, HTC in leaf	84.6	787	2.0E-161	579

SlycM82_Contig27936	956	dbj|AK324002.1| Solanum lycopersicum cDNA, clone: LEFL1069DB07, HTC in leaf	100.0	956	0.0	1895
Spen_Contig24460	829	dbj|AK324259.1| Solanum lycopersicum cDNA, clone: LEFL1074DD07, HTC in leaf	98.7	829	0.0	1556
SlycM82_Contig5233	919	dbj|AK324259.1| Solanum lycopersicum cDNA, clone: LEFL1074DD07, HTC in leaf	99.7	908	0.0	1776
Spen_Contig10948	1003	dbj|AK324357.1| Solanum lycopersicum cDNA, clone: LEFL1076DE03, HTC in leaf	98.2	1003	0.0	1842
SlycM82_Contig8950	975	dbj|AK324357.1| Solanum lycopersicum cDNA, clone: LEFL1076DE03, HTC in leaf	100.0	975	0.0	1933
Spen_Contig5599	1467	dbj|AK326255.1| Solanum lycopersicum cDNA, clone: LEFL2003CD07, HTC in fruit	82.6	149	5.0E-14	89.7
SlycM82_Contig28480	837	dbj|AK328000.1| Solanum lycopersicum cDNA, clone: LEFL2043M09, HTC in fruit	100.0	779	0.0	1544
Spen_k23_122075	392	dbj|AK329702.1| Solanum lycopersicum cDNA, clone: LEFL3154A01, HTC in root	97.1	34	1.0E-05	60
Spen_Contig19819	1120	dbj|AK329780.1| Solanum lycopersicum cDNA, clone: LEFL3156J08, HTC in root	83.3	467	9.0E-80	307
Spen_Contig16750	1772	dbj|AK329780.1| Solanum lycopersicum cDNA, clone: LEFL3156J08, HTC in root	83.1	456	2.0E-75	293
Spen_k43_31086	208	dbj|AP009550.1| Solanum lycopersicum DNA, chromosome 8, clone: C08SLe0129C18, complete	83.7	92	4.0E-07	63.9
Spen_Contig12844	495	emb|AJ009719.1| Solanum tuberosum mRNA for NL25 protein	80.6	495	3.0E-50	208
SlycM82_Contig11465	1603	emb|AM431016.2| Vitis vinifera contig VV78X074727.7, whole genome shotgun sequence	78.7	1452	2.0E-106	396
Spen_Contig733	1611	emb|AM431016.2| Vitis vinifera contig VV78X074727.7, whole genome shotgun sequence	78.7	1490	5.0E-110	408
Spen_k33_222701	1267	gb|AC212314.2| Solanum lycopersicum chromosome 11 clone C11HBa0027B05, complete sequence	81.8	835	4.0E-116	428
Spen_Contig4405	1283	gb|AC212314.2| Solanum lycopersicum chromosome 11 clone C11HBa0027B05, complete sequence	81.8	835	4.0E-116	428
SlycM82_k43_74240	212	gb|AC215434.2| Solanum lycopersicum chromosome 2 clone C02HBa0284G15, complete sequence	100.0	212	2.0E-114	420

SlycM82_k33_143641	212	gb|AC215434.2| Solanum lycopersicum chromosome 2 clone C02HBa0284G15, complete sequence	100.0	212	2.0E-114	420
SlycM82_Contig15299	583	gb|AF243180.1|AF243180 Lycopersicon esculentum dicyanin mRNA, complete cds	100.0	583	0.0	1156
Spen_Contig10606	484	gb|AY303171.1| Solanum bulbocastanum chromosome 8 clone UW177O13, complete sequence	85.3	402	6.0E-85	323
SlycM82_Contig12244	797	gb|BT013250.1| Lycopersicon esculentum clone 134756F, mRNA sequence	100.0	657	0.0E+00	1302
Spen_k23_428173	202	gb|EF514213.1| Solanum tuberosum strain P6/210 contig r1, complete sequence	88.1	168	6.0E-42	179
Spen_k23_389944	236	gb|GU563972.1| Solanum hjertingii R2 late blight resistance protein (Rpi-hjt1.2) gene, complete cds	86.9	206	5.0E-50	206
SlycM82_Contig30152	206	gb|M76670.1|TOMEXTENA L.esculentum extensin (class I) gene, complete cds	97.0	203	7.0E-95	355
SlycM82_Contig30037	232	gb|M76670.1|TOMEXTENA L.esculentum extensin (class I) gene, complete cds	96.5	228	6.0E-105	389
Spen_Contig14951	1978	ref|XM_002271419.1| PREDICTED: Vitis vinifera hypothetical protein LOC100249908, mRNA	80.7	455	2.0E-48	204
Spen_Contig2014	506	ref|XM_002274500.1| PREDICTED: Vitis vinifera hypothetical protein LOC100242630, mRNA	88.4	95	5.0E-18	101
SlycM82_Contig14188	243	ref|XM_002279852.1| PREDICTED: Vitis vinifera hypothetical protein LOC100254571, mRNA	85.8	113	3.0E-17	97.6
Spen_Contig24714	315	ref|XM_002299079.1| Populus trichocarpa predicted protein, mRNA	87.2	94	3.0E-15	91.7
Spen_Contig1146	1268	ref|XM_002328099.1| Populus trichocarpa predicted protein, mRNA	83.7	147	1.0E-17	101
Spen_k33_215008	428	ref|XM_002533498.1| Ricinus communis cytochrome P450, putative, mRNA	89.3	56	9.0E-07	63.9
Spen_Contig7313	428	ref|XM_002533498.1| Ricinus communis cytochrome P450, putative, mRNA	89.3	56	9.0E-07	63.9

BLAST vs REFERENCE TRANSCRIPTOME:

de novo contig ID	contig length	top ITAG2.3 hit	max % identity	alignment length	e-value	bit score
SlycM82_Contig2987	567	Solyc04g071070.2.1	85.2	81	3.0E-10	65.9
SlycM82_Contig26907	236	Solyc10g081980.1.1	91.9	221	8.0E-80	295
SlycM82_Contig30271	604	Solyc06g005060.2.1	90.8	586	0.0	733
SlycM82_k43_149561	663	Solyc10g009570.2.1	100.0	19	8.7E-02	38.2
SlycM82_k43_156096	793	Solyc06g067930.1.1	96.2	26	2.0E-03	44.1
SlycM82_Contig28637	849	Solyc06g067930.1.1	96.2	26	2.0E-03	44.1
SlycM82_Contig25020	2142	Solyc05g006720.1.1	92.7	165	9.0E-60	232
SlycM82_Contig506	1239	Solyc12g044950.1.1	90.0	421	4.0E-141	502
Spen_Contig2092	462	Solyc12g044940.1.1	82.6	432	5.0E-67	254
SlycM82_Contig5991	1244	Solyc12g100230.1.1	79.7	1035	2.0E-102	373
Spen_Contig24133	237	Solyc10g083400.1.1	83.2	238	7.0E-31	133
Spen_Contig4194	1414	Solyc06g060690.2.1	84.5	670	7.0E-140	498
SlycM82_Contig27936	956	Solyc11g018800.1.1	91.6	931	0.0	1227
Spen_Contig24460	829	Solyc06g035520.2.1	80.8	156	2.0E-09	63.9
SlycM82_Contig5233	919	Solyc01g098140.2.1	96.7	30	8.0E-06	52
Spen_Contig10948	1003	Solyc01g005130.2.1	81.3	336	1.0E-38	161
SlycM82_Contig8950	975	Solyc01g005130.2.1	81.3	336	1.0E-38	161
Spen_Contig5599	1467	Solyc07g009440.1.1	85.0	593	2.0E-131	470
SlycM82_Contig28480	837	Solyc05g006730.2.1	90.8	664	0.0	833
Spen_k23_122075	392	Solyc01g091660.2.1	97.1	34	1.0E-08	60
Spen_Contig19819	1120	Solyc06g060190.2.1	82.6	798	7.0E-130	464
Spen_Contig16750	1772	Solyc06g060190.2.1	82.0	1364	0.0	737
Spen_k43_31086	208	Solyc04g082920.2.1	90.2	41	7.0E-06	50.1
Spen_Contig12844	495	Solyc11g011090.1.1	84.1	195	6.0E-33	141
SlycM82_Contig11465	1603	Solyc06g065340.1.1	100.0	20	5.4E-02	40.1
Spen_Contig733	1611	Solyc06g065340.1.1	100.0	20	5.4E-02	40.1
Spen_k33_222701	1267	Solyc11g011090.1.1	81.8	835	5.0E-119	428
Spen_Contig4405	1283	Solyc11g011090.1.1	81.8	835	5.0E-119	428
SlycM82_k43_74240	212	Solyc02g091990.2.1	85.2	155	1.0E-28	125
SlycM82_k33_143641	212	Solyc02g091990.2.1	85.2	155	1.0E-28	125

SlycM82_Contig15299	583	Solyc12g009450.1.1	100.0	18	3.0E-01	36.2
Spen_Contig10606	484	Solyc08g075630.2.1	84.2	450	3.0E-87	321
SlycM82_Contig12244	797	Solyc03g020080.2.1	86.7	360	1.0E-90	333
Spen_k23_428173	202	Solyc09g012040.1.1	81.4	145	1.0E-16	86
Spen_k23_389944	236	Solyc04g009290.1.1	83.7	245	3.0E-45	180
SlycM82_Contig30152	206	Solyc12g038700.1.1	86.4	59	4.0E-07	54
SlycM82_Contig30037	232	Solyc12g098760.1.1	96.4	28	3.0E-05	48.1
Spen_Contig14951	1978	Solyc06g071280.2.1	100.0	20	6.7E-02	40.1
Spen_Contig2014	506	Solyc09g055660.1.1	95.5	22	2.6E-01	36.2
SlycM82_Contig14188	243	Solyc09g082870.1.1	86.0	50	5.0E-04	44.1
Spen_Contig24714	315	Solyc11g072110.1.1	94.6	37	4.0E-08	58
Spen_Contig1146	1268	Solyc04g008760.1.1	78.4	222	5.0E-08	60
Spen_k33_215008	428	Solyc10g083400.1.1	83.5	248	2.0E-47	188
Spen_Contig7313	428	Solyc10g083400.1.1	81.5	248	2.0E-38	159

BLAST vs REFERENCE GENOME:

de novo contig ID	contig length	top CHR2.40 hit	max % identity	alignment length	e-value	bit score
SlycM82_Contig2987	567	SL2.40ch04: 55610112-55610425	87.4	326	2.0E-89	333
SlycM82_Contig26907	236	SL2.40ch10: 62223520-62223300	91.9	221	1.0E-78	295
SlycM82_Contig30271	604	SL2.40ch06: 37558-36973	90.8	586	0.0	733
SlycM82_k43_149561	663	SL2.40ch02: 21802903-21802928	100.0	26	1.0E-04	52
SlycM82_k43_156096	793	SL2.40ch06: 38496082-38496057	96.2	26	3.2E-02	44.1
SlycM82_Contig28637	849	SL2.40ch06: 38496057-38496082	96.2	26	3.4E-02	44.1
SlycM82_Contig25020	2142	SL2.40ch05: 1374480-1374895	92.3	416	2.0E-160	571
SlycM82_Contig506	1239	SL2.40ch12: 45745416-45744731	88.8	686	0.0	749
Spen_Contig2092	462	SL2.40ch12: 45736339-45736621	83.8	284	1.0E-46	190
SlycM82_Contig5991	1244	SL2.40ch12: 45744733-45744274	92.2	460	2.0E-177	626
Spen_Contig24133	237	SL2.40ch10: 62546429-62546197	83.2	238	1.0E-29	133
Spen_Contig4194	1414	SL2.40ch06: 35101940-35102706	84.6	787	4.0E-163	579
SlycM82_Contig27936	956	SL2.40ch11: 9597154-9597549	92.7	396	4.0E-156	555
Spen_Contig24460	829	SL2.40ch06: 21129561-21129407	80.8	156	4.0E-08	63.9
SlycM82_Contig5233	919	SL2.40ch01: 80485142-80485171	96.7	30	2.0E-04	52

Spen_Contig10948	1003	SL2.40ch01: 125049-125384	81.3	336	3.0E-37	161
SlycM82_Contig8950	975	SL2.40ch01: 125049-125384	81.3	336	2.0E-37	161
Spen_Contig5599	1467	SL2.40ch07: 4495698-4496287	85.0	593	3.0E-130	470
SlycM82_Contig28480	837	SL2.40ch05: 1377106-1377418	94.3	313	7.0E-133	478
Spen_k23_122075	392	SL2.40ch01: 77017199-77017232	97.1	34	3.0E-07	60
Spen_Contig19819	1120	SL2.40ch06: 34326346-34327168	83.4	823	5.0E-153	545
Spen_Contig16750	1772	SL2.40ch06: 34326367-34327168	83.5	802	3.0E-152	543
Spen_k43_31086	208	SL2.40ch08: 51113020-51113111	83.7	92	8.0E-09	63.9
Spen_Contig12844	495	SL2.40ch11: 4175593-4175752	84.4	160	4.0E-25	119
SlycM82_Contig11465	1603	SL2.40ch10: 16312464-16312440	100.0	25	1.0E-03	50.1
Spen_Contig733	1611	SL2.40ch10: 16312464-16312440	100.0	25	1.0E-03	50.1
Spen_k33_222701	1267	SL2.40ch11: 4177881-4178704	81.8	835	9.0E-118	428
Spen_Contig4405	1283	SL2.40ch11: 4178704-4177881	81.8	835	9.0E-118	428
SlycM82_k43_74240	212	SL2.40ch02: 47786882-47787036	85.2	155	3.0E-27	125
SlycM82_k33_143641	212	SL2.40ch02: 47787036-47786882	85.2	155	3.0E-27	125
SlycM82_Contig15299	583	SL2.40ch07: 2830556-2830578	100.0	23	6.0E-03	46.1
Spen_Contig10606	484	SL2.40ch08: 56934940-56935385	84.2	450	5.0E-86	321
SlycM82_Contig12244	797	SL2.40ch03: 6933763-6934152	86.2	390	6.0E-93	345
Spen_k23_428173	202	SL2.40ch09: 5325626-5325491	81.4	145	2.0E-15	86
Spen_k23_389944	236	SL2.40ch04: 2738261-2738017	83.7	245	6.0E-44	180
SlycM82_Contig30152	206	SL2.40ch07: 13423062-13422986	92.2	77	2.0E-21	105
SlycM82_Contig30037	232	SL2.40ch07: 13423062-13422983	88.8	80	7.0E-16	87.7
Spen_Contig14951	1978	SL2.40ch04: 42760669-42760689	100.0	21	3.2E-01	42.1
Spen_Contig2014	506	SL2.40ch07: 25663290-25663272	100.0	19	1.2E+00	38.2
SlycM82_Contig14188	243	SL2.40ch09: 63947193-63947144	86.0	50	9.0E-03	44.1
Spen_Contig24714	315	SL2.40ch11: 52455637-52455673	94.6	37	8.0E-07	58
Spen_Contig1146	1268	SL2.40ch04: 2413999-2413778	78.4	222	9.0E-07	60
Spen_k33_215008	428	SL2.40ch10: 62546629-62546388	83.5	248	5.0E-46	188
Spen_Contig7313	428	SL2.40ch10: 62546629-62546388	81.5	248	4.0E-37	159

Table S6. Enriched GO categories in differentially expressed genes identified in the 4 species whole-model analysis.

A. Over-represented GO slim categrories among genes differentially expressed across 4 tomato species

Category	Description	Adjusted P-value
GO:0008152	metabolic process	1.64E-28
GO:0009987	cellular process	4.40E-17
GO:0009058	biosynthetic process	2.28E-15
GO:0006950	response to stress	4.01E-14
GO:0009056	catabolic process	4.01E-14
GO:0006629	lipid metabolic process	1.16E-11
GO:0005975	carbohydrate metabolic process	9.08E-11
GO:0008150	biological_process	2.17E-10
GO:0009628	response to abiotic stimulus	8.30E-10
GO:0019748	secondary metabolic process	9.65E-10
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.48E-07
GO:0006412	translation	1.47E-05
GO:0009835	ripening	2.24E-04
GO:0009607	response to biotic stimulus	4.81E-04
GO:0006810	transport	1.02E-03
GO:0009991	response to extracellular stimulus	1.60E-03
GO:0008219	cell death	2.53E-03
GO:0009875	pollen-pistil interaction	2.53E-03
GO:0019538	protein metabolic process	6.32E-03
GO:0006464	protein modification process	6.47E-03
GO:0007165	signal transduction	7.20E-03
GO:0009605	response to external stimulus	7.20E-03
GO:0009719	response to endogenous stimulus	2.68E-02
GO:0007154	cell communication	2.68E-02
GO:0009790	embryo development	4.23E-02
GO:0015979	photosynthesis	4.77E-02

B. Over-represented GO categrories (full set) among genes differentially expressed across 4 tomato species

Category	Description	Adjusted P-value
GO:0008152	metabolic process	2.74E-25
GO:0055114	oxidation-reduction process	3.37E-17
GO:0046686	response to cadmium ion	3.82E-09
GO:0006952	defense response	7.18E-07
GO:0009651	response to salt stress	3.38E-06
GO:0006633	fatty acid biosynthetic process	3.38E-06
GO:0006508	proteolysis	3.38E-06
GO:0006468	protein phosphorylation	4.04E-06
GO:0055085	transmembrane transport	4.41E-06
GO:0009069	serine family amino acid metabolic process	7.48E-06

GO:0006857	oligopeptide transport	8.04E-06
GO:0005985	sucrose metabolic process	2.21E-04
GO:0006094	gluconeogenesis	4.35E-04

B
(continued).

Category	Description	Adjusted P-value
GO:0070588	calcium ion transmembrane transport	1.48E-03
GO:0009416	response to light stimulus	1.51E-03
GO:0005982	starch metabolic process	1.59E-03
GO:0006855	drug transmembrane transport	1.75E-03
GO:0006568	tryptophan metabolic process	1.92E-03
GO:0006550	isoleucine catabolic process	5.57E-03
GO:0006574	valine catabolic process	5.57E-03
GO:0006096	glycolysis	5.72E-03
GO:0006200	ATP catabolic process	7.38E-03
GO:0006011	UDP-glucose metabolic process	7.81E-03
GO:0018874	benzoate metabolic process	8.03E-03
GO:0009624	response to nematode	9.60E-03
GO:0006522	alanine metabolic process	9.91E-03
GO:0009835	ripening	1.02E-02
GO:0042254	ribosome biogenesis	1.10E-02
GO:0006552	leucine catabolic process	1.23E-02
GO:0019482	beta-alanine metabolic process	1.53E-02
GO:0015846	polyamine transport	1.78E-02
GO:0006531	aspartate metabolic process	2.18E-02
GO:0043086	negative regulation of catalytic activity	2.22E-02
GO:0006629	lipid metabolic process	2.58E-02
GO:0046251	limonene catabolic process	2.58E-02
GO:0006547	histidine metabolic process	2.64E-02
GO:0006554	lysine catabolic process	2.79E-02
GO:0048316	seed development	2.91E-02
GO:0045087	innate immune response	2.91E-02
GO:0042967	acyl-carrier-protein biosynthetic process	2.96E-02
GO:0006412	translation	3.05E-02
GO:0006865	amino acid transport	3.05E-02
GO:0006915	apoptosis	3.05E-02
GO:0009813	flavonoid biosynthetic process	3.05E-02
GO:0015995	chlorophyll biosynthetic process	3.10E-02
GO:0009658	chloroplast organization	3.22E-02
GO:0010025	wax biosynthetic process	3.22E-02
GO:0030001	metal ion transport	3.22E-02
GO:0051258	protein polymerization	3.33E-02
GO:0044237	cellular metabolic process	3.47E-02
GO:0015706	nitrate transport	3.54E-02
GO:0009098	leucine biosynthetic process	3.72E-02
GO:0009099	valine biosynthetic process	3.72E-02
GO:0009644	response to high light intensity	3.75E-02
GO:0006560	proline metabolic process	4.28E-02

GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	4.53E-02
GO:0009607	response to biotic stimulus	4.62E-02
GO:0005975	carbohydrate metabolic process	4.89E-02

Table S7. Over-represented GO slim categories in DE genes from pairwise species comparisons.

A. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. pennellii*

Category	Description	Adjusted P-value
GO:0006412	translation	3.63E-51
GO:0009987	cellular process	9.68E-27
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.66E-22
GO:0009790	embryo development	7.77E-14
GO:0009058	biosynthetic process	3.43E-12
GO:0008152	metabolic process	8.76E-11
GO:0000003	reproduction	2.87E-06
GO:0006950	response to stress	4.55E-06
GO:0009791	post-embryonic development	2.15E-05
GO:0019538	protein metabolic process	2.13E-04
GO:0008150	biological_process	2.13E-04
GO:0006091	generation of precursor metabolites and energy	8.20E-04
GO:0016043	cellular component organization	8.20E-04
GO:0009056	catabolic process	1.03E-03
GO:0009628	response to abiotic stimulus	1.28E-03
GO:0005975	carbohydrate metabolic process	5.24E-03
GO:0009607	response to biotic stimulus	5.77E-03
GO:0019748	secondary metabolic process	5.77E-03
GO:0019725	cellular homeostasis	3.53E-02
GO:0006810	transport	3.56E-02
GO:0015979	photosynthesis	3.77E-02
GO:0006259	DNA metabolic process	4.85E-02

B. Over-represented GO slim categrories among genes expressed lower in *S. pimpinellifolium* than in *S. pennellii*

Category	Description	Adjusted P-value
GO:0008152	metabolic process	7.04E-07
GO:0006629	lipid metabolic process	9.12E-06
GO:0009056	catabolic process	3.44E-04
GO:0015979	photosynthesis	5.00E-03
GO:0009987	cellular process	5.56E-03
GO:0005975	carbohydrate metabolic process	2.42E-02
GO:0019748	secondary metabolic process	2.47E-02

C. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. pimpinellifolium*

Category	Description	Adjusted P-value
GO:0005975	carbohydrate metabolic process	5.93E-08
GO:0008152	metabolic process	9.00E-07
GO:0009835	ripening	6.86E-05
GO:0019748	secondary metabolic process	6.86E-05
GO:0006950	response to stress	5.90E-04
GO:0009056	catabolic process	2.07E-03
GO:0009719	response to endogenous stimulus	6.34E-03
GO:0006629	lipid metabolic process	6.34E-03
GO:0009628	response to abiotic stimulus	6.34E-03
GO:0009605	response to external stimulus	1.13E-02
GO:0007267	cell-cell signaling	4.71E-02

D. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. pimpinellifolium*

Category	Description	Adjusted P-value
GO:0006412	translation	1.01E-30
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.73E-09
GO:0006259	DNA metabolic process	1.18E-06
GO:0009790	embryo development	3.15E-06
GO:0009987	cellular process	1.43E-04
GO:0009058	biosynthetic process	1.60E-04
GO:0000003	reproduction	1.34E-03
GO:0008152	metabolic process	2.43E-03
GO:0009791	post-embryonic development	3.08E-03
GO:0006950	response to stress	8.57E-03
GO:0019538	protein metabolic process	9.04E-03

E. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. pennellii*

Category	Description	Adjusted P-value
GO:0008152	metabolic process	1.45E-13
GO:0005975	carbohydrate metabolic process	1.20E-12
GO:0009987	cellular process	1.76E-12
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.54E-09
GO:0009058	biosynthetic process	3.54E-07
GO:0006412	translation	5.48E-07
GO:0009056	catabolic process	8.22E-07
GO:0019748	secondary metabolic process	4.48E-05
GO:0006950	response to stress	5.20E-05
GO:0009628	response to abiotic stimulus	2.06E-04
GO:0006091	generation of precursor metabolites and energy	2.42E-04
GO:0009790	embryo development	1.31E-03

GO:0006810	transport	3.24E-03
GO:0008150	biological_process	3.44E-03
GO:0009607	response to biotic stimulus	2.00E-02
GO:0000003	reproduction	3.36E-02

F. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	1.62E-07
GO:0006629	lipid metabolic process	1.77E-07
GO:0009875	pollen-pistil interaction	1.91E-02
GO:0019748	secondary metabolic process	2.56E-02
GO:0006950	response to stress	3.03E-02
GO:0009056	catabolic process	4.11E-02

G. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	6.70E-10
GO:0005975	carbohydrate metabolic process	1.39E-05
GO:0019748	secondary metabolic process	2.36E-05
GO:0006629	lipid metabolic process	8.26E-04
GO:0006810	transport	3.77E-03
GO:0009835	ripening	4.02E-03
GO:0009056	catabolic process	1.05E-02
GO:0008219	cell death	1.07E-02
GO:0009628	response to abiotic stimulus	1.08E-02
GO:0006950	response to stress	1.08E-02
GO:0009058	biosynthetic process	1.88E-02
GO:0009606	tropism	2.08E-02
GO:0030154	cell differentiation	2.21E-02
GO:0009991	response to extracellular stimulus	2.21E-02

H. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0006412	translation	9.92E-15
GO:0009987	cellular process	3.11E-10
GO:0008152	metabolic process	7.09E-07
GO:0015979	photosynthesis	1.70E-05
GO:0006259	DNA metabolic process	2.80E-04
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3.30E-04
GO:0009058	biosynthetic process	5.32E-04
GO:0009056	catabolic process	1.43E-02
GO:0006629	lipid metabolic process	3.53E-02
GO:0019538	protein metabolic process	4.48E-02

I. Over-represented GO slim categrories among genes expressed higher in *S. habrochaites* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0006412	translation	2.85E-55
GO:0009987	cellular process	9.98E-31
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.58E-22
GO:0009790	embryo development	1.34E-12
GO:0008152	metabolic process	1.81E-10
GO:0006091	generation of precursor metabolites and energy	1.80E-09
GO:0009058	biosynthetic process	7.37E-08
GO:0008150	biological_process	1.13E-07
GO:0000003	reproduction	1.25E-06
GO:0009056	catabolic process	1.60E-06
GO:0016043	cellular component organization	3.90E-06
GO:0006950	response to stress	4.80E-06
GO:0009791	post-embryonic development	1.22E-05
GO:0019538	protein metabolic process	2.88E-05
GO:0015979	photosynthesis	6.42E-05
GO:0009628	response to abiotic stimulus	1.92E-04
GO:0005975	carbohydrate metabolic process	2.09E-04
GO:0006259	DNA metabolic process	9.26E-04
GO:0009607	response to biotic stimulus	2.88E-03
GO:0006810	transport	4.99E-02

J. Over-represented GO slim categrories among genes expressed lower in *S. habrochaites* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0006629	lipid metabolic process	2.19E-04
GO:0006950	response to stress	2.19E-04
GO:0008152	metabolic process	2.83E-04
GO:0009628	response to abiotic stimulus	1.84E-02
GO:0008219	cell death	4.30E-02
GO:0008150	biological_process	4.42E-02
GO:0009991	response to extracellular stimulus	4.67E-02
GO:0009605	response to external stimulus	4.67E-02
GO:0009875	pollen-pistil interaction	4.67E-02

K. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	2.10E-09
GO:0009058	biosynthetic process	2.32E-06
GO:0019748	secondary metabolic process	9.43E-05
GO:0006950	response to stress	4.17E-03
GO:0005975	carbohydrate metabolic process	5.72E-03

GO:0008219	cell death	7.93E-03
GO:0006810	transport	1.07E-02
GO:0009987	cellular process	1.22E-02
GO:0006629	lipid metabolic process	1.85E-02
GO:0009607	response to biotic stimulus	2.39E-02
GO:0009628	response to abiotic stimulus	3.57E-02
GO:0009056	catabolic process	3.57E-02
GO:0009856	pollination	3.57E-02

L. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	7.04E-07
GO:0006629	lipid metabolic process	9.12E-06
GO:0009056	catabolic process	3.44E-04
GO:0015979	photosynthesis	5.00E-03
GO:0009987	cellular process	5.56E-03
GO:0005975	carbohydrate metabolic process	2.42E-02
GO:0019748	secondary metabolic process	2.47E-02

Table S8. Over-represented GO categories in DE genes from pairwise species comparisons.

A. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. pennellii*

Category	Description	Adjusted P-value
GO:0042254	ribosome biogenesis	3.95E-41
GO:0006412	translation	3.38E-40
GO:0046686	response to cadmium ion	7.47E-15
GO:0009793	embryo development ending in seed dormancy	6.16E-14
GO:0009658	chloroplast organization	1.29E-07
GO:0008152	metabolic process	6.36E-06
GO:0044267	cellular protein metabolic process	1.63E-05
GO:0006522	alanine metabolic process	3.47E-05
GO:0006531	aspartate metabolic process	1.12E-04
GO:0006396	RNA processing	1.12E-04
GO:0006364	rRNA processing	1.12E-04
GO:0006448	regulation of translational elongation	1.87E-04
GO:0009955	adaxial/abaxial pattern formation	5.02E-04
GO:0055114	oxidation-reduction process	5.02E-04
GO:0015976	carbon utilization	6.19E-04
GO:0006144	purine base metabolic process	2.02E-03
GO:0006457	protein folding	2.23E-03
GO:0042026	protein refolding	2.46E-03
GO:0006633	fatty acid biosynthetic process	2.73E-03
GO:0006206	pyrimidine base metabolic process	4.76E-03
GO:0009408	response to heat	5.47E-03
GO:0006184	GTP catabolic process	1.04E-02
GO:0006094	gluconeogenesis	1.04E-02
GO:0009651	response to salt stress	1.04E-02
GO:0009097	isoleucine biosynthetic process	1.12E-02
GO:0051252	regulation of RNA metabolic process	1.16E-02
GO:0042128	nitrate assimilation	1.16E-02
GO:0009098	leucine biosynthetic process	1.16E-02
GO:0009099	valine biosynthetic process	1.16E-02
GO:0006260	DNA replication	1.29E-02
GO:0006544	glycine metabolic process	1.29E-02
GO:0006422	aspartyl-tRNA aminoacylation	1.76E-02
GO:0008283	cell proliferation	1.77E-02
GO:0006268	DNA unwinding involved in replication	2.11E-02
GO:0022900	electron transport chain	2.80E-02
GO:0009086	methionine biosynthetic process	2.99E-02
GO:0006096	glycolysis	3.52E-02
GO:0009607	response to biotic stimulus	4.70E-02

B. Over-represented GO slim categrories among genes expressed lower in *S. pimpinellifolium* than in *S. pennellii*

Category	Description	Adjusted P-value
GO:0006468	protein phosphorylation	1.46E-05

GO:0009069	serine family amino acid metabolic process	4.42E-04
GO:0006629	lipid metabolic process	5.95E-03
GO:0048544	recognition of pollen	1.14E-02
GO:0055114	oxidation-reduction process	1.17E-02

C. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. pimpinellifolium*

Category	**Description**	**Adjusted P-value**
GO:0055114	oxidation-reduction process	6.55E-06
GO:0006857	oligopeptide transport	1.54E-04
GO:0005985	sucrose metabolic process	1.30E-03
GO:0008152	metabolic process	4.02E-03
GO:0005982	starch metabolic process	5.58E-03
GO:0006855	drug transmembrane transport	7.75E-03
GO:0009835	ripening	1.12E-02
GO:0006570	tyrosine metabolic process	1.12E-02
GO:0006547	histidine metabolic process	1.40E-02
GO:0019752	carboxylic acid metabolic process	2.69E-02
GO:0005975	carbohydrate metabolic process	3.44E-02
GO:0006629	lipid metabolic process	3.44E-02
GO:0019482	beta-alanine metabolic process	3.44E-02
GO:0051555	flavonol biosynthetic process	4.40E-02

D. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. pimpinellifolium*

Category	**Description**	**Adjusted P-value**
GO:0006412	translation	3.57E-21
GO:0042254	ribosome biogenesis	4.27E-21
GO:0009793	embryo development ending in seed dormancy	5.92E-06
GO:0009658	chloroplast organization	1.55E-04
GO:0006144	purine base metabolic process	1.44E-03
GO:0046686	response to cadmium ion	4.13E-03
GO:0006206	pyrimidine base metabolic process	6.16E-03
GO:0055114	oxidation-reduction process	1.14E-02
GO:0045037	protein import into chloroplast stroma	1.30E-02
GO:0006531	aspartate metabolic process	4.15E-02
GO:0006268	DNA unwinding involved in replication	4.57E-02

E. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0046686	response to cadmium ion	2.08E-09
GO:0055114	oxidation-reduction process	1.37E-08
GO:0008152	metabolic process	4.99E-08
GO:0005985	sucrose metabolic process	1.81E-04
GO:0015976	carbon utilization	5.91E-04
GO:0005982	starch metabolic process	7.21E-04
GO:0006448	regulation of translational elongation	1.04E-03
GO:0006857	oligopeptide transport	1.14E-03

GO:0009651	response to salt stress	1.85E-03
GO:0009658	chloroplast organization	3.79E-03
GO:0019852	L-ascorbic acid metabolic process	3.79E-03
GO:0006364	rRNA processing	3.79E-03
GO:0006522	alanine metabolic process	4.70E-03
GO:0044267	cellular protein metabolic process	4.70E-03
GO:0044237	cellular metabolic process	6.66E-03
GO:0022900	electron transport chain	7.91E-03
GO:0009793	embryo development ending in seed dormancy	9.11E-03
GO:0006531	aspartate metabolic process	9.67E-03
GO:0009607	response to biotic stimulus	1.11E-02
GO:0006094	gluconeogenesis	2.06E-02
GO:0009399	nitrogen fixation	3.79E-02
GO:0005975	carbohydrate metabolic process	4.11E-02
GO:0046274	lignin catabolic process	4.31E-02

F. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0006468	protein phosphorylation	1.21E-04
GO:0008152	metabolic process	1.21E-04
GO:0055114	oxidation-reduction process	1.21E-04
GO:0006952	defense response	3.29E-03
GO:0009069	serine family amino acid metabolic process	3.29E-03
GO:0045087	innate immune response	1.07E-02

G. Over-represented GO slim categrories among genes expressed higher in *S. lycopersicum* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	2.38E-08
GO:0055114	oxidation-reduction process	1.67E-04
GO:0006952	defense response	3.39E-04
GO:0009813	flavonoid biosynthetic process	3.29E-03
GO:0005985	sucrose metabolic process	4.59E-03
GO:0009416	response to light stimulus	4.88E-03
GO:0005982	starch metabolic process	4.07E-02
GO:0006855	drug transmembrane transport	4.50E-02
GO:0009607	response to biotic stimulus	5.00E-02
GO:0048765	root hair cell differentiation	5.00E-02
GO:0005975	carbohydrate metabolic process	5.00E-02

H. Over-represented GO slim categrories among genes expressed lower in *S. lycopersicum* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0042254	ribosome biogenesis	3.54E-13
GO:0006412	translation	8.39E-13
GO:0006334	nucleosome assembly	5.21E-03
GO:0008152	metabolic process	6.82E-03
GO:0055114	oxidation-reduction process	2.35E-02

GO:0006206	pyrimidine base metabolic process	3.05E-02
GO:0006270	DNA-dependent DNA replication initiation	3.40E-02
GO:0015979	photosynthesis	4.33E-02

I. Over-represented GO slim categrories among genes expressed higher in *S. habrochaites* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0042254	ribosome biogenesis	5.10E-42
GO:0006412	translation	7.44E-40
GO:0046686	response to cadmium ion	2.85E-16
GO:0009793	embryo development ending in seed dormancy	2.00E-11
GO:0009658	chloroplast organization	3.00E-07
GO:0044267	cellular protein metabolic process	4.03E-07
GO:0055114	oxidation-reduction process	2.62E-06
GO:0006364	rRNA processing	2.35E-05
GO:0006144	purine base metabolic process	2.44E-05
GO:0006184	GTP catabolic process	3.54E-05
GO:0006448	regulation of translational elongation	4.97E-05
GO:0042026	protein refolding	5.33E-05
GO:0006094	gluconeogenesis	9.93E-05
GO:0009955	adaxial/abaxial pattern formation	2.04E-04
GO:0006206	pyrimidine base metabolic process	2.14E-04
GO:0006396	RNA processing	2.62E-04
GO:0015976	carbon utilization	2.62E-04
GO:0009408	response to heat	5.71E-04
GO:0006096	glycolysis	5.71E-04
GO:0009651	response to salt stress	1.46E-03
GO:0006522	alanine metabolic process	1.56E-03
GO:0008152	metabolic process	2.27E-03
GO:0022900	electron transport chain	2.66E-03
GO:0019430	removal of superoxide radicals	3.67E-03
GO:0006531	aspartate metabolic process	3.68E-03
GO:0051258	protein polymerization	4.90E-03
GO:0006446	regulation of translational initiation	5.10E-03
GO:0042128	nitrate assimilation	5.24E-03
GO:0042742	defense response to bacterium	5.49E-03
GO:0006260	DNA replication	6.38E-03
GO:0006544	glycine metabolic process	8.39E-03
GO:0006268	DNA unwinding involved in replication	1.06E-02
GO:0006422	aspartyl-tRNA aminoacylation	1.07E-02
GO:0006270	DNA-dependent DNA replication initiation	1.11E-02
GO:0051252	regulation of RNA metabolic process	1.11E-02
GO:0006413	translational initiation	1.37E-02
GO:0006457	protein folding	1.50E-02
GO:0015706	nitrate transport	1.83E-02
GO:0006098	pentose-phosphate shunt	1.93E-02
GO:0043086	negative regulation of catalytic activity	2.09E-02
GO:0000914	phragmoplast assembly	2.10E-02
GO:0009409	response to cold	2.15E-02

GO:0007018	microtubule-based movement	2.18E-02
GO:0006164	purine nucleotide biosynthetic process	3.70E-02
GO:0051131	chaperone-mediated protein complex assembly	4.70E-02
GO:0000059	protein import into nucleus, docking	4.82E-02
GO:0051301	cell division	4.82E-02
GO:0015995	chlorophyll biosynthetic process	4.86E-02
GO:0006529	asparagine biosynthetic process	4.95E-02

J. Over-represented GO slim categrories among genes expressed lower in *S. habrochaites* than in *S. pennellii*

Category	**Description**	**Adjusted P-value**
GO:0006952	defense response	9.31E-05
GO:0008152	metabolic process	3.48E-02
GO:0016126	sterol biosynthetic process	4.74E-02
GO:0006468	protein phosphorylation	4.74E-02
GO:0006633	fatty acid biosynthetic process	4.74E-02
GO:0055085	transmembrane transport	4.74E-02
GO:0009607	response to biotic stimulus	4.74E-02
GO:0006554	lysine catabolic process	4.74E-02

K. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0008152	metabolic process	4.28E-06
GO:0006952	defense response	4.60E-04
GO:0055114	oxidation-reduction process	1.38E-03

L. Over-represented GO slim categrories among genes expressed higher in *S. pimpinellifolium* than in *S. habrochaites*

Category	**Description**	**Adjusted P-value**
GO:0055114	oxidation-reduction process	8.63E-06
GO:0043086	negative regulation of catalytic activity	3.74E-03
GO:0006629	lipid metabolic process	3.74E-03
GO:0008152	metabolic process	3.74E-03

Table S9. Number of differentially expressed genes in pairwise comparisions.

	S. lyc	*S. pim*	*S. hab*	*S. pen*
S. lyc	0	1178	2547	3494
S. pim	1178	0	2026	3859
S. hab	2547	2026	0	3427
S. pen	3494	3859	3427	0

Table S10. Number of differentially expressed genes in pairwise comparisons which separate gene expression values into two groups between species.

	S. lyc	*S. pim*	*S. hab*	*S. pen*
S. lyc	0	373	887	1616
S. pim	373	0	780	1617
S. hab	887	780	0	1565
S. pen	1616	1617	1565	0

Table S11. Number of genes fitting BM2 model using varied thresholds

Filtered by DE adjusted p-value	deltaAIC BM2 best-fit over BM1 and OU	Total	gre	hab	lyc	pen	pim	red	rvg
No	> 4	3912	1042	342	471	770	381	798	108
No	> 7	1755	539	136	187	316	140	401	36
No	> 10	913	368	57	61	119	50	239	19
No	> 20	58	0	18	0	40	0	0	0
0.05	> 4	2100	650	189	191	531	126	338	75
0.05	> 7	977	332	94	94	223	56	154	24
0.05	> 10	498	204	45	39	88	19	93	10
0.05	> 20	42	0	14	0	28	0	0	0
0.01	> 4	1764	554	162	145	472	93	270	68
0.01	> 7	842	292	82	73	205	43	125	22
0.01	> 10	428	179	38	36	78	13	75	9
0.01	> 20	41	0	14	0	27	0	0	0

Table S12. Number of genes in each Coexpression module

	Spen.green	Spen.NA	Spen.purple	Spen.yellow
Slyc.green	852	170	2	0
Slyc.NA	552	2538	153	120
Slyc.orange	6	23	1	0
Slyc.purple	10	120	272	0
Slyc.yellow	1	133	0	144

Table S13. GO enrichment for green module

category	description	p_value	FDR
GO:0009765	photosynthesis, light harvesting	1.91E-23	2.85E-20
GO:0015979	photosynthesis	2.38E-23	2.85E-20
GO:0015995	chlorophyll biosynthetic process	1.27E-19	1.02E-16
GO:0009773	photosynthetic electron transport in photosystem I	1.31E-15	7.84E-13
GO:0019253	reductive pentose-phosphate cycle	2.20E-15	1.06E-12
GO:0015976	carbon utilization	3.60E-15	1.44E-12
GO:0018298	protein-chromophore linkage	2.27E-13	7.78E-11
GO:0006412	translation	3.83E-13	1.15E-10
GO:0042254	ribosome biogenesis	2.63E-10	6.67E-08
GO:0009658	chloroplast organization	2.78E-10	6.67E-08
GO:0016117	carotenoid biosynthetic process	2.93E-09	6.40E-07
GO:0010027	thylakoid membrane organization	4.18E-09	8.36E-07
GO:0009767	photosynthetic electron transport chain	6.91E-09	1.28E-06
GO:0006096	glycolysis	1.95E-08	3.28E-06
GO:0006013	mannose metabolic process	2.05E-08	3.28E-06
GO:0008152	metabolic process	2.28E-08	3.42E-06
GO:0006098	pentose-phosphate shunt	5.47E-08	7.72E-06
GO:0000413	protein peptidyl-prolyl isomerization	6.06E-08	8.09E-06
GO:0006094	gluconeogenesis	9.14E-08	1.16E-05
GO:0046487	glyoxylate metabolic process	3.78E-07	4.54E-05
GO:0045038	protein import into chloroplast thylakoid membrane	6.00E-07	6.86E-05
GO:0006000	fructose metabolic process	7.19E-07	7.85E-05
GO:0042742	defense response to bacterium	1.32E-06	0.000137399
GO:0009853	photorespiration	1.74E-06	0.000174068
GO:0009768	photosynthesis, light harvesting in photosystem I	2.52E-06	0.000242087
GO:0055114	oxidation-reduction process	2.80E-06	0.000258449
GO:0009637	response to blue light	2.95E-06	0.000262437
GO:0006563	L-serine metabolic process	3.67E-06	0.000314681
GO:0009409	response to cold	4.06E-06	0.000336342
GO:0010114	response to red light	6.17E-06	0.000494162
GO:0006566	threonine metabolic process	6.52E-06	0.000505002
GO:0019464	glycine decarboxylation via glycine cleavage system	1.26E-05	0.000945728
GO:0006020	inositol metabolic process	1.49E-05	0.001081834
GO:0006457	protein folding	2.63E-05	0.00185465
GO:0048481	ovule development	4.22E-05	0.002895246
GO:0043085	positive regulation of catalytic activity	5.11E-05	0.00340707
GO:0010275	NAD(P)H dehydrogenase complex assembly	6.41E-05	0.004161631
GO:0010258	NADH dehydrogenase complex (plastoquinone) assembly	6.67E-05	0.004213595
GO:0010190	cytochrome b6f complex assembly	7.71E-05	0.004663422
GO:0010196	nonphotochemical quenching	7.81E-05	0.004663422
GO:0010304	PSII associated light-harvesting complex II catabolic process	7.96E-05	0.004663422

GO:0016120	carotene biosynthetic process	8.21E-05	0.004692707
GO:0009416	response to light stimulus	0.000130788	0.007302812
GO:0016226	iron-sulfur cluster assembly	0.000134447	0.007336513
GO:0019252	starch biosynthetic process	0.000237474	0.012670533
GO:0045454	cell redox homeostasis	0.000242826	0.012674484
GO:0015977	carbon fixation	0.00024834	0.012686463
GO:0010207	photosystem II assembly	0.000275755	0.013793495
GO:0010206	photosystem II repair	0.000282936	0.013863854
GO:0010020	chloroplast fission	0.000301195	0.014187457
GO:0009772	photosynthetic electron transport in photosystem II	0.000301358	0.014187457
GO:0019684	photosynthesis, light reaction	0.000381273	0.01760455
GO:0019288	isopentenyl diphosphate biosynthetic process, mevalonate-independent pathway	0.000605579	0.02694442
GO:0010236	plastoquinone biosynthetic process	0.000611166	0.02694442
GO:0033014	tetrapyrrole biosynthetic process	0.000617219	0.02694442
GO:0006783	heme biosynthetic process	0.000701622	0.029833498
GO:0009793	embryo development ending in seed dormancy	0.00070825	0.029833498
GO:0006352	transcription initiation, DNA-dependent	0.000733094	0.030347568
GO:0009250	glucan biosynthetic process	0.00120283	0.048949068

Table S14. GO slim enrichment for green module

category	description	p_value	FDR
GO:0015979	photosynthesis	8.80E-94	4.22E-92
GO:0006091	generation of precursor metabolites and energy	3.58E-48	8.59E-47
GO:0006412	translation	3.43E-17	5.49E-16
GO:0009058	biosynthetic process	1.55E-13	1.86E-12
GO:0008152	metabolic process	3.71E-12	3.56E-11
GO:0009987	cellular process	1.25E-11	9.97E-11
GO:0006139	nucleobase-containing compound metabolic process	4.30E-10	2.95E-09
GO:0019748	secondary metabolic process	8.86E-07	5.32E-06
GO:0009056	catabolic process	0.002079155	0.01045928
GO:0009628	response to abiotic stimulus	0.002179017	0.01045928
GO:0019725	cellular homeostasis	0.003984236	0.017385755
GO:0009790	embryo development	0.007167	0.028667999
GO:0005975	carbohydrate metabolic process	0.007765541	0.028672767
GO:0019538	protein metabolic process	0.011298128	0.038075981
GO:0016043	cellular component organization	0.011898744	0.038075981

Table S15. GO enrichment for purple module

category	description	p_value	FDR
GO:0006979	response to oxidative stress	5.29E-17	1.27E-13
GO:0006869	lipid transport	1.86E-13	2.24E-10
GO:0055114	oxidation-reduction process	4.99E-09	3.99E-06
GO:0006952	defense response	1.08E-05	0.006457725
GO:0009561	megagametogenesis	9.09E-05	0.0436426

Table S16. GO slim enrichment for purple module

category	description	p_value	FDR
GO:0006950	response to stress	7.14E-06	0.000342696
GO:0008152	metabolic process	0.000107634	0.002583226
GO:0006810	transport	0.000366826	0.005869221
GO:0005975	carbohydrate metabolic process	0.011078457	0.132941488
GO:0009606	tropism	0.067883469	0.651681304

Table S17. GO enrichment for yellow module

category	description	p_value	FDR
GO:0007018	microtubule-based movement	9.88E-25	2.37E-21
GO:0006334	nucleosome assembly	1.28E-12	1.53E-09
GO:0000914	phragmoplast assembly	2.93E-08	2.34E-05
GO:0000079	regulation of cyclin-dependent protein kinase activity	5.90E-08	3.54E-05
GO:0051301	cell division	4.11E-07	0.00019728
GO:0006260	DNA replication	1.02E-06	0.000396094
GO:0046785	microtubule polymerization	1.15E-06	0.000396094
GO:0051225	spindle assembly	1.45E-06	0.000433797
GO:0006270	DNA-dependent DNA replication initiation	2.14E-06	0.000570911
GO:0000910	cytokinesis	4.07E-06	0.00097685
GO:0010342	endosperm cellularization	3.96E-05	0.008553911
GO:0006268	DNA unwinding involved in replication	4.52E-05	0.008553911
GO:0055046	microgametogenesis	4.63E-05	0.008553911
GO:0000723	telomere maintenance	0.000132184	0.020468879
GO:0052096	formation by symbiont of syncytium involving giant cell for nutrient acquisition from host	0.000133746	0.020468879
GO:0000911	cytokinesis by cell plate formation	0.000136402	0.020468879

Table S18. GO slim enrichment for yellow module

category	description	p_value	FDR
GO:0007049	cell cycle	2.36E-19	1.13E-17
GO:0016043	cellular component organization	2.82E-12	6.77E-11
GO:0006259	DNA metabolic process	9.70E-10	1.55E-08
GO:0009987	cellular process	0.000147009	0.001764105
GO:0007610	behavior	0.00654818	0.062862524
GO:0007275	multicellular organismal development	0.031446135	0.25156908
GO:0009653	anatomical structure morphogenesis	0.039847821	0.273242204

Supplemental Table ST19 - Predicted genes involved in wax biosynthesis that show differential expression in the seedling dataset.
All % identities are to Arabidopsis transcripts except for Solyc10g075100.1.1 (LTP2)[+] which is % identity to tomato transcript and CER2[#] which is % identity to Arabidopsis protein sequence.
FDR adjusted P-valu

	Seedling dataset						
Gene ID	**Species Adjusted P-value**	**log2 Fold Change (*S. pen - S. lyc*)**	**AGI Accession No.**	**Gene name**	**Description**	**% Identity**	**Reference**
Solyc03g065250.2.1	0.253	1.324	AT1G02205.2	CER1	Expression of the CER1 gene associated with production of stem epicuticular wax and pollen fertility. Biochemical studies showed that cer1 mutants are blocked in the conversion of stem wax C30 aldehydes to C29 alkanes, and they also lack the secondary alcohols and ketones. These suggested the CER1 protein is a aldehyde decarbonylase.	65.7	Aarts, 1995
Solyc01g088400.2.1	**0.000**	**3.174**	AT1G02205.2	CER1	Expression of the CER1 gene associated with production of stem epicuticular wax and pollen fertility. Biochemical studies showed that cer1 mutants are blocked in the conversion of stem wax C30 aldehydes to C29 alkanes, and they also lack the secondary alcohols and ketones. These suggested the CER1 protein is a aldehyde decarbonylase.	62.42	Aarts, 1995
Solyc08g044260.2.1	**0.000**	**9.261**	AT1G02205.2	CER1	Expression of the CER1 gene associated with production of stem epicuticular wax and pollen fertility. Biochemical studies showed that cer1 mutants are blocked in the conversion of stem wax C30 aldehydes to C29 alkanes, and they also lack the secondary alcohols and ketones. These suggested the CER1 protein is a aldehyde decarbonylase.	60.51	Aarts, 1995
Solyc01g088430.2.1	**0.019**	**0.654**	AT1G02205.2	CER1	Expression of the CER1 gene associated with production of stem epicuticular wax and pollen fertility. Biochemical studies showed that cer1 mutants are blocked in the conversion of stem wax C30 aldehydes to C29 alkanes, and they also lack the secondary alcohols and ketones. These suggested the CER1 protein is a aldehyde decarbonylase.	60.1	Aarts, 1995
Solyc05g054490.2.1	**0.000**	**1.885**	AT3G55360.1	CER10	Enoyl-CoA reductase is involved in all very long chain fatty acids (VLCFA) elongation reactions that are required for cuticular wax, storage lipid and sphingolipid metabolism.	80.97	Zheng, 2005
Solyc02g085870.2.1	**0.003**	**1.975**	AT1G68530.1	CER6	Involved in wax biosynthesis; required for elongation of C24 very-long-chain fatty acids	83.5	Millar, 1999; Fiebig, 2000
Solyc02g063140.2.1	**0.000**	**3.270**	AT1G68530.1	CER6	Involved in wax biosynthesis; required for elongation of C24 very-long-chain fatty acids	82.9	Millar, 1999; Fiebig, 2000
Solyc05g009270.2.1	**0.000**	**2.808**	AT1G68530.1	CER6	Involved in wax biosynthesis; required for elongation of C24 very-long-chain fatty acids	73.48	Millar, 1999; Fiebig, 2000
Solyc08g067260.2.1	**0.003**	**1.208**	AT2G26250.1	FDH	Epidermis-specific, encodes a putative beta-ketoacyl-CoA synthase. probably involved in the synthesis of long-chain lipids found in the cuticle.	69.15	Yephremov, 1999
Solyc12g087980.1.1	0.305	1.810	AT4G24510.1	CER2	Involved in C28 to C30 fatty acid elongation	36.00[#]	Negruk,
Solyc10g075110.1.1	**0.000**	**2.570**	AT2G38540.1	ATLTP1	Non-specific lipid transfer protein. Binds calmodulin in a Ca2+-independent manner. Localized to the cell wall. Specifically expressed in L1 epidermal layer.	51.69	Thoma, 1994; Trevino, 1998
Solyc10g075100.1.1	**0.000**	**2.456**	NA	LpLTP2	Non-specific lipid transfer protein. Binds calmodulin in a Ca2+-independent manner. Localized to the cell wall. Specifically expressed in L1 epidermal layer.	95.00[+]	Trevino, 1998
Solyc11g065350.1.1	0.056	1.154	AT3G21090.1	CER5-like	ABC transporter family protein; Identical to White-brown complex homolog protein 15 (White-brown complex homolog protein 22) (WBC15) [Arabidopsis Thaliana] (GB:Q8RWI9;GB:Q9LJC3); similar to CER5 (ECERIFERUM 5), ATPase, coupled to transmembrane movement of substances [Arabidopsis thaliana] (TAIR:AT1G51500.1).	68.8	Pighin, 2004
Solyc11g065360.1.1	0.000	6.923	AT3G21090.1	CER5-like	ABC transporter family protein; Identical to White-brown complex homolog protein 15 (White-brown complex homolog protein 22) (WBC15) [Arabidopsis Thaliana] (GB:Q8RWI9;GB:Q9LJC3); similar to CER5 (ECERIFERUM 5), ATPase, coupled to transmembrane movement of substances [Arabidopsis thaliana] (TAIR:AT1G51500.1).	62.39	Pighin, 2004

Solyc05g047420.2.1	0.008	0.740	AT3G60500.1	CER7	3' exoribonuclease family protein; similar to 3'-5'-exoribonuclease/ RNA binding [Arabidopsis thaliana] (TAIR:AT3G12990.2); similar to exosome component 9 [Gallus gallus] (GB:NP_001030000.1); similar to Os02g0550700 [Oryza sativa (japonica cultivar-group)] (GB:NP_001047096.1); contains InterPro domain Exoribonuclease; (InterPro:IPR001247)	63.3	Hooker, 2007
Solyc01g079240.2.1	0.033	1.332	AT2G47240.2	CER8	long-chain-fatty-acid--CoA ligase family protein / long-chain acyl-CoA synthetase family protein; similar to long-chain-fatty-acid--CoA ligase, putative / long-chain acyl-CoA synthetase, putative [Arabidopsis thaliana] (TAIR:AT4G11030.1).	63.58	Lu, 2009
Solyc06g074390.2.1	0.000	0.132	AT4G33790.1	CER4	acyl CoA reductase, putative; similar to acyl CoA reductase, putative / male-sterility protein, putative [Arabidopsis thaliana] (TAIR:AT5G22500.1).	61.22	Rowland, 2006
Solyc11g067190.1.1	0.009	0.547	AT4G33790.1	CER4	acyl CoA reductase, putative; similar to acyl CoA reductase, putative / male-sterility protein, putative [Arabidopsis thaliana] (TAIR:AT5G22500.1).	60.08	Rowland, 2006
Solyc03g117800.2.1	0.044	0.947	AT5G57800.1	CER3	encodes a transmembrane protein with similarity to the sterol desaturase family at the N-terminus and to the short-chain dehydrogenase/reductase family at the C-terminus. Mutant analyses indicate this protein is involved in cuticle membrane and wax biosynthesis.	65.44	Rowland, 2007
Solyc03g116610.2.1	0.001	0.229	AT1G15360.1	SHN1	encodes a member of the ERF (ethylene response factor) subfamily B-6 of ERF/AP2 transcription factor family. The protein contains one AP2 domain. There are 12 members in this subfamily including RAP2.11. This gene is involved in wax biosynthesis.	65.22	Aharoni, 2004

Datasets

Dataset S1.Polymorphisms

Coding_polymorphisms_and_effects
Noncoding_polymorphisms
Polymorphisms_in_genes.no_effect_calculated
SNPs_in_common_with_potato

Dataset S2. Genes inferred to be under coding sequence positive selection. Tomato genes are identified by their ITAG number, and their putative *A. thaliana* homologs are specified by their AGI number, gene symbol, and description. For each gene, the best-fit nucleotide substitution model used for inference is specified in Hy-Phy notation.

Dataset S3. Genes under directional selection for gene expression level. Genes are considered to be under directional selection when their expression levels are better fit by a 2-rate Brownian motion model (BM2) than either BM1 or Ornstein–Uhlenbeck (OU) models.

Dataset S4. Transcript abundance and differential expression. This excel sheet contains the differential expression analysis datasheets in the following tabs:

Tissue_Species_Diff_Exp. Differential expression in *S. lycopersicum* var. M82 and *S. pennellii.* For each gene in the matched ITAG set, raw and multiple-testing corrected P-values for differential expression between species, between tissues, and for a different tissues expression between the two species. Also provided are the average expressions and the differences between the two species (log2 of fitted read counts).

4_Species_Diff_Exp_WM. Differential expression in *S. lycopersicum* var. M82 (SLY), *S. pimpinellifolium* (SPI), *S. habrochaites* (SHA), and *S. pennellii* (SPE). Multiple-testing corrected P-values are given for a model including all 4 species analyzed together. Significant genes in this table are those with evidence for there being an effect of the species on expression.

4_Species_Diff_Exp_Pairwise. For genes with evidence of a species-level effect ($P < 0.01$), pairwise comparisons between *S. lycopersicum* var. M82 (SLY), *S. pimpinellifolium* (SPI), *S. habrochaites* (SHA), and *S. pennellii* (SPE) were made. This table provides multiple-testing corrected P-values and log2 fold-change expression differences for each pairwise comparison performed.

IL4_3_Diff_Exp. Pairwise log2 gene expression differences between IL4-3, *S. lycopersicum* var. M82, and *S. pennellii* (PEN). Genes that were found to be differentially expressed between M82 and PEN in this experiment were further classified as follows. Genes within the introgressed region are classified as having PEN-like expression (PENNgenoPennExpr), M82-like expression (PENNgenoM82Exp), or

expression not like either parent (PENNgenoINT). Genes not in the introgressed region are classified as having M82-like expression (M82genoM82expr), PEN-like expression (M82genoPennExpr), or expression not like either parent (M82genoINT.csv).

Dataset S5. File giving module membership for each gene from co-expression network analysis.

Dataset S6. GO annotation. Tab-delimited text file with merged results from Blast2GO and ITAG2.3 GO terms. Only genes with at least one GO term are shown. Warning: opening this file in Excel causes leading zeros to be lost and is not recommended.

Dataset S7. GOslim annotation. Tab-delimited text file with merged results from Blast2GO using plant GOslim terms. Only genes with at least one GO term are shown. Warning: opening this file in Excel causes leading zeros to be lost and is not recommended.

Dataset S8. Arabidopsis annotation. Best BLAST hit from a blastp analysis of ITAG2.3 predicted proteins against Arabidopsis TAIR10 annotated proteins. Results were filtered to only retain hits with more than 50% identity across the alignment, with more than 75% of the ITAG2.3 protein participating in the alignment and with an e-value < 1e-20. ITAG 2.3 gene models not listed in this table did not pass the filter criteria.

Dataset S9. Primer and PCR information. Information related to validating SNPs, indels, and differential expression patterns, including primers used for PCR, enzymes used for CAPS, annealing temperatures for RT-PCR/qRT-PCR and validation results. The file contains six tabs: Legend_CAPS, CAPS, Legend_SSR, SSR, RT-PCR, and qRT-PCR.