Sequence diversity in three tomato species: SNPs, markers, and molecular evolution
José M Jiménez-Gómez and Julin N Maloof*

Address: Department of Plant Biology, College of Biological Sciences, University of California Davis, Davis, CA, 95616, USA
Email: José M Jiménez-Gómez - jmjimenez@ucdavis.edu; Julin N Maloof* - jnmaloof@ucdavis.edu
* Corresponding author

Abstract

Background: Tomato species are of significant agricultural and ecological interest, with cultivated tomato being among the most common vegetable crops grown. Wild tomato species are native to diverse habitats in South America and show great morphological and ecological diversity that has proven useful in breeding programs. However, relatively little is known about nucleotide diversity between tomato species. Until recently limited sequence information was available for tomato, preventing genome-wide evolutionary analyses. Now, an extensive collection of tomato expressed sequence tags (ESTs) is available at the SOL Genomics Network (SGN). This database holds sequences from several species, annotated with quality values, assembled into unigenes, and tested for homology against other genomes. Despite the importance of polymorphism detection for breeding and natural variation studies, such analyses in tomato have mostly been restricted to cultivated accessions. Importantly, previous polymorphisms surveys mostly ignored the linked meta-information, limiting functional and evolutionary analyses. The current data in SGN is thus an under-exploited resource. Here we describe a cross-species analysis taking full-advantage of available information.

Results: We mined 20,000 interspecific polymorphisms between Solanum lycopersicum and S. habrochaites or S. pennellii and 28,800 intraspecific polymorphisms within S. lycopersicum. Using the available meta-information we classified genes into functional categories and obtained estimations of single nucleotide polymorphisms (SNP) quality, position in the gene, and effect on the encoded proteins, allowing us to perform evolutionary analyses. Finally, we developed a set of more than 10,000 between-species molecular markers optimized by sequence quality and predicted intron position. Experimental validation of 491 of these molecular markers resulted in confirmation of 413 polymorphisms.

Conclusion: We present a new analysis of the extensive tomato EST sequences available that represents the most comprehensive survey of sequence diversity across Solanum species to date. These SNPs, plus thousands of molecular makers designed to detect the polymorphisms are available to the community via a website. Evolutionary analyses on these polymorphism uncovered sets of genes potentially important for the evolution and domestication of tomato; interestingly these sets were enriched for genes involved in response to the environment.
Background

Tomato (*Solanum lycopersicum*) is the second most popular vegetable crop in the world [1]. In addition, tomato is being developed as a model organism and is more closely related to important crops like lettuce and coffee than other models such as Arabidopsis, poplar or rice [2]. Tomato also has an interesting natural history, which includes a single domesticated species and a number of wild relatives that have wide morphological variability and are adapted to very diverse environments [3,4]. The comparative study of tomato wild species can help us identify key genetic factors involved in domestication and will benefit breeding programs.

Despite the advent of high throughput genomic techniques and bioinformatics, comprehensive genome-wide information remains unavailable for most species, including tomato. Since the tomato genome sequence is currently incomplete and the microarray platforms for this species do not feature most of the loci predicted to exist [5-7], genetic studies have focused on the analysis of particular loci and segregating populations. These studies led to in-depth information on a few loci and uncovered the potential usefulness of the natural variation existing in related species [3,4]. A major goal for tomato geneticists is the acquisition of comprehensive genome-wide information that can be used in the improvement of resistance, quality, aspect, flavor and growth in cultivated varieties [8].

A large amount of genetic and molecular information is available for tomato, most of which has been deposited at the SGN [http://sgn.cornell.edu](http://sgn.cornell.edu). In this database there are more than 320,000 ESTs from several tomato species. This abundant sequence repository has been used to develop polymerase chain reaction-based (PCR-based) molecular markers to build on the original, time consuming and cost ineffective restriction fragment length polymorphisms (RFLPs) [9-12].

Recently, bioinformatic surveys of genome sequences have revealed the importance of SNPs in shaping evolution and also in serving as molecular markers [13]. In tomato, initial genomic work on SNP discovery involved de novo sequencing of EST libraries and comparison to the existing databases [14], or faster and less expensive computer-aided mining of the available sequences [15,16]. In these analyses only the information from the plain sequence was used. As a consequence there were high numbers of both false positives created by sequencing errors, and false negatives resulting from data that did not meet the conditions set by the researchers, such as minimum number of sequences or sequence similarity surrounding a SNP. Furthermore, low rates of validation were often obtained, in part because of lack of information on sequence quality and on intron position. For example, if a predicted marker spans an intron it will be difficult to detect by PCR.

There is now an increasing wealth of resources in the tomato databases and similar repositories such as sequence qualities, unigene assemblies, open reading frame (ORF) predictions, sequence similarity with other plant species and localization to genetic maps [17]. The use of this metadata allows for more sophisticated experimental designs that increase the quality of SNP prediction and information on the SNP’s effects. For example, by comparing tomato and *Arabidopsis thaliana* databases a set of markers was developed that targeted loci conserved throughout evolution both in sequence and copy number [18,19]. This Conserved Orthologous Set (COS) has been proven an invaluable resource for comparative and evolutionary studies among plants [20-22]. Similarly, a recent publication used intron positions conserved between tomato and *Arabidopsis thaliana* to detect intraspecific SNPs in noncoding regions, demonstrating the possibilities of bioinformatics predictions [23].

Despite the availability of sequence-associated metadata and the need for functional genomic studies in tomato, there is surprisingly little information about the relevance and levels of variation found in coding regions. Previous SNP mining efforts are either based on noncoding regions or have no information about the effect of the polymorphisms on protein sequence. Moreover, most SNPs mined from EST sequences have little probability of being functionally important, since the majority are expected to fall in the un-translated regions (UTR), where the polymorphism rate increases more than five fold in comparison with coding regions, [23]. Information about the number, localization and type of non-synonymous polymorphisms will help unravel useful information about the evolution of genes that may have adaptive significance [24], serving as a primer for more profound studies on natural variation of interesting traits.

To perform reliable evolutionary and phylogenetic analyses sufficient polymorphism data can be obtained from the sequences of the wild species already available. However, most studies have focused on finding intraspecific molecular markers between cultivated tomato cultivars. The few attempts to infer rates of polymorphism and selection on tomato genes among its wild relatives are reduced to small sets of genes likely providing biased estimations [25].

To fill the gap in knowledge about sequence similarities and differences among cultivated and wild tomato species, we mined the EST and unigene tomato database from SGN in search for substitutions and insertions/deletions between and within the three species with highest representation: *S. lycopersicum*, *S. habrochaites* and *S. pennellii*.
We used additional metadata from the database to effectively predict SNPs and infer levels of polymorphism in coding versus noncoding regions and in gene families. We surveyed this dataset for signatures of sequence evolution, selection and/or adaptation using the McDonald-Kreitman test [26] and codon-based maximum likelihood analyses [27,28]. Based on the polymorphisms detected we also developed a set of specific molecular markers and a website to make these available to the community at http://www.plb.ucdavis.edu/labs/maloof/TomatoSNP/index.html.

Results and discussion

EST assembly
The unigene database version 200607 build 1 from SGN consists of 239,172 ESTs grouped into 34,829 unigenes, each one containing between 1 and 1087 ESTs, with an average of 6.9 ESTs per unigene (data not shown). Unigenes can be formed by EST sequences from any tomato race or species, although most unigenes (87%) contain sequences from a single species (Figure 1).

We mined sequence assemblies of *S. lycopersicum* (L) ESTs for intraspecific polymorphisms and assemblies containing sequences from both *S. lycopersicum* and *S. habrochaites* (LxH) or *S. lycopersicum* and *S. pennellii* (LxP) for interspecific SNPs. To reduce the complexity of each analysis we reassembled the unigenes using only ESTs from the species relevant to the analysis. The consensus sequence for each reassembled contig was then aligned to the original unigene sequence and, when available (88.89% of the unigenes, data not shown), to the predicted coding sequence (CDS). In the cases where a predicted CDS was available, each nucleotide in the assembly was annotated as 5' UTR, CDS or 3' UTR. We assigned quality scores to each position in the assemblies by calculating separately for each species the sum of the qualities of all nucleotides at that particular position.

Several filters were applied to diminish the number of false positive SNPs predicted in our analyses. Assembly regions that did not align with unigene sequences were removed. We also discarded positions where only a single sequence was found and, in the interspecific analyses, assembly regions where sequences from only one of the species assayed were detected. Using the remaining portions of the assemblies we determined an optimum quality threshold for each analysis at which the average sequence quality was maximized (see Methods, Figure 2). Assembly positions that did not pass the quality threshold were removed, leaving for SNP mining a total of 4,712 unigenes spanning 1,736 Kb of interspecific assemblies and 19,159 unigenes including 11,058 Kb of intraspecific *S. lycopersicum* data (Table 1).

In every analysis the majority of nucleotides assembled were located in predicted coding regions. Between 2.8 and 6.2% of the positions could not be assigned to a gene region due to the lack of predicted CDS (Table 1). The average quality of the nucleotides considered in the assemblies was always highest in the 5' regions and lowest in the 3' regions (Figure 3). This could be explained by the bias towards sequencing the ESTs from the 5', as an average of 95.87% of the sequences from each library were readings from the top strand (data not shown). Interestingly, assemblies from unigenes for which there is no predicted CDS had overall average qualities similar to the 5' region (Figure 3). This raises the possibility that those unigenes contain mostly 5' UTR sequence, in turn explaining why ESTscan [29,30] was not able to predict a CDS.

**SNP mining**

We mined the resulting sequence assemblies for nonconsecutive SNPs taking into account the position of the SNP
with respect to the predicted gene regions. In total 40,834 substitutions and 8,266 insertions/deletions (indels) were detected (summarized in Table 1; see Additional files 1, 2, 3 & 4 for full data). Several observations confirm the performance of our SNP detection algorithms. As shown in Table 1, indels appear between 2.6 and 8.7 times more frequently in UTRs than in the CDS. Insertion and deletion of nucleotides are likely to produce changes in the open reading frames disrupting correct translation, therefore selection against them is expected in coding regions. Similarly, predicted noncoding regions yielded between 1.8 and 2.9 times more SNPs per kilobase analyzed than coding regions (Table 1), as reported in previous tomato analyses [15,19,23,25]. We next used the estimated CDS for each unigene to calculate the codon position for each SNP. In translated regions the majority of the mutations are located in the third position of the codon (Figure 4). This result was expected since selective pressure in coding regions reduces the number of non-synonymous substitutions and the redundancy of the genetic code is mainly in the third base of the codons. We then calculated hypothetical codons in the UTRs by extending the ORF from the origin of transcription in the predicted CDS. The third position bias disappeared in these hypothetical noncoding regions (Figure 4) suggesting that on average the SNP, CDS, and codon position predictions are accurate. In addition, when looking at all SNPs in coding regions

| Analysis | LxP | LxH | L |
|----------|-----|-----|---|
| CDS      | 2249| 2463| 19159|
| UTR      | 116.8| 203.6| 1962.1|
| NA       | 21.5| 29.7| 649.6|
| CDS      | 860.1| 3501/ | 14455/ |
| UTR      | 8457/ | 349/ | 8229/ |
| NA       | 300/ | 64/ | 919/ |
| SNPs/indels | 92/30 | 614| 1980|
| Unigenes with SNPs | 1341| 1934| 5084|
| SNP rate | 0.8852| 0.334| 0.2605|
| SNP rate per region | 0.1946| 0.5726| 0.1755|

Only nonconsecutive SNPs are displayed. NA - gene region is not available for unigenes in which a CDS could not be predicted. SNP rates are calculated as the number of SNPs per 100 bp.

Quality threshold estimation. EST assemblies were analyzed at different quality thresholds by removing those positions of the assemblies whose sum of qualities were below the threshold (see Methods). The average quality of the sequences in the remaining positions was plotted (y-axis) against the range of quality thresholds tested (x-axis). L – *S. lycopersicum* assemblies, LxH – assemblies including *S. habrochaites* and *S. lycopersicum* ESTs, LxP – assemblies including *S. lycopersicum* and *S. pennelli* ESTs. Maximum average sequence quality is achieved with thresholds of 40 for LxH and LxP and 50 for the L assemblies.

Differences in average sequence quality between gene regions. Average quality of the sequences mined for SNPs in each analysis grouped by estimated gene region. Error bars represent standard deviation. LxP – Assemblies including *S. lycopersicum* and *S. pennelli* ESTs, LxH – Assemblies including *S. habrochaites* and *S. lycopersicum* ESTs, L – *S. lycopersicum* assemblies.

---

**Table 1: Unigenes, SNPs, and SNP rates**

| Analysis | LxP | LxH | L |
|----------|-----|-----|---|
| Unigenes analyzed | 2249| 2463| 19159|
| Kb analyzed | 504.9| 116.8| 1962.1|
| SNPs/indels | 206| 534| 3006|
| Unigenes with SNPs | 1341| 1934| 5084|
| SNP rate | 0.8852| 0.334| 0.2605|
| SNP rate per region | 0.1946| 0.5726| 0.1755|
together, the percentage of non-synonymous SNPs is 46.37%, similar to genome-wide analyses in Arabidopsis (45.34%) [31] and humans (46.46%) [32].

Regarding the differences between intra and interspecific analyses, we found only 26.5% of the analyzed loci to be polymorphic when mining a single species versus the 69.5% when analyzing pairs (Table 1). For interspecific analysis the SNP rates were comparable to those published before between S. pennellii and S. lycopersicum (1.02 to 1.61 SNPs/100 bp) making no distinction between coding and non coding regions [14,33]. The same holds true for S. lycopersicum intraspecific analyses, where reported SNP rates range between 0.0117 and 0.585 SNPs/100 bp on all EST sequences analyzed regardless of their location in the gene [14-16,25].

SNP representation in Gene Ontology classes
We reasoned that certain gene families might be more variable among tomato cultivars and species than others. For example, since tomato species are adapted to diverse environments, genes involved in environmental response might accumulate a higher number of non-synonymous SNPs due to selection. We used Gene Ontology (GO) categories [34], which group genes into functionally related classes, to assess the differential polymorphic rates of uni-genes encoding specific classes of proteins. We assigned GO categories to each nucleotide position in the assemblies based on the closest Arabidopsis thaliana homolog, and looked for categories with over- or under-representation of non-synonymous SNPs. To achieve higher power we grouped together the data from the interspecific analyses. The larger number of sequences available in the intraspecific analysis allowed greater statistical power, although the most over- and underrepresented gene classes are in agreement in both analyses (Figure 5). As expected, GO categories related to environment interaction, such as responses to stress and abiotic stimulus, had an over-representation of non-synonymous polymorphisms both between and within species (p < 0.001, Figure 5). On the other hand, categories involved in basic biological processes and transcription regulation showed a relative lack of polymorphisms, as had been found in Arabidopsis thaliana accessions [31]. Surprisingly, genes encoding ribosomal proteins also presented more SNPs that expected, perhaps due to the challenge of distinguishing homologs from paralogs in this gene family, which shows complex patterns of copy number variants in Arabidopsis [35].

McDonald-Kreitman test
The neutral evolution theory holds that most within species polymorphisms and between species differences have little fitness consequence [36]. As a result, the ratio of non-synonymous to synonymous substitutions should be similar for a particular gene both within and between species. Deviations from this can be a sign of non-neutral evolution [26]. To examine this prediction, we con-
structured alignments of the wild and cultivated consensus sequences containing only the good quality positions and the SNPs predicted by our algorithms, and surveyed each alignment for signatures of positive selection using the McDonald-Kreitman test [26]. With this method we tested 1425 unigenes from the *S. lycopersicum* and *S. habrochaites* analysis and 924 unigenes from the *S. lycopersicum* and *S. pennellii* analysis. Before correction for multiple testing we found significant excess of non-synonymous mutations (p < 0.05) in 16 unigenes in the LxH analysis and 3 unigenes in the LxP analysis (data not shown). However, none of those unigenes survived the Benjamini and Hochberg correction for multiple testing [37].

**Maximum likelihood codon-substitution models**

The McDonald-Kreitman test compares variation between and within two species. An alternative approach is to simultaneously analyze all the sequences in a phylogenetic tree [38]. Following this method, we surveyed the unigenes for signals of positive selection using maximum likelihood estimates from codon-substitution models [27,28]. We reasoned that genes important for domestication might show a higher rate of non-synonymous to synonymous substitutions specifically in the *S. lycopersicum* lineage. To test this idea, we fit two different models for each unigene using alignments of the coding regions from cultivated tomato, *Arabidopsis thaliana* and *S. habrochaites* or *S. pennellii*. One of the models used assumed similar non-synonymous/synonymous rates (d_n/d_s) in every branch of the evolutionary tree (one-branch rate model). The second model allowed for different d_n/d_s estimates in the branch leading to *S. lycopersicum* relative to the rest of the tree (two-branch rates model). For each gene we used a likelihood ratio test to ask if the two-branch model fit significantly better than the one-branch model. Support for two-branch rates is suggestive of directional selection on the branch leading to domesticated tomato, therefore raising the possibility of identifying genes important for domestication [38]. Alternatively, *S. lycopersicum*-specific d_n/d_s ratios could occur due to natural selection acting...
Molecular marker design
Using the information gathered from the assemblies we developed a set of molecular markers for detecting the predicted SNPs. Successful design of PCR based molecular markers from EST sequences requires the ability to avoid amplifying introns. Since many Arabidopsis intron positions are conserved in tomato [23], we used this information (see Methods) to inform the design of molecular markers for the polymorphisms identified in the interspecific analyses. First, we tested our intron predictions by designing three primer pairs surrounding predicted introns. One of those produced a band 700 bp greater than expected without introns and the other two did not amplify, suggesting the existence of introns where predicted (data not shown). We did not carry out similar analysis for the intraspecific SNPs due to the lack of information within SGN regarding the *S. lycopersicum* accessions used. Although it is well known that *S. lycopersicum* ESTs in SGN come from several cultivars, the information for each individual EST library is not available. For each high-quality, interspecific SNP we developed a database containing suggested primers to amplify the SNP region, restriction enzymes to detect the polymorphism, and predicted fragment sizes before and after digestion of each allele. This information can be accessed at [http://www.plb.ucdavis.edu/labs/maloof/TomatoSNP/index.html](http://www.plb.ucdavis.edu/labs/maloof/TomatoSNP/index.html).

We validated a subset of markers by using 491 primer pairs designed to amplify fragments from *S. lycopersicum* and from *S. pennellii*, *S. habrochaites*, or both. We calculated the size differences between the amplified products cut with the appropriate restriction enzymes when needed or uncut in the case of indels. From the 491 primer pairs, 6 pairs amplified fragments that were smaller than expected and were discarded. Another 48 pairs failed to amplify fragments in at least one species, leaving us with 437 primer pairs that would test 281 polymorphisms between *S. lycopersicum* and *S. habrochaites* and 228 polymorphisms between *S. lycopersicum* and *S. pennellii*. Among these, 30 primers pairs yielded bands that were bigger than expected, probably due to the existence of non-conserved introns, nevertheless 23 of these were polymorphic as expected after restriction enzyme digestion. For the remaining amplifications we found 87% (220 of 261 LxH and 193 of 210 LxP) of the molecular markers to work as predicted. It is worth noting that 10% of the successful markers showed bands corresponding to both alleles in at least one of the species tested, suggesting heterozygosity in the lines used or the amplification of fragments from a family of genes sharing that sequence. Markers that failed could be due to undetected SNPs in

| Analysis | Unigene | Annotation | p-value | Sites | d/ds | d/ds^1 | d/ds^2 | lnL | lnL2 |
|----------|---------|------------|---------|-------|------|--------|--------|-----|------|
| LxH      | SGN-U314303 Aldehyde dehydrogenase | 0.00132 | 491   | 0.0665 | 0343 | 998.99 | -3184.6 | -3173.1 |
| LxH      | SGN-U317105 Unknown protein | 0.00132 | 286   | 0.2677 | 0058 | 145.99 | -1821.6 | -1810.1 |
| LxH      | SGN-U315058 Tropinone reductase | 0.00211 | 275   | 0.1361 | 0836 | 998.99 | -1902.2 | -1891.5 |
| LxH      | SGN-U315796 Alcohol dehydrogenase | 0.000276 | 399   | 0.0924 | 0036 | 3.34   | -1058.9 | -1049.3 |
| LxH      | SGN-U317449 Fe(II)/ascorbate oxidase | 0.000276 | 441   | 0.0569 | 0019 | 193.75 | -2443.7 | -2378.8 |
| LxH      | SGN-U317668 Hypothetical protein | 0.000276 | 253   | 0.2771 | 0053 | 3.34   | -1058.9 | -1049.1 |
| LxH      | SGN-U319889 Putative kinesin light chain | 0.000417 | 431   | 0.1566 | 0012 | 121.86 | -1966.2 | -1957.8 |
| LxH      | SGN-U314953 Unknown protein | 0.01001 | 176   | 0.2677 | 0058 | 0.52   | -2158.9 | -2151.8 |
| LxH      | SGN-U318509 Hypothetical protein | 0.03006 | 547   | 0.1026 | 0017 | 0.52   | -2158.9 | -2151.8 |
| LxH      | SGN-U36796 Hypothetical protein | 0.00276 | 253   | 0.2771 | 0053 | 3.34   | -1058.9 | -1049.1 |
| LxH      | SGN-U317449 Ubiquitin extension protein 2/605 ribosomal protein | 0.000099 | 155   | 0.0580 | 0017 | 998.99 | -512.5  | -500.2 |

LxP – *S. lycopersicum* and *S. pennellii* analysis, LxH – *S. lycopersicum* and *S. habrochaites* analysis. 1Annotation – based on the annotation from the best BLAST hit in Arabidopsis thaliana. 2p-value – Multiple testing corrected p-values for the differences in the maximum likelihood estimates of the two models fit. 3Sites – number of codons analyzed. 4d/ds – d/ds for the model where one d/ds ratio is estimated for the entire evolutionary tree. 5d/ds^1 – d/ds for the model with one d/ds ratio. 6d/ds^2 – d/ds for the model with two d/ds ratios. The very high d/ds ratios observed for some genes are a consequence of relatively little diversity between wild and domesticated tomato and are likely to be an overestimate of the true rate. 7lnL – likelihood estimate for the model with one d/ds ratio. 8lnL2 – likelihood estimate for the model with two d/ds ratios.
the sequence that modify the predicted restriction sites, or errors in SNP prediction. All tested markers are available in the Additional file 5.

Conclusion
We report in this work more than forty-nine thousand inter and intraspecific polymorphisms mined from the EST databases of the cultivated and two wild species of tomato. By taking advantage of the additional information linked to each sequence we were able to more accurately estimate the quality and the position of each SNP with respect to the coding region, which allowed us to distinguish those polymorphisms more likely to have phenotypic effects. Comparison of the sequences to homologs from better characterized species also allowed us to functionally classify the predicted unigenes and perform gene evolution analysis and tests for positive selection. We were able to suggest candidate genes and gene families that may be related to domestication of the cultivated tomato and/or environmental adaptation of wild species, providing hypotheses for more involved evolutionary studies. The information obtained was also used to design a set of markers that we make available to the community via a website. To our knowledge this is the first time that substantial meta-information including quality values, open reading frame predictions and homology to genes in Arabidopsis thaliana, has been used on tomato sequences to perform a pre-genomic analysis of gene variability and evolution.

Methods
Sequence data
Fasta files containing version 200607 build 1 of the unigene sequences, EST sequences, EST qualities and estimated coding region for each unigene were downloaded from SGN. ESTs in this database originated from the sequencing of at least 43 S. lycopersicum cDNA libraries belonging to at least two different accessions, 2 S. pennellii and 2 S. habrochaites cDNA libraries plus some individual cDNAs. Unigenes are the consensus sequences of these ESTs assembled with cap3 software as described in [17]. CDSs for the unigenes were calculated by SGN with ESTs-can software [29,30]. This software returns an optimum open reading frame based on Markov models and the nucleotide usage found in coding regions. Lukas Mueller at SGN kindly generated a custom list of all the unigenes and their constituent ESTs. The highest BLAST hit of every unigene versus the Arabidopsis thaliana genome was bulk queried and downloaded from SGN. Arabidopsis thaliana sequences were downloaded from TAIR (version 20080412, http://www.arabidopsis.org)

Assembly construction
We mined intraspecific SNPs in S. lycopersicum (L) ESTs and interspecific SNPs between S. lycopersicum and S. habrochaites (LxH) and between S. lycopersicum and S. pennellii (LxP) ESTs. Intraspecific analysis of S. pennellii and S. habrochaites ESTs were not performed as the EST libraries for those species were developed from a single accession.

For each unigene we used cap3 with relaxed parameters (-p 66 -b 99 -e 200) to assemble the EST sequences from the species participating in each of the analyses into analysis-specific contigs. Using these assemblies we calculated the sum and average of the qualities for each nucleotide call at every position and obtained analysis-specific consensus sequences. These consensus sequences were aligned to the original unigene sequence and the predicted CDS to estimate the beginning and end of the coding region. Every step in this process was carried out and verified using custom Perl scripts.

Perl and R [40] scripts were developed to estimate the quality score thresholds for sequence inclusion in SNP discovery. Once a threshold was determined we exclusively considered the positions in the assemblies above the threshold as follows. For interspecific analyses we required the sum of all the qualities of the nucleotides from each species to be over the threshold. For the intraspecific analysis we developed an algorithm that partitions all qualities at a given position into two groups with the minimum difference of the sums. We then considered only those positions in the assemblies at which the sum of the qualities of both groups was over the threshold. The threshold was defined as the quality score that maximizes the average sequence quality calculated separately for each species (interspecific analyses), or for each group of qualities (intraspecific analysis) taking into account only those positions over the threshold (Figure 2). We determined a quality threshold score of 50 for the intraspecific S. lycopersicum analysis and of 40 for both interspecific analyses (Figure 2).

SNP discovery
We defined SNPs in the intraspecific analysis as any assembly position where two and only two different bases were registered and for which the sum of qualities for each nucleotide call was over the quality threshold imposed for the analysis. For the interspecific analysis we considered SNPs whose positions within the assemblies presented a single and different nucleotide call for each species, and whose sum of qualities was greater than the imposed quality threshold. For quantification, SNPs that were consecutive in the assemblies were counted as a single polymorphism. SNP rates were calculated as the number of non-contiguous SNPs per 100 bases.

Amino acid translation, SNP codon position and transition/transversion ratios were evaluated with custom Perl and R scripts that analyzed the EST assemblies, the unigene sequence and the predicted CDS sequences.
Differential representation of SNPs in GO terms
Each nucleotide in the EST assemblies was assigned one or more GO categories based on the terms from the homologous Arabidopsis thaliana locus. To increase the accuracy of the test, we used only the parts of the assemblies that corresponded to coding regions and only those SNPs that has been predicted to produce amino acid changes. GO categories text files (version 20080712) were downloaded from TAIR [34]. For the interspecific analyses, we pooled the nucleotide calls for both (LxH and LxP) analyses. For duplicate loci we removed the locus with the shortest sequence. R scripts were developed to calculate over and under-representation of SNPs in nucleotide pools of each GO category versus all SNPs/nucleotides detected in the analysis using Fisher’s exact test.

Tests for selection
We performed McDonald-Kreitman tests and estimated Maximum likelihood from codon-substitution models on those unigenes that contained S. lycopersicum together with S. pennellii or S. habrochaites sequences. For the McDonald-Kreitman test we built fasta files for each unigene with the estimated CDS for the wild species alleles and two S. lycopersicum alleles differing in the SNPs identified in the L analysis. To maintain the fidelity of the analysis, those positions that did not pass the quality threshold using the methods described above were substituted with the ‘unknown’ character and not considered in the subsequent analysis. The number of synonymous and non-synonymous substitutions and p-values were calculated using the previously described MK.pl Perl script [41].

Two maximum likelihood codon-substitution models were fit to test the hypothesis of existence of positive selection in each unigene [27,28]. First, we fit the null model with a single d_0/d_2 ratio with equal ratios in every branch. The second model allowed for two d_0/d_2 ratios: one for the S. lycopersicum lineage and one for the rest of the tree. Then, a likelihood ratio test of the hypothesis of two branch rates was calculated by comparing the likelihood values from both models as in [42]. To do this we created Perl scripts that used ClustalW [43] to align the cultivated and wild species predicted CDS and the homologous Arabidopsis thaliana coding sequences. Alignments whose sum of qualities were not over the imposed quality threshold were substituted with the ‘unknown’ character. We also removed the parts of the alignments that lacked sequences from any of the three species. We constructed a phylogenetic tree of the three species and used PAML (v 3.14 [44]) to calculate the maximum likelihood of the models.

The resulting p-values from these experiments were corrected for multiple testing using the Benjamini and Hochberg algorithm in the Bioconductor multtest package [37,45].

Molecular marker design
For each polymorphic unigene in the interspecific analysis we aligned its predicted protein sequence with the protein sequence of its Arabidopsis best BLAST hit using standalone BLAST and custom Perl scripts. We calculated intron positions in the unigene based on intron positions in the Arabidopsis CDS. For each SNP we designed primers using Primer3 [46] with the unigene sequences as input, adjusting the program to design the primers within the predicted exon where the SNP was located.

We used Bioperl to generate virtual PCR fragments for each SNP allele based on the primers designed, and to find restriction endonucleases that would differentially cut the fragments, thus creating molecular markers. A set of these molecular markers was tested on genomic DNA from S. lycopersicum VF36, S. pennellii LA716 and S. habrochaites LA1347. Touchdown PCR was performed in a MJ Research PTC-200 Thermocycler with a starting annealing temperature of 58°C, which decreased 0.5°C per cycle for 15 cycles and stayed constant at 55°C for 30 cycles. Extension time was 40 seconds and denaturizing steps were performed for 30 seconds at 96°C. PCR products were digested with the appropriate restriction enzymes to detect the polymorphisms. We developed a database and a website holding the molecular marker information for each interspecific SNP.

All R and Perl scripts are available by request.

Abbreviations
CDS: Coding sequence; COS: Conserved orthologous set; EST: Expressed sequence tag; GO: Gene ontology; L: S. lycopersicum intraspecific analysis; LxH: S. lycopersicum versus S. habrochaites interspecific analysis; LxP: S. lycopersicum versus S. pennellii interspecific analysis; ORF: Open reading frame; PCR: Polymerase chain reaction; RFLP: Restriction fragment length polymorphism; SGN: Solanaceae genomics network; SNP: Single nucleotide polymorphism; UTR: Untranslated regions.

Authors’ contributions
JMJG wrote the manuscript, conceived and designed the study and the website hosting the results. JMJG also performed the assemblies, SNP mining, evolutionary analysis and molecular marker design and essays. JNM conceived and participated in the design of the study, and helped to draft the manuscript. All authors read and approved the final manuscript.
Acknowledgements

We thank Lukas Mueller for his very helpful information on the SGN datasets and for generating the list relating unigene and ESTs names. We thank Katherine S. Pollard for her advice on the evolutionary analyses and Dan Koenig and Matt Jones for helpful comments on the manuscript. Thanks to the PerlMonks for invaluable code and educational discussions on partition algorithms. This work was supported by NSF GPR grants DBI-0227103 and DBI-0820854.

References

1. FAQOSTT [http://faostat.fao.org]
2. International Solanaceae Genomics Project (SOL): Systems Approach to Diversity and Adaptation [http://www.sgn.cornell.edu/solanaceae-project/index.pl]
3. Fooal MR: Genome mapping and molecular breeding of tomato. Int J Plant Genomics 2007, 2007:64358.
4. Zamir D: Improving plant breeding with exotic genetic liber-
5. International Tomato Sequencing Project Home [http://www.sgn.cornell.edu/about/tomato_sequencing.pl]
6. GeneChip® Tomato Genome Array [http://www.affymetrix.com/products_services/arrays/specific/tomato.affx]
7. Tomato Functional Genomic Database [http://red.tbi.cornell.edu/]
8. Barone A, Chiussano ML, Ercolano MR, Giuliano G, Grandillo S, Fruscian L: Structural and functional genomics of tomato. Int J Plant Genomics 2008, 2008:820274-820274.
9. Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S: Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. TAG Theoretical and applied genetics 2005, 111(2):291-312.
10. Arshchenkova T, Ganal MW: Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. Theor Appl Genet 2002, 104(2-3):229-235.
11. Bai Y, Feng X, Hulst R van der, Lindhout P: A set of simple PCR markers converted from sequence specific RFLP markers on tomato chromosomes 9 to 12. Molecular Breeding 2004, 13(3):281-287.
12. Jimenez-Gomez JM, Alonso-Blanco C, Borja A, Anastasio G, Angosto T, Lozano R, Martinez-Zapater JM: Quantitative genetic analysis of flowering time in tomato. Genome 2007, 50(3):303-315.
13. Rafalski A: Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 2002, 5(2):94-100.
14. Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D: Expressed sequence tags from the laboratory-grown miniature tomato (Lycopersicon esculentum) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cul-
15. Yang W, Bai X, Kabelka E, Eaton C, Kamoun S, Knaap E van den, David F, David F: Discovery of single nucleotide polymorphisms in Lycopersicon esculentum by computer aided analysis of expressed sequence tags. Molecular Breeding 2004, 14(1):21-34.
16. Labate JA, Baldo A: Tomato SNP Discovery by EST Mining and Resequencing. Molecular Breeding 2005, 16(4):343-349.
17. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Birns J, Lin C, Wright MH, Ahrens R, Wang Y, et al: The SOL Genomics Net-
18. Fulton TM, Hoeven R Van der, Eennatta NT, Tanksley SD: Identification, analysis, and utilization of conserved ortholog set markers for comparative genetics in higher plants. Plant Cell 2002, 14(7):1457-1467.
19. Wu F, Mueller LA, Crouzillat D, Periard V, Tanksley SD: Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single Copy, Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade. Genetics 2006, 174(3):1407-1420.
20. Bauer P, Thiel T, Klatte M, Brezczky Z, Brumbachova T, Hell R, Grosse I: Analysis of Sequence, Map Position, and Gene Expression
Reveals Conserved Essential Genes for Iron Uptake in Arabidopsis and Tomato. Plant Physiol 2004, 136(4):4169-4183.

21. Beraldi D, Picarella ME, Sorettsi GP, Mazzucato A: Fine mapping of the parthenocarpic fruit (pat) mutation in tomato. TAG Theoretical and applied genetics 2004, 108(2):209-216.

22. Nakitandwe J, Trognitz F, Trognitz B: Reliable allele detection using SNP-based PCR primers containing Locked Nucleic Acid: application in genetic mapping. Plant Methods 2007, 3:2-2.

23. Van Deunze AE, Stoffel K, Buell RC, Kozik A, Liu J, Knaap E van der, Francis D: Diversity in conserved genes in tomato. BMC Genomics 2007, 8(1):465.

24. Ehrenreich IM, Purugganan MD: The molecular genetic basis of plant adaptation. Am J Bot 2006, 93(7):953-962.

25. Nesbitt TC, Tanksley SD: Comparative Sequencing in the Genus Lycopersicon: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes. Genetics 2002, 162(1):365-379.

26. McDonald JH, Kreitman M: Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991, 351(6328):652-654.

27. Yang Z, Swanson WJ: Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes. Mol Biol Evol 2002, 19(1):49-57.

28. Yang Z: Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998, 15(5):568-573.

29. Iseli C, Jongeneel CV, Bucher P: ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999:138-148.

30. Lottaz C, Iseli C, Jongeneel CV, Bucher P: Modeling sequencing errors by combining Hidden Markov models. Bioinformatics 2003, 19(Suppl 2):ii103-112.

31. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al.: Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 2007, 317(5836):338-342.

32. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: Whole-genome patterns of common DNA variation in three human populations. Science 2005, 307(5712):1072-1079.

33. Sullivan-Pollatschek S, Kashkush K, Shats H, Hillel J, Lavi U, Suliman-Pollatschek M, Gintzler AR, Garty M, Durand F: Estimation and mapping of AFLP, SSRs and SNPs in Lycopersicon esculentum. BMC Plant Biol 2008, 8(1):92.

34. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 1994, 22(22):4673-4680.

44. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997, 13(5):555-556.

45. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Bio 2004, 5(10):R80.

46. Rozen S, Skalaktsky H: Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 2000, 132:365-386.