Molecular Phylogenetics and Evolution 126 (2018) 293–302

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Comparative transcriptomics and genomic patterns of discordance in Capsiceae (Solanaceae)

Daniel Spalink[a,*], Kevin Stoffel[b], Genevieve K. Walden[a,d], Amanda M. Hulse-Kemp[b,c], Theresa A. Hill[b], Allen Van Deynze[b], Lynn Bohs[a]

[a] Department of Biology, University of Utah, Salt Lake City, UT, USA
[b] Department of Plant Sciences, University of California, Davis, CA, USA
[c] USDA-ARS, Genomics and Bioinformatics Research Unit, Raleigh, NC, USA
[d] Plant Pest Diagnostics Center, California Department of Food and Agriculture, 3294 Meadowview Road, Sacramento, CA 95832-1448 USA

ARTICLE INFO

Keywords:
BUCKy
Capsicum
Discordance
Genome structure
Lycianthes
SNaQ

ABSTRACT

The integration of genomics and phylogenetics allows new insight into the structure of gene tree discordance, the relationships among gene position, gene history, and rate of evolution, as well as the correspondence of gene function, positive selection, and gene ontology enrichment across lineages. We explore these issues using the tribe Capsiceae (Solanaceae), which is comprised of the genera *Lycianthes* and *Capsicum* (peppers). In combining the annotated genomes of *Capsicum* with newly sequenced transcriptomes of four species of *Lycianthes* and *Capsicum*, we develop phylogenies for 6747 genes, and construct a backbone species tree using both concordance and explicit phylogenetic network approaches. We quantify phylogenetic discordance among individual gene trees, measure their rates of synonymous and nonsynonymous substitution, and test whether they were positively selected along any branch of the phylogeny. We then map these genes onto the annotated *Capsicum* genome and test whether rates of evolution, gene history, and gene ontology vary significantly with gene position. We observed substantial discordance among gene trees. A bifurcating species tree placing *Capsicum* within a paraphyletic *Lycianthes* was supported over all phylogenetic networks. Rates of synonymous and nonsynonymous substitution varied 41-fold and 130-fold among genes, respectively, and were significantly lower in pericentromeric regions. We found that results of concordance tree analyses vary depending on the subset of genes used, and that genes within the pericentromeric regions only capture a portion of the observed discordance. We identified 787 genes that have been positively selected throughout the diversification history of Capsiceae, and discuss the importance of these genes as targets for investigation of economically important traits in the domesticated peppers.

1. Introduction

The field of phylogenetics has been revolutionized in this era of genomics and bioinformatics, both in terms of the numbers of markers being used for tree reconstruction, and more fundamentally, the kinds of questions that can be addressed with information from hundreds to thousands of genes. These include, for example, measuring discordance among gene trees, determining the causes of gene discordance, testing whether evolutionary histories are more accurately characterized as trees or networks, measuring rate heterogeneity amongst lineages and loci, and identifying lineage-specific instances of positive or negative selection (Stephens et al., 2015; Scornavacca and Galtier, 2016; Solís-Lemus and Ané, 2016; Sanderson et al., 2017). By integrating these analyses with mapped and annotated genomes, we can now push these investigations further and ask: How do gene histories vary with gene position on chromosomes? How do substitution rates vary according to gene position? Do gene functions vary with genealogy? And, do genes selected along a lineage collectively represent an enrichment of specific gene ontologies?

In this study, we target this synthesis of genomics with phylogenetics using the tribe Capsiceae, which is comprised of *Capsicum* L. and *Lycianthes* Hassl. (Dunal) (Solanaceae; Olmstead et al., 2008). The pepper genus *Capsicum* forms the foundation of a multibillion-dollar industry, has transformed global culinary cultures for hundreds of years, and serves as a model system for ecological, genomic, and developmental evolution (Haak et al., 2012). Transcriptome and genome sequences and abundant genetic resources are available for multiple species in the genus, reflecting its global importance (Kim et al., 2008;

* Corresponding author.
E-mail address: D.Spalink@utah.edu (D. Spalink).

https://doi.org/10.1016/j.ympev.2018.04.030
Received 10 November 2017; Received in revised form 20 April 2018; Accepted 20 April 2018
Available online 25 April 2018

Ashrafi et al., 2012; Góngora-Castillo et al., 2012; Park et al., 2012; Liu et al., 2013; Kim et al., 2014; Qin et al., 2014; Fernandez-Pozo et al., 2015; Hill et al., 2015). *Capsicum* is one of only a handful of genera of land plants for which full genomes have been assembled and annotated for more than one species. Given this abundance of genomic resources and the cultural, scientific, and economic importance of *Capsicum*, it is perhaps surprising that very little is known about the evolutionary diversification of *Capsicum* beyond the few species that have been cultivated, or within the broader context of Solanaceae. Indeed, the relationship between *Capsicum* and *Lycianthes* remains unresolved, with phylogenetic evidence suggesting that *Capsicum* may be embedded within *Lycianthes* (Särkinen et al., 2013). Thus, even the taxonomic status of *Capsicum* and its relatives is problematic. Resolving these relationships has far-reaching implications for understanding the diversification of the clade and the evolution and ecology of pungency, and will provide resources for improving crop diversity and agricultural sustainability (Albert and Chang, 2014; Brozynska et al., 2015; Fernandez-Pozo et al., 2015).

Capsicum and *Lycianthes* have remained taxonomically separated because of their geographical, ecological, and morphological distinctiveness. With ca. 35 species of annual and perennial herbs and shrubs, *Capsicum* (Fig. 1a) is native only to Mexico and Central and South America, with the exception of *Capsicum annuum* var. *glabriusculum,* which also extends into the USA. The capsaicinoids that result in fruit pungency are expressed to varying degrees in all but six species, though only five species have been domesticated (Carrizo García et al., 2016). In comparison, pungency is lacking altogether in *Lycianthes* (Fig. 1b). With ca. 250 species, *Lycianthes* is one of the largest genera in Solanaceae, occurring from Mexico to South America, Asia and the South Pacific. *Lycianthes* exhibits a diversity of habits including small trees, vines, shrubs, herbs, and epiphytes. Several *Lycianthes* species produce edible berries and are grown as ornamental shrubs (Nee, 1991; Williams, 1993; Barboza, 2013). *Capsicum* flowers have stamens that dehisce by longitudinal slits, and offer an energy-intensive nectar reward to primarily bee pollinators (Kristjansson and Rasmussen, 1991; Jarlan et al., 1997; Meisels and Chiasson, 1997; Raw, 2000; Cauich et al., 2015). *Lycianthes* stamens, on the other hand, dehisce by terminal pores and are buzz-pollinated exclusively by bees, which receive pollen as their only reward (Dean, 2004). Many *Lycianthes* species are nocturnal or crepuscular bloomers, which may offer a competitive advantage by allowing access to a wider diversity of pollinators (Smith and Knapp, 2002).

Attempts to resolve the Capsiceae phylogeny have relied exclusively on a few plastid and nuclear ribosomal genes, or at most, these in combination with a single low-copy nuclear gene (Walsh and Hoot, 2001; Särkinen et al., 2013; Carrizo García et al., 2016). Support for the paraphyly of *Lycianthes* with respect to *Capsicum* is weak in these published studies, suggesting that additional data are necessary to more fully understand the evolutionary processes involved in the diversification of this clade. Several such processes could be involved. For example, rapid diversification along the backbone of the phylogeny could simply require more informative characters to provide better resolution and support. In this case, the paraphyly of *Lycianthes* may indeed reflect the true history of these lineages. On the other hand, rapid diversification could also result in the incomplete sorting of individual genes, in which case the gene trees obtained through the limited sequence data available to date might not reflect the dominant species topology. Further still, the concatenation methods used in prior analyses could obscure instances of past hybridization and introgression events, resulting in a bifurcating tree that oversimplifies the relationships among the two genera. In any of these scenarios, an improved phylogenetic hypothesis is critical to identify genes that have been selected through *Capsicum* and *Lycianthes* evolution, to polarize the reconstruction and quantify the divergence of key traits through time, and to update the taxonomy of *Capsicum* and *Lycianthes.*

Here we present four new transcriptomes from species spanning major clades of *Lycianthes* and the non-pungent *Capsicum* clade to broaden the genomic resources available for the tribe. Our objectives are three-fold. First, by incorporating genomic data from the pungent species of *Capsicum*, we construct gene trees from nearly 7000 loci to calculate the extent of discordance amongst genes, identify the causes of this discordance, and estimate the Capsiceae backbone species tree. Second, given the global importance of this clade, identifying genomic patterns in the divergence of *Capsicum* and *Lycianthes* could yield valuable insights into the evolution of culturally or economically significant traits. To this end, we measured rates of synonymous and nonsynonymous nucleotide substitution in these genes and identified which ones may have experienced positive selection during the diversification of the clade. Third, as recombination rates are suppressed in the pericentromeric regions in *Capsicum* (Qin et al., 2014; Hill et al., 2015), we expect these genes to evolve more slowly than those towards the chromosome ends. For the same reason, we expect to find less discordance among genes close to the chromosome centers compared to those at the chromosome ends. To explore these issues, we incorporated an annotated genome of *Capsicum annuum* and tested the hypothesis that the diversity of gene histories should be greater towards

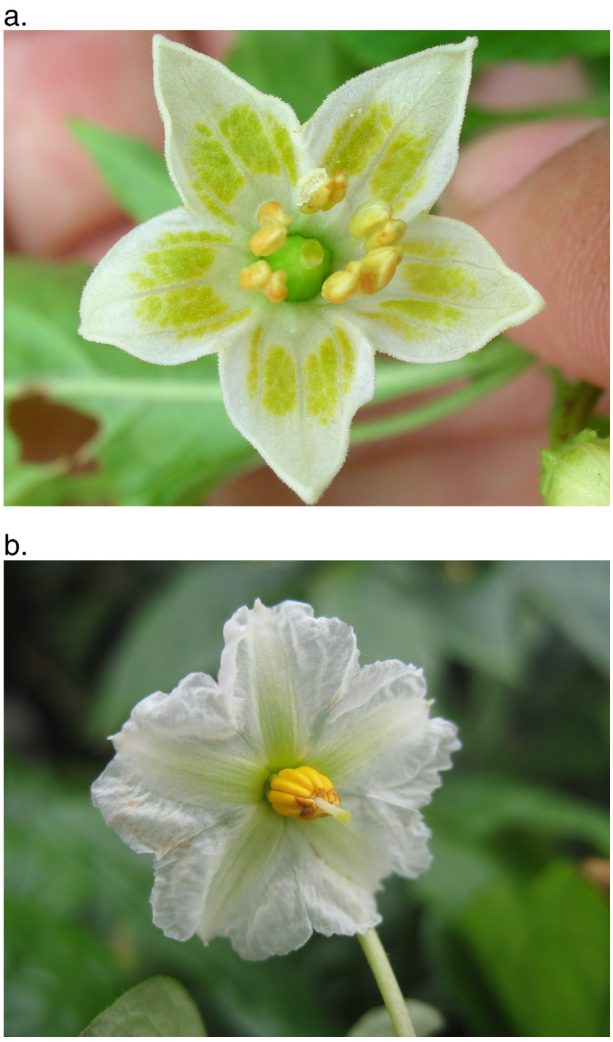

Fig. 1. Flowers of *Capsicum baccatum* (Photo Credit: G. Barboza) (a) and *Lycianthes asarifola* (b). All species of *Capsicum* have stamens with longitudinally dehiscing anthers, whereas all species of *Lycianthes* release pollen through terminal pores. *Capsicum* species also present a nectar reward to pollinators, which is evident on the lowermost petal in 1a. The buzz pollinated flowers of *Lycianthes* present pollen as the only reward.

chromosome ends than pericentromeric regions. Ultimately, we present data to expand our understanding on the evolution within Capsiceae and provide foundational genomic resources for more targeted phylogenetic studies within this dynamic clade.

2. Materials and methods

2.1. Sampling and library construction

Three species from the major clades of *Lycianthes* (Särkinen et al., 2013) and the non-pungent clade of *Capsicum* (Carrizo García et al., 2016) were selected for transcriptome sequencing. Fresh leaf, root, flower, and flower bud tissues were collected and frozen in liquid nitrogen from University of Utah greenhouse individuals of *L. saltensis* Bitter, *L. biflora* (Lour.) Bitter, and *C. rhomboideum* (Dunal) Kuntze, and from leaves, roots, and rhizomes from a UC-Davis individual of *L. asarifolia* (Kunth & C.D. Bouché) Bitter. Preserved specimens of these plants are housed at UC-Davis (DAV) and the Garrett Herbarium (UT; SI Table S1). RNA extractions were conducted with Qiagen RNeasy Plant Mini Kits (Qiagen, Valencia, CA) on 100 mg of tissue pooled from all sampled organs, and yields were quantified using a Qubit 2.0 Fluorimeter (Invitrogen, Carlsbad, CA). Libraries were constructed from cDNA following Zhong et al. (2011). All four paired-end libraries were sequenced on a single Illumina HiSeq 4000 (Illumina, San Diego, CA) lane at UC Davis.

2.2. Transcriptome assembly

Paired-end reads were trimmed of adapter sequences, quality filtered, and assembled *de novo* using the commercially available CLC Genomics Workbench 8.5.1 (Qiagen, Valencia, CA). We tested the importance of three parameters on the quality of assemblies. First, we varied the number of nucleotides trimmed from the 5′ end of the sequence from 13 to 15 nucleotides. Second, we filtered reads based on quality, using thresholds of 0.05, 0.02, and 0.01. Third, we altered the k-mer size used for *de-Brujin* graph-based assembly from 20 to 50 nucleotides by 10-nucleotide intervals. For each transcriptome, we selected the combination of parameters that maximized the average N50 length. Plastome transcripts were assembled separately with the Map To Reference assembler of Geneious 9.1.7 (Kearse et al., 2012), using the complete *C. chinense* Jacq. plastome sequence (Park et al., 2016) as a reference. Individual plastome coding DNA sequences (CDS) were parsed to eliminate low-coverage intergenic spacer regions.

Gene ontology (GO) annotation of nuclear and plastid contigs. Contigs from the *de novo* assemblies were annotated using Blast2Go 4.0.7 (Conesa et al., 2005). A BLASTX search was conducted on a local non-redundant protein database constructed from all Solanaceae sequences on GenBank using an expectation value of 1.0E-25, which was followed by mapping and annotation of the GO terms and protein functions associated with the BLAST results. We conducted a second round of Blast2Go analyses using a local database constructed from the *C. annuum* cv. CM334 genome, a Mexican landrace hot pepper (Criollo de Morelos 334; Kim et al., 2014), to associate our *de novo* contigs with *Capsicum* gene names. Contigs that did not blast to a single *C. annuum* gene were eliminated from all downstream analyses. To eliminate potential paralogs, we also removed all contigs that blasted to more than one *C. annuum* gene with equally high expectation values. As gene content and synteny are unusually conserved throughout species in Solanaceae (Park et al., 2012; Hill et al., 2015), we also used this second round of BLAST results to estimate the positions of the *de novo* contigs onto the 12 chromosomes of *C. annuum*.

2.3. Phylogenetic analyses

We used gene tree, species tree, and phylogenetic network approaches to reconstruct the phylogeny of *Capsicum* and *Lycianthes*. We broadened our taxonomic sampling to include *C. frutescens* L. and *C. annuum* L., two pungent peppers with available genome or transcriptome sequences (Ashrafi et al., 2012; Park et al., 2012; Liu et al., 2013; Kim et al., 2014), and *Solanum tuberosum* L., which was used as an outgroup. For these analyses, we only considered genes for which sequence data were available for all seven taxa. GenBank data were downloaded using the {rentrez}package (Winter, 2017) in R (R Core Team, 2013), fetching only the top BLAST hit. Genes were initially aligned with MAFFT 7.305b (Katoh and Standley, 2013). All alignments were processed with trimAl 1.2rev59 (Capella-Gutierrez et al., 2009) to remove introns and eliminate entire alignments containing clearly mismatched or divergent sequences (e.g., < 50% match). This was done to ascertain comparisons of genes across all species. Alignments were visually inspected and manually adjusted in Geneious 9.1.7 to ensure that all genes were in reading frame.

Gene trees were constructed using MrBayes 3.2.4 (Ronquist et al., 2012). For each gene, we selected best fitting models of nucleotide substitution using Bayesian Information Criterion through jModelTest 2.1.4 (Darriba et al., 2012), which used four chains and two runs of 1,000,000 generations each in the Bayesian tree estimation, and removed 10,000 generations as burn-in period. Plastid genes were concatenated prior to the phylogenetic analyses. We then tallied which of the possible 945 rooted, bifurcating tree topologies was recovered by each gene, eliminating trees that were not fully resolved. Output from MrBayes was used for Bayesian concordance analysis using BUCKy 1.4.3 (Ane et al., 2006; Larget et al., 2010), a species-tree approach that determines the proportion of the genome for which a clade is supported. In the construction of this species tree, no assumptions are made regarding the cause of incongruence among loci, whether incomplete lineage sorting, hybridization, horizontal gene transfer, etc. This species tree can then be compared to a population tree, which is constructed using concordance factors and assumes that all incongruence results from incomplete lineage sorting. To account for the alternate possibility that incongruence among loci is the result of ancient hybridization events rather than incomplete lineage sorting, we developed an explicit phylogenetic network using the maximum pseudolikelihood approach SNaQ (Solís-Lemus and Ané, 2016), as implemented in the PhyloNetworks package in Julia 0.5 (Bezanson et al., 2017). We then conducted goodness-of-fit tests to determine whether phylogenetic networks with as many as three reticulations better explain gene discordance than the coalescent tree. We used the TICR pipeline (Stenz et al., 2015) along with the R package {phylolm} (Ho and Ane, 2014) to run BUCKy, prepare input files for SNaQ, and to conduct goodness-of-fit tests.

These analyses were based on every gene for which we could make unambiguous alignments. We were interested in determining whether gene position should be considered while designing phylogenetic analyses that incorporate numerous nuclear genes. As rates of recombination decrease near centromeres (Qin et al., 2014; Hill et al., 2015), we specifically tested whether randomizing the genes selected for the analysis captures genome-scale heterogeneity better than genes that are restricted to either pericentromeric zones or chromosome ends. We created two datasets constructed using the 5% of genes closest to and furthest from the centromere centers (hereafter referred to as “pericentromeric” and “paracentric,” respectively), and compared this to a distribution of 100 sets of randomly selected genes. Centromere locations were inferred by reference to a high-density genetic map of *C. annuum* (Qin et al., 2014). For both analyses, we identified all primary and secondary splits that were present in at least 5% of gene trees and recorded their concordance factors. We then constructed a null distribution of splits and concordance factors by analyzing 100 sets of 200 randomly selected genes. We tested whether the pericentromeric and paracentric datasets captured all the secondary splits of the random datasets, and whether concordance factors departed significantly from the null distributions.

2.4. Rates of evolution, positive selection, and gene ontology enrichment

We measured rates of substitution and identified instances of positive selection using a subset of genes, which were selected using three criteria: (i) genes are expressed in all seven species, (ii) no stop codons are present within the reading frame, and (iii) edited alignments are at least 500 nucleotides long. We then examined whether positively selected genes represented an enrichment of gene ontologies (GO), which would suggest that these subsets of genes express functional unity. We similarly tested for GO enrichment among all genes representing each observed evolutionary history to determine the extent to which specific gene functions have diverged independently of the species tree topology.

We used the codeml program in PAML 4.8 (Yang, 2007) to measure rates of synonymous (dS) and nonsynonymous (dN) substitution. We then tested whether substitution rates of genes varied categorically according to their associated gene tree topologies using analysis-of-variance (ANOVA) and Tukey-Kramer tests. Second, we tested the hypothesis that rates of substitution should increase with distance from centromeres. We measured the distance of each gene from the centromere centers and weighted these distances by the total length of the respective chromosomes. Third, we tested the hypothesis that phylogenies constructed from neighboring genes should become increasingly dissimilar with increased distance from centromeres, reflecting the higher rates of substitution expected in these regions. Both hypotheses were tested used Pearson's product-moment correlation analyses.

We used a time-calibrated chronogram developed using BEAST 2.4.0 (Bouckaert et al., 2014) to transform dS and dN to absolute rates of substitution in units of substitution site^{-1} year^{-1}. For efficiency, we used 26 genes for the BEAST analysis, with two genes randomly selected from each chromosome and two genes from unplaced scaffolds. To allow for rates of evolution to vary among lineages, we used a relaxed clock under a log-normal distribution and a Birth Death model. As no reliable fossils were available to provide minimum dates for this analysis, we used four secondary, normally distributed priors on nodes throughout the tree, with minimum dates reflecting those in a Solanaceae-wide chronogram (Särkinen et al., 2013). We placed a prior on the root node with a mean offset of 19.13 million years (my), a prior on the most recent common ancestor (MRCA) of *Capsicum* and *Lycianthes* with a mean offset of 13.23 my, a prior on the *Capsicum* stem node with a mean offset of 12.83 my, and a prior on the *Capsicum* crown with a mean offset of 9.8 my. We ran the MCMC for 100 million generations, assessed convergence and effective sampling using Tracer 1.6 [http://tree.bio.ed.ac.uk/software/tracer/], and eliminated 20% of generations as burn-in. We used the following formula to calculate substitution rates: r = d/2T, where d represents the synonymous or nonsynonymous distance between two species and T represents the age of their MRCA.

To identify genes with putative positive selection throughout the diversification of Capsiceae, we used the branch-site test of positive selection (Zhang, 2005) as implemented in codeml (Yang, 2007). This model allows the ratio of nonsynonymous to synonymous mutations (dN/dS, or ω) to vary across specified branches and among sites. We compared a null model, where ω is fixed and no positive selection can occur, to an alternative model, where positive selection can occur along specified branches in the phylogeny, and tested for significant improvement in the more complex model using likelihood ratio tests at the 1% significance level ($2\Delta l > 5.99$). We tested for positive selection on all branches of the ingroup phylogeny. Recent studies have indicated that the branch-site test of positive selection can lead to false positives, especially in genes where multiple mutations have occurred in a single codon (Venkat et al., 2017). Results from this analysis should therefore be viewed with caution.

We performed two analyses to identify instances of significant enrichment of GO terms throughout the diversification of Capsiceae. First, we tested whether positively selected genes on each branch of the phylogeny represented GO enrichment. Second, we parsed genes according to their evolutionary history, and tested whether these gene categories represented significant enrichment. For these analyses, we used the singular enrichment analysis (SEA) tool of AgriGO (http://bioinfo.cau.edu.cn/agriGO/), with *Solanum* as the reference background and significance assessed using Fischer's tests (Du et al., 2010).

3. Results

3.1. De novo transcriptome assembly

Illumina sequencing yielded ~97 M, 95.7 M, 88.9 M, and 74.7 million reads for *L. asarifolia, L. biflora, L. saltensis,* and *C. rhomboideum*, respectively. Raw reads are deposited in the NCBI Sequence Read Archive (SRP129039). For each transcriptome, the average length and N50 scores of contigs were maximized when 15 bp were trimmed from 5′ ends, nucleotides were filtered at a quality threshold of 0.02, and the assembly k-mer size was set to 40 nucleotides. Increasing the quality stringency from 0.05 to 0.02 or 0.01 resulted in the recovery of ~3000 fewer contigs, and contig length increased under the more stringent settings. As the quality threshold of 0.02 had the maximum average contig length, all subsequent analyses are based on the assembly under these parameters. Among the four species, we assembled an average of 78,317 contigs, which had an average length of 830 base pairs (bp) and N50 value of 1,353 bp. A total of 15,392 nuclear contigs were expressed in all four new transcriptomes (SI Fig. S1). Summary statistics for each species are presented in SI Fig. S2 and Table 1.

A total of ~0.32 million (M), 0.19 M, 0.11 M, and 0.10 M reads mapped to the *Capsicum chinense* plastome for *L. asarifolia, L. biflora, L. saltensis,* and *C. rhomboideum*, respectively. The average sequencing depth ranged from 63 to 289 reads per site. We recovered a total of 77 of 134 plastid genes without any plastome enrichment in the library preparation, with the alignment of concatenated exons totaling 59,448 bp.

3.2. Gene annotation

Among the contigs from the four transcriptomes sequenced in this study, between 48.7% (*L. biflora*) and 58.1% (*C. rhomboideum*) successfully mapped to genes in a local database of downloaded Solanaceae sequences (Table 2). Of these, over 60% were annotated with specific GO terms. On average, 3.3 GO terms were associated with each of these contigs, with some having as many as 31 distinct annotations. GO terms are classified under three primary domains: the functional activity of the associated protein (F), the biological process in which this function is involved (P), and the location in the cell where this process occurs (C) (Hill et al., 2008). The annotations associated with the successfully mapped contigs included 1380F, 1872P, and 409C unique GO terms (SI Tables S2–S4). Among all sampled genes, the most common protein function was ATP binding, the most common process was oxidation-reduction, and the most common locations were integral membrane components. A total of 24, 24, 154, and 120 GO terms were uniquely assigned to *C. rhomboideum*, *L. asarifolia, L. biflora,* and *L. saltensis*, respectively.

Table 1
Assembly statistics for each *de novo* assembly. L = *Lycianthes*, C = *Capsicum*.

Statistics	*L. asarifolia*	*L. biflora*	*L. saltensis*	*C. rhomboideum*
Total assembled nucleotides	64,551,915	83,421,976	64,800,004	63,169,143
Number of unigenes	83,668	77,529	81,598	70,337
Average GC content (%)	38.9	38.5	39.2	39.1
Minimum contig length	163	163	162	164
Maximum contig length	15,861	16,970	16,961	16,638
Average contig length	772	858	794	898
N50	1234	1442	1277	1464

Table 2
Gene Ontology (GO) summary statistics, showing the number of contigs assembled for each species, the number and percent of these contigs that had significant BLAST hits to other Solanaceae sequences, the number and percent of contigs annotated with GO terms, the maximum and average number of GO terms associated with contigs for each species, and the number of percentage of *de novo* contigs that were successfully mapped to the Mexican landrace *Capsicum annuum* CM334 (CA) genome.

Species	# Contigs	# Blast Hits	% Blast Hits	# with GO Terms	% with GO Terms	Max # GO Terms	Average # GO Terms	# CA Hits	% CA Hits
L. asarifolia	83,618	42,432	51	25,135	30	27	3.26	40,507	48
L. biflora	77,481	37,744	49	22,083	29	31	3.32	36,089	47
L. saltensis	81,909	44,860	55	27,643	34	23	3.20	41,616	51
C. rhomboideum	70,259	40,829	58	25,180	36	24	3.26	38,796	55
Average	78,317	41,466	53	25,010	32	26	3.26	39,252	50

3.3. Phylogenetics

Of the 15,392 genes expressed in all four transcriptomes, 11,219 produced significant BLAST hits with *C. annuum*, *C. frutescens,* and *S. tuberosum*. After quality control filtering and trimming of these genes, 6747 alignments representing 10,187,553 total bp were retained for phylogenetic analyses. The alignment of 77 concatenated plastid genes included an additional 59,448 bp. All alignments are available for download from Mendeley Data.

Consensus trees from the phylogenetic analyses of the nuclear genes supported 25 distinct topologies (Fig. 2). Of the 6747 genes analyzed, 969 produced trees that were not fully resolved. The most common topology was supported by 2349 genes and placed *L. biflora* sister to the remainder of *Lycianthes* and a monophyletic *Capsicum*, with *C.*

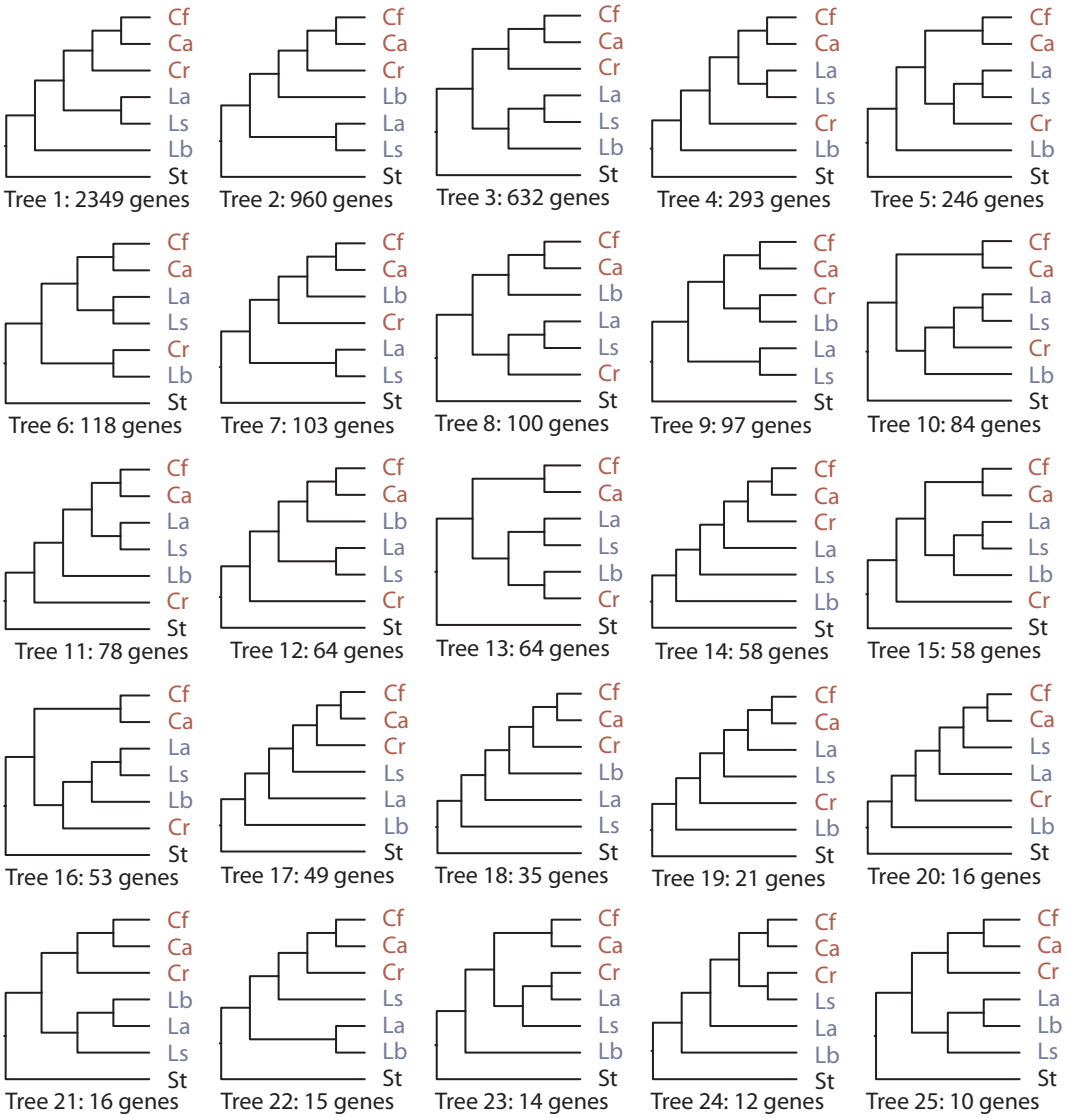

Fig. 2. Twenty-five distinct evolutionary histories recovered by the 5545 analyzed genes. The most common topology, which places *Capsicum* within a paraphyletic *Lycianthes*, is also the species tree. *St = Solanum tuberosum*, *La = Lycianthes asarifolia, Lb = Lycianthes biflora, Ls = Lycianthes saltensis, Ca = Capsicum annuum, Cf = Capsicum frutescens*, *Cr = Capsicum rhomboideum.*

rhomboideum sister to the pungent peppers (*C. annuum* and *C. frutescens*). The second most common topology, supported by 960 genes, also placed *Capsicum* within a paraphyletic *Lycianthes*, with *L. asarifolia* and *L. saltensis* forming a clade sister to all remaining species. The third most common topology, supported by 632 genes, resolved *Capsicum* and *Lycianthes* as reciprocally monophyletic, with *C. rhomboideum* sister to the pungent peppers and *L. biflora* sister to *L. asarifolia* + *L. saltensis*. The remaining topologies, supported by as many as 293 genes and as few as 10 genes, varied primarily in the placement of *L. biflora* and *C. rhomboideum*, rendering one, both, or neither genera monophyletic (Fig. 2). The plastome phylogeny was identical to Tree 3, supporting both genera as monophyletic.

The concordance and population BUCKy trees both supported the most common gene tree topology (Fig. 2, Tree 1; Fig. 3), with a monophyletic *Capsicum* and *L. biflora* sister to all ingroup taxa. All secondary splits involved the alternate placements of *L. biflora* and *C. rhomboideum*, with 19.3% of the genome supporting *L. biflora* as sister to *Capsicum* and 11% supporting reciprocally monophyletic genera. The identical concordance and population trees suggest that this incongruence is best attributed to incomplete lineage sorting. We tested this further by comparing the concordant species tree to an explicit phylogenetic network constructed using SNaQ, allowing up to three ancestral hybridization events. Goodness-of-fit tests rejected the phylogenetic networks as a significant improvement over the bifurcating species tree ($X^2 = 3.89$, $p = 0.274$).

We found that both the pericentromeric and paracentric datasets had higher concordance factors (CF) for the primary splits as compared to the distribution of randomized datasets (Fig. 4). Though all CFs were within the 95% confidence interval of the randomized datasets, pericentromeric genes had higher CFs for the second and third primary splits by more one standard deviation (Fig. 4a). Similar patterns were observed among secondary splits, where the pericentromeric and paracentric datasets tended to have higher or lower CFs as compared to the randomized distribution (Fig. 4b). One secondary split, which was supported by 8.1% of genes in the full analysis and an average of 8.9% of genes in the randomized datasets, was not observed in the pericentromeric dataset.

3.4. Molecular dating and rates of evolution

Our BEAST analysis achieved convergence and effective sampling sizes of over 200 for all parameters (SI Fig. S3). After filtering out alignments containing internal stop codons, we measured rates of synonymous and nonsynonymous substitution in a total of 5492 nuclear genes, which had a mean length of 1336 bp. Across all nuclear genes and pairwise lineage comparisons, average dS was 0.16 substitutions site^{-1}, while dN was 0.03 substitutions site^{-1}, indicating that neutral substitutions were more than four times as common as mutations leading to changes in amino acids (Fig. 3; Table 3a). Rates of substitution varied among nuclear genes, with dS ranging from 0.0180 to 0.7000 substitutions site^{-1} and dN ranging from 0.0003 to 0.1740 site^{-1} (Fig. 3). Substitution rates were generally much lower amongst the 77 plastid genes included in our analysis, with dS ranging from 0 to 0.114 substitutions site^{-1}, and dN ranging from 0 to 0.073 substitutions site^{-1}. We found no correlation between the topologies of gene trees and their rate of evolution. However, we found that 16.3% of the variance in dS is explained by the distance of genes from the centromere center ($p < 0.01$), with more paracentric genes having higher average rates of synonymous substitution. By comparison, only 4.3% of variance in dN was explained by gene position.

Incorporating the dates from the BEAST chronogram, we estimated an average of 7.41×10^{-9} synonymous, and 1.41×10^{-9} nonsynonymous, substitutions site^{-1} year^{-1} across all lineages (Table 3b). While dS and dN within the pungent *Capsicum* species were each more than an order of magnitude lower than the average across all lineages (0.014 and 0.0027 substitutions site^{-1}, respectively), the absolute rates of

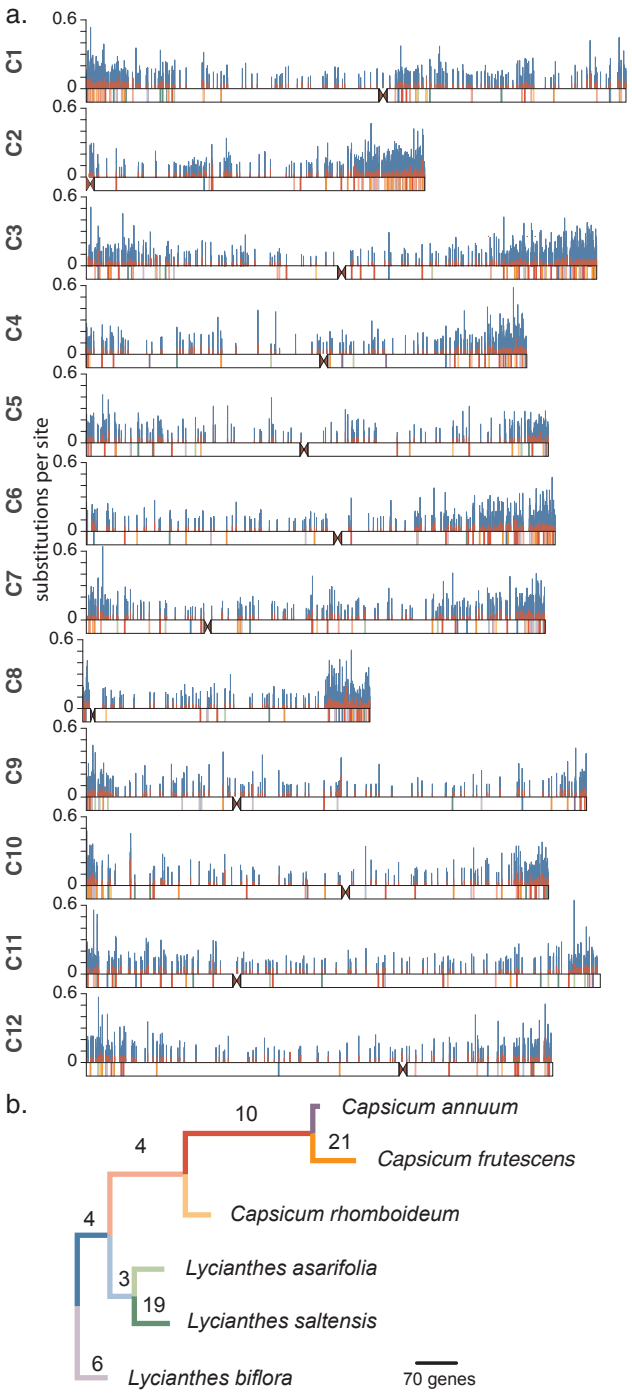

Fig. 3. Distribution of sampled genes across the twelve chromosomes of CM334 *Capsicum annuum* (C1-C12) and their associated rates of synonymous (dS) and nonsynonymous (dN) substitution. (A) Chromosomes are depicted by horizontal rectangles, with the position of the centromere indicated by the constriction. The histograms above each chromosome indicate dS (blue) and dN (red) for each measured gene. Genes that have been positively selected at some point through the diversification of Capsiceae are depicted by colored bars within the chromosomes. Colors correspond to the branch of the phylogeny in (B) on which the gene was positively selected. (B) Cladogram with branch lengths proportional to the number of genes identified as positively selected along each lineage. Branch labels indicate the number of GO terms that were significantly enriched among the positively selected genes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

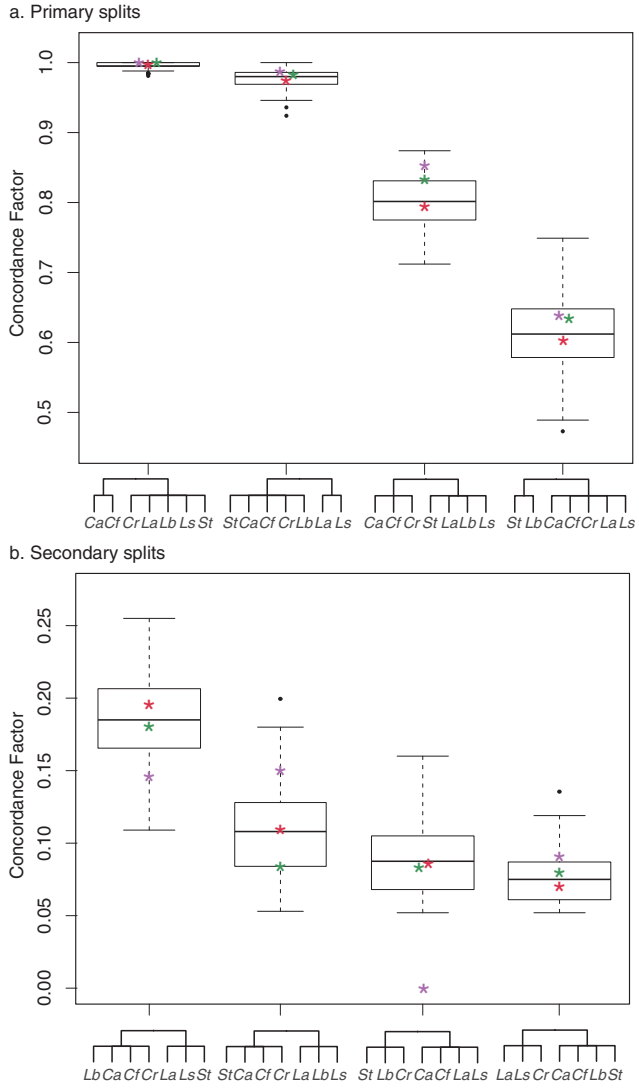

Fig. 4. Distributions of concordance factors associated with primary (A) and secondary (B) splits from randomly selected BUCKy datasets. Boxplots depict the distribution of concordance factors supporting each split among all randomized datasets. These are compared to the concordance factors calculated when using all available genes (red asterisk), the 5% of genes closest to the centromeres of the twelve chromosomes (purple asterisk), and the 5% of genes furthest from the centromeres (green asterisk). The pericentromeric genes generally have concordance factors that are more than one standard deviation away from the randomized datasets, and in one case failed to recover a secondary split. This pattern was less apparent with the paracentric dataset. In all cases, the total evidence analysis recovered concordance factors within one standard distribution of the averaged randomized datasets. Splits are visually represented with cladograms. *St = Solanum tuberosum*, *La = Lycianthes asarifolia*, *Lb = Lycianthes biflora*, *Ls = Lycianthes saltensis*, *Ca = Capsicum annuum*, *Cf = Capsicum frutescens*, *Cr = Capsicum rhomboideum*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

synonymous and nonsynonymous substitution since their estimated divergence 0.11 Ma were the fastest amongst all lineages (Table 3).

Out of 5492 genes tested, we identified 787 genes that exhibited significant positive selection on at least one branch of the ingroup (Fig. 3). GO terms associated with these genes included 268 function (F), 352 process (P), and 95 cellular component (C) annotations. The number of selected genes varied among branches from 17 genes (stem branch of *C. annuum*) to 297 genes (stem lineage of pungent *Capsicum*), with 179 genes selected on more than one branch of the phylogeny. Of the 787 genes exhibiting positive selection, fifteen have been previously identified as having some importance in domestication. Of these, six are directly involved in the biosynthesis or regulation of capsaicin (CA01g17970, CA02g27850, CA04g10340, CA09g14690, CA10g15980, and CA10g20920), three are involved in meristem branching (CA02g04650, CA03g30380, and CA10g05080), five are involved in the regulation of organ growth (CA00g84260, CA03g28640, CA05g17330, CA08g03060, and CA12g02850), and one is involved in gibberellin signaling (CA12g02780).

We observed significant enrichment of GO terms in several analyses. First, we found that positively selected genes represented GO term enrichment on all branches of the phylogeny (Fig. 3). Second, we found that genes representing nineteen of twenty-five of the observed evolutionary histories also represented GO term enrichment (Fig. 2). For example, GO terms involving cis-trans isomerase and aldehyde and ketone group transferase activities, enzymes associated with the fatty acid and vanillin pathways, were overrepresented in the genes that show *Capsicum* and *Lycianthes* as reciprocally monophyletic genera.

4. Discussion

4.1. The backbone Capsiceae phylogeny

We developed a strongly supported Capsiceae backbone species tree, with a topology that is consistent with results from previous studies (Särkinen et al., 2013; Carrizo García et al., 2016). For a dataset with only seven taxa, however, we observed high rates of incongruence among gene trees, which represented 25 distinct and strongly supported evolutionary histories (Fig. 2). We compared two possible explanations for this pattern, and found that incomplete lineage sorting (ILS) explains the incongruence between the species tree and the remaining 24 genealogies significantly better than horizontal gene transfer resulting from ancestral introgressive hybridizations (Solís-Lemus and Ané, 2016). As other recent genome-level phylogenetic analyses have found that hidden alignment errors are responsible for more discordance than ILS (e.g., Scornavacca and Galtier, 2016), we verified our results after checking all alignments manually. Our molecular dating analysis is consistent with what would be expected if ILS was responsible for gene tree discordance, indicating that *L. biflora,* the remaining *Lycianthes* clade, *C. rhomboideum,* and pungent *Capsicum* lineages all diverged within 3 million years (SI Fig. S3). This rapid succession of cladogenetic events could have prevented the fixation of lineage-specific alleles and resulted in different genes depicting distinct histories (Takahashi et al., 2001; Maddison and Knowles, 2006; Chung and Ané, 2011). Indeed, the primary location of incongruence among genes is along this backbone of the phylogeny. Whereas the *Lycianthes asarifolia* + *L. saltensis* and pungent *Capsicum* clades are consistently supported by nearly all genes, the relative position of these clades in relation to *C. rhomboideum* and *L. biflora* is far more variable. We suspect that total genomic levels of incongruence are likely higher than we were able to capture using these analyses. The tissues that were sampled for transcriptome sequencing varied among the four species, and the plants themselves were raised under different environments and sampled at different growth stages. As these analyses only incorporated genes that were expressed in all four species, which likely excluded highly divergent genes.

Our genome-level support for the paraphyly of *Lycianthes* in relation to *Capsicum* has important implications for the taxonomy of Capsiceae. As *Capsicum* takes nomenclatural priority over *Lycianthes*, we expect the circumscription of *Capsicum* to either expand, and/or *Lycianthes* to be split into segregate genera if monophyly is to serve as a foundation for taxonomy. Our dataset provides an opportunity, however, to serve as a basis for understanding the genetics of the morphological traits that have been historically used to unify *Lycianthes* as a genus. These traits include, for example, the poricidal anthers and absence of nectaries common to all *Lycianthes* (Fig. 1). *Capsicum*, as well as most of the close relatives of Capsiceae (e.g., *Witheringia, Physalis, Iochroma, Salpichroa,*

Table 3
Average rates of substitution site^{-1} and substitution site^{-1} year^{-1}, calculated using pairwise comparisons of alignments of coding sequences. In each matrix, the upper right triangle depicts nonsynonymous substitutions (dN) and the lower left triangle depicts synonymous substitutions (dS). A. dN and dS of 5492 nuclear genes. B. Absolute rates of substitution of nuclear genes. C. dN and dS of plastid genes. D. Absolute rates of substitution of plastid genes.

	L. asarifolia	*L. biflora*	*L. saltensis*	*C. annuum*	*C. frutescens*	*C. rhomboideum*
A. Substitutions per site (nuclear)						
L. asarifolia	–	1.62E−01	2.08E−02	3.49E−02	3.63E−02	2.89E−02
L. biflora	3.20E−02	–	3.00E−02	3.48E−02	3.62E−02	2.87E−02
L. saltensis	1.09E−02	1.51E−01	–	3.31E−02	3.46E−02	2.69E−02
C. annuum	1.84E−01	1.82E−01	1.74E−01	–	2.70E−03	1.42E−01
C. frutescens	1.93E−01	1.90E−01	1.82E−01	1.41E−02	–	1.51E−01
C. rhomboideum	1.51E−01	1.48E−01	1.40E−01	2.60E−02	2.75E−02	–
B. Substitutions per site per year (nuclear)						
L. asarifolia	–	6.52E−09	1.30E−09	1.44E−09	1.49E−09	1.27E−09
L. biflora	1.28E−09	–	1.22E−09	1.37E−09	1.42E−09	1.17E−09
L. saltensis	6.80E−09	6.17E−09	–	1.39E−09	1.45E−09	1.20E−09
C. annuum	7.82E−09	7.26E−09	7.48E−09	–	1.24E−08	7.44E−09
C. frutescens	8.10E−09	7.49E−09	7.76E−09	5.47E−08	–	7.77E−09
C. rhomboideum	6.68E−09	6.03E−09	6.25E−09	1.35E−09	1.41E−09	–
C. Substitutions per site (plastome)						
L. asarifolia	–	2.28E−02	9.81E−03	7.09E−03	7.14E−03	1.02E−02
L. biflora	8.14E−03	–	1.29E−02	9.32E−03	9.37E−03	1.29E−02
L. saltensis	2.17E−02	2.74E−02	–	1.15E−02	1.15E−02	1.51E−02
C. annuum	2.35E−02	2.56E−02	2.84E−02	–	4.81E−05	1.84E−02
C. frutescens	2.36E−02	2.57E−02	2.84E−02	2.79E−04	–	1.84E−02
C. rhomboideum	2.40E−02	2.58E−02	2.77E−02	9.78E−03	9.82E−03	–
D. Substitutions per site per year (plastome)						
L. asarifolia	–	8.64E−10	5.92E−10	2.93E−10	2.95E−10	4.23E−10
L. biflora	3.08E−10	–	4.88E−10	3.53E−10	3.55E−10	4.89E−10
L. saltensis	1.31E−09	1.04E−09	–	4.75E−10	4.77E−10	6.24E−10
C. annuum	9.73E−10	9.70E−10	1.17E−09	–	2.18E−10	8.93E−10
C. frutescens	9.75E−10	9.72E−10	1.17E−09	1.27E−09	–	8.95E−10
C. rhomboideum	9.92E−10	9.79E−10	1.15E−09	4.75E−10	4.77E−10	–

Nectouxia; Särkinen et al., 2013), all have nectaries and longitudinally dehiscing anthers. Our species tree topology, then, would require at least two separate shifts in the states of these characters to explain their distribution across the phylogeny: first, a shift to poricidal anthers and loss of nectaries on the branch leading to MRCA of Capsiceae, and second, a reversion back to longitudinally dehiscing anthers and presence of nectaries on the branch leading to the MRCA of *Capsicum*. On the other hand, we have identified 769 genes that reconstruct a monophyletic *Lycianthes*. Ancestral state reconstructions using phylogenies from these genes would only require a single evolutionary shift (on the branch leading to the MRCA of *Lycianthes*) to explain the distribution of anthers and nectaries across Capsiceae. While this would present a more parsimonious reconstruction, it could also very well be the case that these genes themselves are involved in the expression of these hemiplasious traits. Thus, while the species tree may present the best summary of overall species relationships, it is not necessarily the best topology to determine the evolutionary history of specific traits or the genes that control their expression (Mendes et al., 2016). More broadly, poricidal anthers and buzz pollination have also evolved in the genus *Solanum*. Whether the multiple appearances of these syndromes in Solanaceae are truly convergent, or reflect the incomplete sorting of associated genes throughout the diversification of the family, remains a key question. This issue of convergence vs. hemiplasy is even more complicated in complex traits that depend on the expression of many genes. For example, genes in our dataset that are associated with the fatty acid and vanillin pathways, which are involved in the development of capsaicin (Liu et al., 2013; Kim et al., 2014), are represented by at least 12 distinct genealogies. An analysis of the evolution of traits associated with these pathways using only the species tree would likely mischaracterize the complexity of these traits; rather, a more thorough analysis would incorporate the historical discordance in the very genes involved in this expression.

Genealogies associated with enrichment of GO terms could similarly be used as a starting point to link gene history with key functional traits involved in domestication, and to broaden the gene pool of these traits in domesticated species by incorporating genotypes of wild relatives (Albert and Chang, 2014; Brozynska et al., 2015). In our dataset, genes associated with 19 of the 25 gene tree topologies (Fig. 2) represent the enrichment of dozens of GO terms, indicating possible non-randomness to the inheritance of alleles and associated genetic functions.

4.2. Genomic structure of substitution rate heterogeneity

We observed substantial heterogeneity in substitution rates among genes, with dS varying ~41-fold and dN varying ~130-fold (Table 3). This level of heterogeneity is much lower than the highest reported instances of rate variance, which is as high as a 340-fold range of synonymous substitutions in *Ajuga* (Lamiaceae) (Zhu et al., 2014). Averaged over all genes and lineages, absolute rates of substitution varied 9-fold, with the highest rate of 5.47×10^{-8} observed since the divergence of the pungent peppers (Table 3). On average, we found that plastid genes appear to be under greater evolutionary constraints than nuclear genes, with the latter evolving about 5.8 times faster than the former, on average (Table 3). While this general pattern is consistent with all reports of substitution rates in other plant lineages, most of these indicate only a 3–4 fold difference in rates between chloroplast and nuclear genes (Wolfe et al., 1987; Drouin et al., 2008; Smith et al., 2014; Zhu et al., 2014).

In addition to the observed differences in rates of substitution between nuclear and plastid genes, rates among nuclear genes vary with their physical location on chromosomes (Fig. 3). In this dataset, the distance of genes from the centromeres alone explains up to 16.3% of the variance in synonymous substitution rates, with rates tending to increase towards chromosome ends (Hill et al., 2015). Two independent research groups have reported strong repression of recombination in pericentromeric regions in *Capsicum* (Qin et al., 2014; Hill et al., 2015),

which would result in reduced rates of mutation as the mutagenic effects of recombination would be minimized in these regions (Baker et al., 2014; Lercher and Hurst, 2002). It should be noted, however, that these conclusions are dependent on high-quality reference genome sequences that have accurately ordered contigs in pseudochromosomes to provide correct relative positions of genes.

This chromosomal structure of rate heterogeneity has implications for experimental design in the age of genomics. As genes near chromosome ends tend to evolve more rapidly, these may be sought as targets for resolving recent, rapidly diversifying, or intraspecific clades. On the other hand, those more slowly evolving genes closer to the centromeres could be useful for resolving deeper nodes in a phylogeny. Our results suggest, however, that this simplified approach to gene selection could result in overconfidence of species tree topologies. While we found that gene history per se is independent of the position of genes on chromosomes, the phylogenetic distance between the trees from adjacent genes increases as those genes approach chromosome ends, even though the physical distance between those genes decreases. Targeting genes close to centromeres and avoiding genes near chromosome ends may then seem like an attractive approach to reduce phylogenetic noise. Our BUCKy randomization tests, however, indicate that such an approach would result in increased confidence of primary splits while simultaneously failing to capture the full spectrum of supported secondary splits, even if many genes are used (Fig. 4). We suggest that phylogenomic experimental design should not only focus on numbers of genes, rates of gene evolution, and gene tree congruence, but should also sample genes throughout the length of the chromosome. We acknowledge, however, that this approach may be impractical if the study system includes distant relatives, where the unambiguous alignment of rapidly evolving genes may not be possible.

4.3. Positive selection through diversification

We identified 787 genes that have experienced instances of positive selection throughout the diversification of Capsiceae (Fig. 3), which could be the subject of further research to determine their functional roles and utility in domestication. Nearly 70% of these genes were selected since the divergence of the *Capsicum* clade, ~40% were selected since the divergence of the pungent *Capsicum* clade, and ~15% were selected since the divergence of *C. annuum* and *C. frutescens*. The disproportionately high number of genes selected along these branches likely reflects the hundreds of years of artificial selection that these species have undergone, for example, to optimize fruit yield, shape, size, maturity, and dehiscence, and plant architecture and dormancy (Albert and Chang, 2014). For example, three genes (CA02g22280, CA05g11830, CA12g21490) involved in the response to heat shock (GO:0031072), which affects many pathways including those in flowering, stress response, and pollination, (Guo et al., 2015), have been positively selected through the diversification of the pungent *Capsicum* species but have not yet been identified as important for domestication. Genes that have been positively selected within *Lycianthes* but not in *Capsicum*, including those representing GO enrichment, may be equally important as targets of investigation to uncover the genetic basis for desirable traits of our crops' wild relatives.

Acknowledgements

This work was funded by NSF DEB 1457366 to LB and NSF DEB 1457351 to AVD. The authors thank the University of California-Davis Genome Center for maintaining high-performance computational resources utilized in the project. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ympev.2018.04.030.

References

Albert, V.A., Chang, T.H., 2014. Evolution of a hot genome. Proc. Natl. Acad. Sci. 111, 5069–5070. http://dx.doi.org/10.1073/pnas.1402378111.

Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2006. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426. http://dx.doi.org/10.1093/molbev/msl170.

Ashrafi, H., Hill, T., Stoffel, K., Kozik, A., Yao, J., Chin-Wo, S.R., van Deynze, A., 2012. De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. BMC Genom. 13, 571. http://dx.doi.org/10.1186/1471-2164-13-571.

Baker, K., Bayer, M., Cook, N., Dreißig, S., Dhillon, T., Russell, J., Hedley, P.E., Morris, J., Ramsay, L., Colas, I., Waugh, R., Steffenson, B., Milne, I., Stephen, G., Marshall, D., Flavell, A.J., 2014. The low-recombining pericentromeric region of barley restricts gene diversity and evolution but not gene expression. Plant J. 79, 981–992. http://dx.doi.org/10.1111/tpj.12600.

Barboza, M.E., 2013. Lycianthes. In: Anton, A.M., Zuloaga, F.O. (Eds.), Barboza, G.E. (coord.) Flora Argentina vol. 13, Solanaceae. IOBDA-IMBIV, CONICET: Buenos Aires & Córdoba, Argentina, pp. 25–30.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B., 2017. Julia: a fresh approach to numerical computing. SIAM Rev. 59, 65–98. http://dx.doi.org/10.1137/141000671.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10, e1003537–e1003546. http://dx.doi.org/10.1371/journal.pcbi.1003537.

Brozynska, M., Furtado, A., Henry, R.J., 2015. Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol. J. 14, 1070–1085. http://dx.doi.org/10.1111/pbi.12454.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. http://dx.doi.org/10.1093/bioinformatics/btp348.

Carrizo García, C., Barfuss, M.H.J., Sehr, E.M., Barboza, G.E., Samuel, R., Moscone, E.A., Ehrendorfer, F., 2016. Phylogenetic relationships, diversification and expansion of chili peppers (*Capsicum,* Solanaceae). Ann. Bot. 118, 35–51. http://dx.doi.org/10.1093/aob/mcw079.

Cauich, O., Quezada Euán, J.J.G., Ramírez, V.M., Valdovinos-Nuñez, G.R., Moo-Valle, H., 2015. Pollination of habanero pepper (*Capsicum chinense)* and production in enclosures using the stingless bee Nannotrigona perilampoides. J. Apic. Res. 45, 125–130. http://dx.doi.org/10.1080/00218839.2006.11101330.

Chung, Y., Ané, C., 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst. Biol. 60, 261–275. http://dx.doi.org/10.1093/sysbio/syr003.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676. http://dx.doi.org/10.1093/bioinformatics/bti610.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772. http://dx.doi.org/10.1038/nmeth.2109.

Dean, E.A., 2004. A taxonomic revision of *Lycianthes* series *Meizonodontae* (Solanaceae). Bot. J. Linn. Soc. 145, 385–424. http://dx.doi.org/10.1111/j.1095-8339.2004.00296.x.

Drouin, G., Daoud, H., Xia, J., 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49, 137–141. http://dx.doi.org/10.1016/j.ympev.2008.09.009.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., Su, Z., 2010. agriGO: a GO analysis toolkit for the agricultural community. Nucl. Acids Res. 38, W64–W70. http://dx.doi.org/10.1093/nar/gkq310.

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., Mueller, L.A., 2015. The Sol Genomics Network (SGN)–from genotype to phenotype to breeding. Nucl. Acids Res. 43, D1036–D1041. http://dx.doi.org/10.1093/nar/gku1195.

Góngora-Castillo, E., Fajardo-Jaime, R., Fernández-Cortes, A., Jofre- Garfias, A.E., Lozoya-Gloria, E., Martínez, O., Ochoa-Alejo, N., Rivera-Bustamante, R., 2012. The *Capsicum* transcriptome DB: a "hot" tool for genomic research. Bioinformation 8, 43–47. http://dx.doi.org/10.6026/97320630008043.

Guo, M., Lu, J.-P., Zhai, Y.-F., Chai, W.-G., Gong, Z.-H., Lu, M.-H., 2015. Genome-wide analysis, expression profile of heat shock factor gene family (CaHsfs) and characterisation of CaHsfA2 in pepper (*Capsicum annuum* L.). BMC Plant Biol. 1–20. http://dx.doi.org/10.1186/s12870-015-0512-7.

Haak, D.C., McGinnis, L.A., Levey, D.J., Tewksbury, J.J., 2012. Why are not all chilies hot? A trade-off limits pungency. Proc. R. Soc. B: Biol. Sci. 279, 2012–2017. http://dx.doi.org/10.1098/rspb.2011.2091.

Hill, D.P., Smith, B., McAndrews-Hill, M.S., Blake, J.A., 2008. Gene Ontology annotations: what they mean and where they come from. BMC Bioinf. 9, S2–S9. http://dx.doi.org/10.1186/1471-2105-9-S5-S2.

Hill, T., Ashrafi, H., Chin-Wo, S.R., Stoffel, K., 2015. Ultra-high density, transcript-based genetic maps of pepper define recombination in the genome and synteny among related species. G3: Genes. http://doi.org/10.1534/g3.115.020040/-/DC1.

Jarlan, A., de Oliveira, D., Gingras, J., 1997. Pollination of sweet pepper (*Capsicum annuum* L.) in greenhouse by the syrphid fly *Eristalis tenax* (L.). Acta Horticulturae 335–340. http://dx.doi.org/10.17660/actahortic.1997.437.41.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. http://dx.doi.org/10.1093/molbev/mst010.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199.

Kim, H.-J., Baek, K.-H., Lee, S.-W., Kim, J., Lee, B.-W., Cho, H.-S., Kim, W.T., Choi, D., Hur, C.-G., 2008. Pepper EST database: comprehensive in silico tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome. BMC Plant Biol. 8, 101. http://dx.doi.org/10.1186/1471-2229-8-101.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J.M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., Jung, K., Lee, G.-W., Oh, S.-K., Bae, C., Kim, S.-B., Lee, H.-Y., Kim, S.-Y., Kim, M.-S., Kang, B.-C., Jo, Y.D., Yang, H.-B., Jeong, H.-J., Kang, W.-H., Kwon, J.-K., Shin, C., Lim, J.Y., Park, J.H., Huh, J.H., Kim, J.-S., Kim, B.-D., Cohen, O., Paran, I., Suh, M.C., Lee, S.B., Kim, Y.-K., Shin, Y., Noh, S.-J., Park, J., Seo, Y.S., Kwon, S.-Y., Kim, H.A., Park, J.M., Kim, H.-J., Choi, S.-B., Bosland, P.W., Reeves, G., Jo, S.-H., Lee, B.-W., Cho, H.-T., Choi, H.-S., Lee, M.-S., Yu, Y., Do Choi, Y., Park, B.-S., van Deynze, A., Ashrafi, H., Hill, T., Kim, W.T., Pai, H.-S., Ahn, H.K., Yeam, I., Giovannoni, J.J., Rose, J.K.C., Sørensen, I., Lee, S.-J., Kim, R.W., Choi, I.-Y., Choi, B.-S., Lim, J.-S., Lee, Y.-H., Choi, D., 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat. Genet. 46, 270–278. http://dx.doi.org/10.1038/ng.2877.

Kristjansson, K., Rasmussen, K., 1991. Pollination of sweet pepper (*Capsicum annuum* L.) with the solitary bee *Osmia cornifrons* (Radoszkowski). Acta Horticulturae 173–179. http://dx.doi.org/10.17660/actahortic.1991.288.24.

Larget, B.R., Kotha, S.K., Dewey, C.N., Ane, C., 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26, 2910–2911. http://dx.doi.org/10.1093/bioinformatics/btq539.

Lercher, M.J., Hurst, L.D., 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 18, 337–340.

Liu, S., Li, W., Wu, Y., Chen, C., Lei, J., 2013. De novo transcriptome assembly in chili pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids. PLoS ONE 8, e48156. http://dx.doi.org/10.1371/journal.pone.0048156.

Olmstead, R.G., Bohs, L., Migid, H.A., Santiago-Valentin, E., Garcia, V.F., Collier, S.M., 2008. A molecular phylogeny of the Solanaceae. Taxon 57, 1159–1181.

Maddison, W., Knowles, L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30. http://dx.doi.org/10.1080/10635150500354928.

Meisels, S., Chiasson, H., 1997. Effectiveness of *Bombus impatiens* cr. as pollinators of greenhouse sweet peppers (*Capsicum annuum* L.). Acta Horticulturae 425–430. http://dx.doi.org/10.17660/actahortic.1997.437.56.

Mendes, F.K., Hahn, Y., Hahn, M.W., 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol. Biol. Evol. 33, 3299–3307. http://dx.doi.org/10.1093/molbev/msw197.

Nee, M.. 1991. The systematics of lesser known edible Solanaceae of the New World. In: Hawkes, J.G., Lester, R.N., Nee, M., Estrada, N., (Eds.), Solanaceae III: Taxonomy, Chemistry, Evolution. Royal Botanic Gardens, Kew, UK. pp. 365–568.

Park, H.-S., Lee, J., Lee, S.-C., Yang, T.-J., Yoon, J.B., 2016. The complete chloroplast genome sequence of *Capsicum chinense* Jacq. (Solanaceae). Mitochondrial DNA Part B 1, 164–165. http://dx.doi.org/10.1080/23802359.2016.1144113.

Park, M., Park, J., Kim, S., Kwon, J.K., Park, H.M., 2012. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. Plant J. http://dx.doi.org/10.1111/j.1365-313X.2011.04851.x.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., Yang, Y., Wu, Z., Mao, L., Wu, H., Ling-Hu, C., Zhou, H., Lin, H., Gonzalez-Morales, S., Trejo-Saavedra, D.L., Tian, H., Tang, X., Zhao, M., Huang, Z., Zhou, A., Yao, X., Cui, J., Li, W., Chen, Z., Feng, Y., Niu, Y., Bi, S., Yang, X., Li, W., Cai, H., Luo, X., Montes-Hernandez, S., Leyva-Gonzalez, M.A., Xiong, Z., He, X., Bai, L., Tan, S., Tang, X., Liu, D., Liu, J., Zhang, S., Chen, M., Zhang, L., Zhang, L., Liao, W., Zhang, Y., Wang, M., Lv, X., Wen, B., Liu, H., Luan, H., Yang, S., Wang, X., Xu, J., Li, X., Li, S., Wang, J., Palloix, A., Bosland, P.W., Li, Y., Krogh, A., Rivera-Bustamante, R.F., Herrera-Estrella, L., Yin, Y., Yu, J., Hu, K., Zhang, Z., 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc. Natl. Acad. Sci. 111, 5135–5140. http://dx.doi.org/10.1073/pnas.1400975111.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raw, A., 2000. Foraging behaviour of wild bees at hot pepper flowers (*Capsicum annuum*) and its possible influence on cross pollination. Ann. Bot. 85, 487–492. http://dx.doi.org/10.1006/anbo.1999.1090.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542. http://dx.doi.org/10.1093/sysbio/sys029.

Sanderson, M.J., Nicolae, M., McMahon, M.M., 2017. Homology-aware phylogenomics at gigabase scales. Systematic Biology syw104–14. http://doi.org/10.1093/sysbio/syw104.

Särkinen, T.S., Bohs, L., Olmstead, R.G., Knapp, S., 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol. Biol. 13, 1–11. http://dx.doi.org/10.1186/1471-2148-13-214.

Scornavacca, C., Galtier, N., 2016. Incomplete lineage sorting in mammalian phylogenomics. Systematic Biology syw082–9. http://doi.org/10.1093/sysbio/syw082.

Smith, D.R., Arrigo, K.R., Alderkamp, A.-C., Allen, A.E., 2014. Massive difference in synonymous substitution rates among mitochondrial, plastid, and nuclear genes of Phaeocystis algae. Mol. Phylogenet. Evol. 71, 36–40. http://dx.doi.org/10.1016/j.ympev.2013.10.018.

Smith, S.D., Knapp, S., 2002. The natural history of reproduction in *Solanum* and *Lycianthes* (Solanaceae) in a subtropical moist forest. BBO 32, 1–12. http://dx.doi.org/10.1017/S0968044602000051.

Solís-Lemus, C., Ané, C., 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12, e1005896–e1005921. http://dx.doi.org/10.1371/journal.pgen.1005896.

Stenz, N.W.M., Larget, B., Baum, D.A., Ané, C., 2015. Exploring tree-like and non-tree-like patterns using genome sequences: an example using the inbreeding plant species *Arabidopsis thaliana* (L.) Heynh. Syst. Biol. 64, 809–823. http://dx.doi.org/10.1093/sysbio/syv039.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., Malmberg, R.L., 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol. Phylogenet. Evol. 85, 76–87. http://dx.doi.org/10.1016/j.ympev.2015.01.015.

Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol. Biol. Evol. 18, 2057–2066.

Tung Ho, L.S., Ane, C., 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol. 63, 397–408. http://dx.doi.org/10.1093/sysbio/syu005.

Venkat, A., Hahn, M.W., Thornton, J.W., 2017. Multi-nucleotide mutations cause false inferences of positive selection. http://doi.org/10.1101/165969.

Walsh, B.M., Hoot, S.B., 2001. Phylogenetic relationships of *Capsicum* (Solanaceae) using DNA sequences from two noncoding regions: the chloroplast atpB- rbcL spacer region and nuclear waxy introns. Int. J Plant Sci. 162, 1409–1418. http://dx.doi.org/10.1086/323273.

Williams, D.E., 1993. *Lycianthes moziniana* (Solanaceae): an underutilized Mexican food plant with "new" crop potential. Econ. Bot. 47, 387–400. http://dx.doi.org/10.1007/bf02907353.

Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. 84 (24), 9054–9058. http://dx.doi.org/10.1073/pnas.84.24.9054.

Winter D., 2017. rentrez: Entrez in R. R package version 1.1.0. https://CRAN.R-project.org/package=rentrez.

Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591. http://dx.doi.org/10.1093/molbev/msm088.

Zhang, J., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479. http://dx.doi.org/10.1093/molbev/msi237.

Zhong, S., Joung, J.-G., Zheng, Y., Chen, Y.-R., Liu, B., Shao, Y., Xiang, J.Z., Fei, Z., Giovannoni, J.J., 2011. High-throughput Illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc 2011, 940–949. http://dx.doi.org/10.1101/pdb.prot5652.

Zhu, A., Guo, W., Jain, K., Mower, J.P., 2014. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. Mol. Biol. Evol. 31, 1228–1236. http://dx.doi.org/10.1093/molbev/msu079.