Evolution of the *rbcS* gene family in Solanaceae: concerted evolution and gain and loss of introns, with a description of new statistical guidelines for determining the number of unique gene copies

Ryan J. Miller

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Richard G. Olmstead, Chair

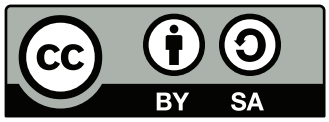
Willie J. Swanson

Vladimir N. Minin

Program Authorized to Offer Degree:

Biology

Ryan J. Miller



Dedication

For my sister, Meredith,

my son, Jarek,

and his mother, Kimberly,

for inspiring me,

lifting me up when necesary,

and supporting me.

Acknowledgements

I would like to thank my advisor, Dick Olmstead, for introducing me to this project and for his insights on science and writing, and his steadfast support throughout. I also thank my committee members, especially Vladimir Minin, for enriching my enjoyment in science.

Thanks to all my teachers: Ken Karol, for his patient lessons on PCR and life, Dave Tank, for sharing his love of phylogenetics, Valerie Soza, Yao-wu Yuan, Pat Lu-Irving, and John Chao, for worrying about me and listening to all my confusion.

Thanks to the faculty in Biology, especially Keiko Torii, for believing in me and bringing me to Seattle, and Alison Crowe, who trained me in the most amazing manipulation of DNA I'll ever perform, and Liz Van Volkenburgh for her frank conversations on plants. Thanks to Arnie Bendich and Delene Oldenburg for encouragement, and to David Giblin for his patience teaching herbarium practices to a molecular biologist .

Thanks to the faculty, post-docs, and grad students of the Evol/Syst seminar for conversations on deep time, selection, species and everything else.

Thanks to my cohort for friendship and for sharing such different passions for biology.

Thanks to Ben Hall and the Biology department, for the significant level of support I received from the Hall Royalty Plant Biology Fellowship and from many teaching assistantships. Thanks, as well, for support from the

NIH/NHGRI Genome Sciences Training Grant and the Melinda Denton Writing Fellowship.

Thanks to Alison Colwell, for her initial work on the research described in Chapter 1, and to Hanhla Phan and Tauras Vilgalys, for contributing sequence data for the project and all their hardwork.

Also, thanks to Lynn Bohs and the other donors of DNA samples that made this study possible.

University of Washington

ABSTRACT

Evolution of the *rbcS* gene family in Solanaceae:
concerted evolution and gain and loss of introns,

with a description of new statistical guidelines
for determining the number of unique gene copies

Ryan J. Miller

Chair of the Supervisory Committee:
Professor of Biology and Herbarium Curator, Burke Museum Richard G. Olmstead
Biology

Concerted evolution is a pattern of gene evolution resulting from mechanisms that homogenize gene copies within a lineage. Two mechanisms have been identified that may homogenize *rbcS* copies within Solanaceae. Selection was examined as a mechanism of homogenization through separation of synonymous and nonsynonymous substitutions, through the branch site test of positive selection, and codon usage. Strong negative selection and selection for codon usage

were identified, but positive selection also was found to be contributing to homogenization of paralogs in one lineage within *Solanum*. No evidence was found for gene conversion. The results support a role for positive selection as a mechanism in concerted evolution and highlight the danger of inferring a species tree from any set of genes undergoing homogenization. Chapter 1, Supplementary Table 4 (attached as separate file) lists ENc and 3rd codon position composition.

Among land plants *rbcS* copies contain two introns at homologous locations. In addition to 2-intron *rbcS* copies, Solanaceae lineages have *rbcS* copies with three introns. *rbcS* copies with 3-introns at homologous positions to other 3-intron Solanaceae copies was identified from *Cestrum*. Phylogenetic analyses indicate this novel, third intron may have originated from a locus of tandemly repeated rbcS copies with 2-introns. Numerous sequence motifs similar to transposable elements and containing direct repeats and inverted repeats were identified. These sequence features may contribute to a high divergence in intron sequence between these 3-intron *rbcS* copies.

Gene duplication has long been thought to play an important evolutionary role and many genes are now known to differ in copy number between closely related lineages. PCR, cloning, and sequencing is a commonly employed method to examine gene copies from a group of taxa but a standard statistical method to determine whether the actual number of gene copies has been sequenced is lacking from most studies. Simulations indicate the lower bounds for the number of clones necessary to find a given number of unique gene copies and a parametric bootstrap test provides researchers with a method to gauge whether more copies remain unidentified.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

CHAPTER 1:Selection-mediated concerted evolution of *rbcS* copies in *Solanum abutiloides*

Introduction

Concerted evolution is the unexpected pattern of gene evolution resulting from processes that cause gene copies to evolve collectively in a non-independent manner (Zimmer et al. 1980). Evolving collectively creates a pattern where gene copies within one species are more similar to each other than to any corresponding orthologous copy in another species. The pattern of concerted evolution is common to many multigene families and is thought to occur by three general mechanisms: 1) duplication/loss, 2) nonhomologous recombination, and 3) selection (Hood et al. 1975; Ohta 1983; Nei et al. 1997). The *rbcS* multigene family exhibits a pattern of concerted evolution when copies from relatively distant species are compared. *rbcS* encodes the small subunit of ribulose-1,5-bisphosphate carboxylase oxygenase (Rubisco) and generally consists of two to eight copies within flowering plants. Two mechanisms, selection (Pichersky et al. 1986) and nonhomologous recombination (Meagher et al. 1989), have been postulated as the primary homogenizing mechanism for *rbcS* copies. Our purpose is to evaluate the pattern of concerted evolution among *rbcS* copies to determine which hypothesis is better supported.

Differential gene duplication and loss along species lineages can create a pattern of concerted evolution (Nei et al. 1997). This birth-and-death model of gene evolution explains apparent similarity between gene copies, not as caused by homogenization, but by expansion and loss of different copies between species. This model predicts the presence of many pseudogenes and has been shown to explain the evolution of major histocompatibility complex genes and immunoglobulin genes (Nei et al. 1997), among histone genes (Rooney et al. 2002), as well as ubiquitin genes (Nei et al. 2000).

Nonhomologous recombination can homogenize gene copies through unequal crossing over

when copies are arrayed in tandem (Smith 1976) or by gene conversion. Unequal crossing over can occur between nonhomologous copies on the same chromosome when sister chromatids exchange DNA or between nonhomologous copies on either homologous or nonhomologous chromosomes during meiosis and has been proposed as the primary mechanism homogenizing rRNA genes (Brown et al. 1972; Eickbush and Eickbush 2007). Gene conversion can also lead to homogenization of unlinked gene copies through the nonreciprocal transfer of DNA sequence from one locus to another. In yeast, gene conversions are modeled to occur through a double strand break repair process that uses DNA sequence from a donor copy to repair the break in the recipient gene copy (Pâques and Haber 1999; Zickler and Kleckner 1999). Evidence of gene conversion has been found in many taxa, in plants it has been shown to occur between nucleotide-binding site leucine-rich repeat (NBS-LRR), receptor-like kinases (RLK), and receptor-like protein genes in *Arabidopsis thaliana* (Mondragon-Palomino and Gaut 2005), between RPP8 gene copies in three species of *Arabidopsis* (Kuang et al. 2008), among multigene family members in sorghum and rice (Wang et al. 2007), and ten gene conversions were identified by next-gen sequencing of the four meiotic products from an *Arabidopsis* hybrid (Lu et al. 2012).

Studies of gene conversion implicate a mechanism that is dependent on high similarity between recipient and donor. Gene conversion tracts can vary tremendously in size from as small as dozens of bases to as long as six thousand bases, with mean tract size of 370 ± 750 in humans (Benovoy and Drouin 2009). Sequence divergence can inhibit pairing of nonhomologous regions during meiosis and thus act as a barrier to gene conversion (Teshima and Innan 2004). Hence, a balance must be maintained between gene conversion and nucleotide substitution for gene conversion to continue.

Selection can act to constrain DNA sequence change and prevent nonsynonymous substitutions or to promote substitutions at specific sites in all gene copies. In the absence of selection, independent substitutions at each gene copy are expected to increase divergence between copies, and the ratio of expected nonsynonymous substitutions/synonymous substitutions (d_N/d_S) would equal one. For selection to act as a homogenizing mechanism between copies within a species, positive selection could act to promote the same particular substitutions in independent gene copies that have previously diverged at those sites. If enough sites are homogenized between copies within one species, a pattern of concerted evolution by positive selection could emerge. This pattern would be typified by sites with nonsynonymous substitutions when orthologous copes are compared. When those same sites are examined between the paralogous copies homogenized by positive selection, unique synonymous codons would be expected due to the independent pathway of substitutions each copy experienced. Homogenization by selection has been discussed as a mechanism for homogenization (Hood et al. 1975), but evidence is generally lacking. Most studies inferring selection also invoke additional mechanisms acting in tandem, such as birth-and-death evolution (Nei et al. 2000; Eirín-López et al. 2004) or gene conversion (Mondragon-Palomino and Gaut 2005) to be responsible for creating the pattern of concerted evolution.

Codon usage bias can be considered a particular case of the selection hypothesis that deserves attention when evaluating concerted evolution because it can affect divergence levels between gene copies and skew the rate of synonymous substitutions. In every genome examined, a nonrandom usage of synonymous codons has been found where preferred codons for an amino acid occur more often than other synonymous codons. Preferred codons can differ between lineages (Duret and Mouchiroud 1999) although among eudicots codon preference is highly

conserved. The same 21 preferred codons were identified in *Arabidopsis thaliana* and *Silene latifolia* (Qiu et al. 2011) and these were very similar to preferred codons in *Populus* species (Ingvarsson 2008), and in *Nicotiana tabacum* and *Pisum sativum* (Kawabe and Miyashita 2003).

Differences in codon usage bias can be due to selection for translational efficiency/accuracy or different mutational biases or a combination of the two (Hershberg and Petrov 2008). Selection is supported when codon usage bias correlates with frequencies of tRNA abundance and expression such that highly expressed genes show higher levels of bias (Duret 2002). Mutational biases are supported when codon bias correlates to GC content (Ikemura 1985). Mutational biases and processes like GC biased gene conversion can be gene specific and can significantly bias substitution patterns among vertebrates (Berglund et al. 2009). If mutational bias is the cause of codon bias, the composition of G and C or A and T nucleotides at synonymous positions should be proportional (Wright 1990). It is often unknown whether taxon-specific preferences for synonymous codons are due to translational efficiency or mutational biases but codon usage bias can decrease divergence between genes (Sharp and Li 1987) by maintaining similarities between gene copies and can increase sequence divergence between taxa when preferred codons are different by species (Lin et al. 2006). Comparisons across a gene family in different species indicate more codon usage differences between genes than differences in codon usage between species (Zhang et al. 2009).

Previous studies (Pichersky et al. 1986; Meagher et al. 1989; Clegg et al. 1997) identified concerted evolution among the gene copies encoding the small subunit of Rubisco (*rbcS*) in three disparate species of Solanaceae (petunia, tobacco, and tomato) where a higher degree of similarity is found between gene copies within a species than when orthologous copies are compared between species. However, this pattern is not observed between two closely related

species in *Solanum* (tomato and potato), where orthologous copies are more similar to each other than are the copies within each species. Thus, in tomato and potato, evidence of concerted evolution has not appeared in the time since their divergence. Two hypotheses have been proposed for these contrasting patterns. In 1986, Pichersky and colleagues compared nucleotide and amino acid similarities between orthologous and paralogous sequences from tomato and tobacco. They found that nucleotide differences between species were not much different from nucleotide differences between copies of the same species and it was only at the amino acid level that copies within a species showed a greater level of similarity. They concluded that selection acting on the Rubisco small subunit amino acid sequence was homogenizing coding DNA among copies within a species. Subsequently, Meagher et al. (1989) dismissed selection as a mechanism due to the required time and/or rate of substitution needed to generate that many coding changes at every gene copy and suggested interchromosomal gene conversion as the mechanism homogenizing *rbcS* copies (Meagher et al. 1989).

The *rbcS* gene encodes a peptide of approximately 180 amino acids. Eight small subunits encoded by *rbcS* interact with eight large subunits encoded by the chloroplast gene *rbcL* in the chloroplast to form the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). Rubisco catalyzes CO_2 fixation during photosynthesis and is the most abundant protein on the planet (Dean et al. 1989). The first ~70 amino acids of the small subunit form the transit sequence for import into the chloroplast and are removed during or after transport to the organelle (Chua and Schmidt 1978). The transit peptide sequences for *rbcS* vary substantially more than those encoding the mature peptide encoding region. Comparisons of transit sequences reveal a conserved nine amino acid segment near the middle that is important for interaction with components of the chloroplast membrane (Mishkind et al. 1985). The mature peptide of the small

subunit interacts with the large subunit to form the Rubisco holoenzyme. Areas of interaction between neighboring small subunits and between small subunits and large subunits are extensive (Schneider et al. 1992). Although no small subunit amino acid residues contribute to the active site, the small subunit has dramatic effect on catalytic activity through undetermined interactions distant from the active site (Schneider et al. 1992). The region of the small subunit most often implicated as contributing to differences in Rubisco catalytic efficiency between species is a 22 amino acid loop encoded at the beginning of *rbcS* exon 3 (Spreitzer and Salvucci 2002).

Within Solanaceae, structural differences between *rbcS* loci enable discrimination of the individual loci despite the effects of concerted evolution. The *rbcS* coding structure is interrupted by two introns in most dicots. A unique locus has been identified in Solanaceae that is easily identifiable by the interruption of the third exon by a novel intron(Dean et al. 1989). The two other loci contain the common two-intron structure, one consists of a single copy and the other is characterized by a series of short tandemly repeated copies (e.g., three in *Solanum* and five in *Petunia* (Dean et al. 1989)). The intron differences between copies allows for easy identification of orthologous loci among Solanaceae and rules out the birth-and-death process that often homogenizes other gene copies.

The physically adjacent *rbcS* copies in the tandem repeat locus show much higher levels of sequence similarity than do inter-locus comparisons. The process of unequal crossing over has been shown to effectively homogenize tandemly arrayed copies and has been suggested to be the mechanism responsible for homogenizing the repeated copies at the tandem repeat locus (Sugita et al. 1987).

rbcS genes are highly transcribed and can exhibit distinct expression patterns indicative of sub-functionalization. In *Arabidopsis*, four *rbcS* copies are expressed differently in both

development and tissue type (Sawchuk et al. 2008). In tomato, all copies are expressed at high levels in leaf tissue but are expressed differently during development (Sugita and Gruissem 1987).

Our purpose is to evaluate the pattern of concerted evolution among the three *rbcS* loci in *Solanum* by evaluating each of the following potential homogenizing mechanisms: positive selection, gene conversion, and codon usage bias. To accomplish this we have sampled *rbcS* copies from taxa of increasing divergence from tomato and potato, where there is no pattern of concerted evolution (except for copies from the tandem repeat locus), with the goal of locating where the homogenization signal is identifiable and then examining the sequences using phylogenetic techniques to determine which hypothesis, selection or gene conversion, is better supported.

Materials and Methods

DNA extraction, PCR, and sequencing

10 species of *Solanum* representing lineages of increasing phylogenetic distance from tomato and potato, were included in this study (Supplementary Table 1) (Weese and Bohs 2007). Sequences from one species of *Jaltomata*, the sister genus to *Solanum* (Olmstead et al. 2008), were included (Supplementary Table 1). Sequences from *Capsicum*, sister to the clade of *Solanum* and *Jaltomata*, were also included (Olmstead et al. 2008). Available *rbcS* sequences from public databases were downloaded as follows: five genomic sequences from GenBank (www.ncbi.nlm.nih.gov/) (Benson et al. 2012) for both *Solanum lycopersicum* and *S. tuberosum*,

two full length unigenes from Sol Genomics Network (SGN) (www.solgenomics.net) (Bombarely et al. 2011) for *Capsicum annuum*, three Genome v0.3 scaffolds containing unique and complete copies of *rbcS* from *Nicotiana benthamiana*, eight processed reads of Genome TGI:v.1 from *Nicotiana tabacum*, and two *Petunia hybrida* Mitchell sequences from GenBank (Turner et al. 1986). Five *rbcS* coding sequences from *Petunia hybrida* Mitchell sequences (Dean et al. 1987) not available from GenBank were manually entered from the original publication.

Total genomic DNA was extracted from field-collected, silica-gel dried tissue using the modified 2x CTAB method (Doyle and Doyle 1987) and purified using Wizard minicolumns (Promega, Madison, Wisconsin, USA). The polymerase chain reaction (PCR) was performed to amplify various regions of the nuclear gene for the small subunit of ribulose bisphosphate carboxylase oxygenase (*rbcS*). PCR was conducted in 25 µL volumes with annealing temperatures of 48–60°C using primers from Supplementary Table 2. Amplified *rbcS* products were either gel isolated or cloned before direct sequencing. For gel isolation, bands were separated on a 1% agarose gel and purified using the Promega Wizard SV Gel and PCR Clean-up System (Promega, Madison, Wisconsin, USA). Otherwise, *rbcS* products were cleaned by precipitation from a 20% polyethylene glycol 8000/NaCl solution and washed with 70% EtOH prior to cloning. PCR products were cloned using the TOPO TA Cloning Kit for Sequencing (Invitrogen, Carlsbad, CA, USA) and 12-100 clones (Supplementary Table 3) were PCR amplified with vector primers for direct sequencing. Both strands were sequenced using either DYEnamic ET Terminator (GE Healthcare, Piscataway, NJ, USA) or BigDye Terminator v3.1 (Applied Biosystems Inc, Foster City, CA, USA) cycle sequencing kits on ABI model 377 or 3100 automated DNA sequencers. Sequence data were proofed, edited, and contigs assembled using

Sequencher v.4.7 (Gene Codes Corp., Ann Arbor, MI, USA). Sequences were grouped by similarity and consensus sequences were constructed from clones that differed by less than 1% (Supplementary Table 3).

Phylogenetic analyses

Sequences were manually aligned using the program MacClade 4.08 OSX (Sinauer Associates, Inc, Sunderland, MA, USA). Introns could not be aligned across species nor between loci within the same species and were removed before further analysis. Phylogenetic analyses were performed using a maximum likelihood inference method. DT-Model Select (Minin et al. 2003) was used to determine the best-fit model of sequence evolution. For both the full *rbcS* coding region and the mature peptide coding region the TVMef model (variable transversion rates, equal transition rates with equal nucleotide frequencies) with discrete gamma distributed rate variation (Yang 1994) model of sequence evolution was indicated as the best-fit to the data; the SYM (symmetrical model with 6 rate classes) and discrete gamma distributed rate variation model was the best-fit for the region encoding the transit peptide. The same models were supported as the best-fit in jModelTest v.2 by both the Akaike Information Criterion and Bayesian Information Critertion (Darriba et al. 2012).

Maximum likelihood (ML) analyses were performed using the program GARLI version 2.0 (Zwickl 2006). Forty searches were performed under default parameters (including four categories of discrete approximation of gamma-distributed rate heterogeneity, starting trees created by stepwise-addition with 50 attachment branches evaluated for each taxon, branch-length optimization started at 0.5, reduced 20 times to a minimum of 0.01). To verify

convergence, 20 searches were started from stepwise-addition trees and 20 from random trees. Free model parameters were estimated for each search replicate and automatically terminated after 20,000 generations without improvement in the topology score. One thousand bootstrap (Felsenstein 1985) repetitions were performed with the same parameters as above, but with two replicates per search and all starting trees created by stepwise-addition. Each pseudoreplicate was automatically stopped after 10,000 generations without improvement in the topology score and bootstrap proportions were calculated by computing a majority rule consensus tree with SumTrees (Sukumaran and Holder 2010).

Trees were rooted between the *Petunia* sequences and the rest of the sequences. This rooting is justified on the basis of 1) *Petunia* is the appropriate outgroup among the sampled species (Olmstead et al. 2008), and 2) preliminary analyses of *rbcS* sequences that included sequences from more distantly related families (Pichersky et al. 1986; Meagher et al. 1989). Trees were visualized using the program FigTree (http://tree.bio.ed.ac.uk/software/figtree).

Topology testing

Topologies inferred from different functional regions of *rbcS* were compared using the approximately unbiased test (Shimodaira 2002). Site-wise log-likelihoods under the previously used substitution rate models were obtained using PAUP* version 4.0b10 (Swofford 2002) with model parameters estimated by Garli-2.0 (Zwickl 2006) and imported into CONSEL (Shimodaira and Hasegawa 2001). 10,000 bootstrap replicate log-likelihoods were generated to obtain P values for each topology comparison and topologies with P values < 0.05 were rejected.

Gene conversion and recombination

Two methods were used to screen the aligned data for signals of recombination, GENECONV and the SBR/GARD method implemented on the Datamonkey webserver (Pond and Frost 2005). GENECONV (http://www.math.wustl.edu/~sawyer/geneconv/) (Sawyer 1999) tests for recombination by identifying significant clusters of identical substitutions between pairs of sequences. Default settings were used except as follows: mismatch penalties set to 1 (Gscale=1), only silent sites were analyzed (seqtype=SILENT). Statistical significance was assessed globally with 10,000 random permutations of the alignment. Pairwise P values are based on comparison of each fragment with the maximum length fragment expected from each pair by chance (Sawyer 1999). GENECONV has been shown to give false positives when sequences are <80% identical (Posada and Crandall 2001), however, all sequences in the dataset fall within a range of 80-99% similar. Following previous recommendations, only conversion tracts with a global P value of ≤ 0.05 were considered (Drouin 2002).

Single Breakpoint Recombination (SBR)/Genetic Algorithms for Recombination Detection (GARD) (Pond et al. 2006) uses a likelihood-based algorithm to search for neighboring gene fragments with discordant phylogenetic signal and assesses goodness of fit for the best estimate of breakpoint number. The HKY85 (Hasegawa et al. 1985) nucleotide substitution model was determined by the model selection tool on the Datamonkey server and used for both SBR and GARD with gamma rate variation and three rate classes.

Codon analyses

Effective Number of Codons (ENC) measures the degree of bias away from equal use of synonymous codons (Wright 1990). ENC does not require knowledge of preferred codons but

measures bias in codon usage by reducing from a value of 61 where all codons are used equally. Low ENC values indicate a nonuniform codon usage in a sequence. CodonW (http://codonw.sourceforge.net/) was used to calculate ENC and nucleotide composition indices at third codon position synonymous sites (Peden 2000). GC content at the third position of synonymous codons (GC3) was used to calculate the expected ENC (Wright 1990) for each sequence using the corrected version of the original formula (Peden 2000).

The codon substitution model FMutSel (Yang and Nielsen 2008) included in PAML4 (Yang 2007) was used to examine mutational biases and selection at silent sites under a ML framework. The FMutSel model is compared to a null model, FMutSel0, that incorporates only mutational biases and lacks the codon fitness parameters included in FMutSel that models selection on synonymous sites. A Likelihood Ratio Test between these two models is performed with 41 degrees of freedom.

A package of programs for examining codon usage (http://www.life.illinois.edu/gary/programs/codon_usage.html) was used to calculate codon usage distances between sequences (Davis and Olsen 2010). Distance is calculated as the sum of the differences in relative codon usage frequencies. For each set of synonymous codons, distance values range between 0, where synonymous codon usage is identical between sequences, and 1, when no overlap in codon usage exists. Total distance in codon usage between two sequences is the square root of the sum of the square of the individual distances for all 18 synonymous codon groups. The neighbor program included within the Phylip v3.69 software package (http://evolution.genetics.washington.edu/phylip.html) was used to build a neighbor-joining tree of synonymous codon usage distances. The tree was arbitrarily rooted at the branch leading to the locus 3 sequences of *Petunia hybrida*.

Positive selection analysis

The program PAML4 was used to test for positive selection (Yang 2007). The branch-site model A compares a null hypothesis tree, where the branch under consideration is allowed to evolve without constraint ($d_N/d_S = 1$), to a tree where the same branch has a proportion of sites evolving under positive selection ($d_N/d_S > 1$) (Zhang et al. 2005). The branch leading to all copies from *Solanum abutiloides* was examined using this test. A hierarchical Likelihood Ratio Test was used to compare these two models; twice the difference between the two log likelihood scores was compared to a χ^2 distribution with 1 degree of freedom to reject the null hypothesis ($p<0.05$). When positive selection was detected, the Bayes empirical Bayes (BEB) procedure was used to estimate the posterior probability that a site evolved under positive selection.

Phylogenetic patterns of synonymous and nonsynonymous substitutions

The R package, markovjumps (http://www.stat.washington.edu/vminin/markovjumps/) (O'Brien et al. 2009), was used with default settings to calculate synonymous and nonsynonymous distances using robust counting for a codon alignment of the *rbcS* mature peptide. A phylogenetic tree was constructed from each set of distances using the neighbor joining function in the R package, APE v3.0-2 (Paradis et al. 2004). Significance was assessed with 1,000 bootstraps by resampling with replacement entire codons from the alignment. Bootstrap values were summarized and mapped onto the neighbor joining tree using SumTrees (Sukumaran and Holder 2010).

Results

Gene trees

The ML tree inferred from the full *rbcS* coding sequence (Figure 1) reveals a pattern of concerted evolution where gene copies from different loci within the same species group together. Paralagous copies form species clades for *Solanum abutiloides* and *Petunia hybrida*. Paralogs in *Petunia hybrida* form a clade (96% bootstrap support) sister to all remaining sequences. Paralogs in *Solanum abutiloides* form a clade (98%) sister to a clade of locus 2 sequences from *Solanum dulcamara and S. jasminoides* species (69% for the inclusive clade) and nested within a clade containing locus 2 sequences from *Solanum* and *Jaltomata* (68%). Other well-supported clades include the locus 2 sequences from the two *Nicotiana* species (96%), the locus 3 sequences from the two *Nicotiana* species (96%), and locus 1 sequences from *Solanum* species excluding *Solanum abutiloides* (81%).

The ML tree inferred from the transit sequence portion of *rbcS* (Figure 2) shows a pattern of relationships more consistent with orthologous relationships. Two main clades are identifiable: sequences from locus 1 (57%) are sister to the remaining sequences, which form two clades, a clade of locus 2 sequences (42%), including *Solanum abutiloides* and *Petunia hybrida*, and a clade of locus 3 sequences (26%). Each clade includes the relevant sequences for both *Petunia hybrida* and *Solanum abutiloides*. These clades were used to label the putative locus for each *rbcS* copy (*). Other copies were identified to a locus by structural characteristics (locus 1: single copy with two introns, locus 2: copy with three introns, locus 3: tandemly repeated copies with two introns).

The ML tree inferred from the mature peptide portion of *rbcS* (Figure 3) reveals a topology similar to the analysis from the full peptide (Figure 1) such that paralogous sequences

often form clades. The sequences from *Petunia hybrida* form a clade (100%bp) sister to the remaining sequences. All sequences from *Solanum abutiloides* form a clade (99%) nested within a clade containing locus 2 sequences from *Solanum* and *Jaltomata* species (50%). Two clades of *Nicotiana* sequences for locus 2 (94%) and locus 3 (77%) are sister to each other (59%). The *Nicotiana* clade for both loci is placed within a poorly-supported clade of locus 2 sequences from *Solanum* and *Jaltomata* and all the *Solanum abutiloides* sequences (25%). The locus 3 and locus 1 sequences of *Solanum* and related genera form a clade (40%). Within this clade numerous species-specific clades are formed. Locus 3 sequences form a *Jaltomata grandiflora* clade (87%), a *Solanum tuberosum* clade (44%), a *Solanum lycopersicum* clade (68%), and a *Solanum appendiculatum* clade (71%). There is a locus 1 clade (45%) that includes sequences from four species of *Solanum*, but does not include other putative locus 1 sequences (*i.e.*, locus 1 sequences identified using the transit peptide (Figure 2): *Solanum herculeum* 00.08 1*, *S. phaseoloides* 00.20 1*, *S appendiculatum* 4 1*, *S. abutiloides* 06.19 1*, *S. dulcamara* 04.12 1*, *S. appendiculatum* 3 1*).

The NJ tree inferred from synonymous distances between sequences for the mature peptide portion of *rbcS* (Figure 4) reveals a general pattern where orthologous sequences form clades. Locus 2 sequences from *Solanum* and *Nicotiana* form a clade (28%). Two sub-clades are included within: all *Nicotiana* 2 sequences (92%) and all *Solanum* 2 sequences including paralogs from *S. abutiloides* (52%). Locus 1 and locus 3 sequences from *Solanum*, *Jaltomata grandiflora*, and *Capsicum anuum* form a clade (44%).

The tree was rooted with *Petunia hybrida* locus 3 sequences to increase readability, but this clade could be grouped with the clade of *Nicotiana* locus 3 sequences (76%). The other two loci from *Petunia hybrida* form a clade (61%), as do the remaining sequences (65%) (Figure 4).

The NJ tree inferred from nonsynonymous distances between sequences for the mature peptide portion of *rbcS* (Figure 5) reveals a pattern of concerted evolution. *Petunia hybrida* sequences form a clade (100%) that was used to root the remaining sequences. A clade of all the *S. abutiloides* sequences (98%) is contained within a clade of locus 2 sequences from *Solanum* (46%).

Topology testing

Tree topologies from the two functional regions of *rbcS* were statistically different ($p < 0.0001$) and each alternative topology was rejected from the other by the approximately unbiased test (Table 1).

Testing for recombination between sequences

Two methods were used to screen the aligned data for signals of recombination, GENECONV (Sawyer 1999) and the SBR/GARD method (Pond et al. 2006) for recombination detection as implemented on the Datamonkey webserver (Pond and Frost 2005). In the alignment for the full coding sequence for *rbcS*, SBR and GARD both identified a breakpoint near the junction between the nucleotides encoding the transit sequence peptide and the mature peptide (100% support for cAIC, BIC). Neither method found support for further breakpoints when applied to only the coding sequence for the mature peptide.

GENECONV identified recombination between sequences from locus 3. On the full *rbcS* coding alignment, GENECONV identified no potential recombination. However, when the same analysis was conducted on genera specific groups of sequences and species specific groups of

sequences an inner fragment was identified starting in the third codon position of the 9th codon from the end of the transit sequence through all the codons for the mature peptide sequence between *Solanum tuberosum* 3a and 3c sequences (simulated P value = 0.015). Analyses of the alignment for the coding sequence of the mature peptide grouped by species identified an inner fragment for the full length of the coding sequence except for the last 25 codons between *Solanum abutiloides* 3.1 and 3.2 sequences (simulated P value = 0.027). GENCONV does not identify fragments when sequences are identical so many locus 3 sequences could not be tested.

Codon usage

ENc is a general measure of codon bias with values that represent the number of equally used codons that would generate the same codon usage bias as the one observed. A common method to distinguish between codon bias and mutation bias is to compare the ENc (Wright 1990) to GC3 content. ENc and GC3 values and four indices for nucleotide composition at third codon position synonymous sites (G3, C3, A3, and T3) were calculated using CodonW (Peden 2000). Each third codon position synonymous site nucleotide index is the frequency observed proportional to the maximum possible usage of that nucleotide without changing amino acid composition (Peden 2000). Expected ENc (eENc) for a given GC3 content was calculated for each sequence along with the difference between the observed and expected values with respect to the expected values. When ENc approaches eENc values a codon bias is determined by the underlying mutational bias and without selection on codon usage.

Values for ENc range between 37 – 55, average GC between 45-51% content (Supplementary Table 4). Additional values of nucleotide composition at the 3rd codon position

for each sequence, as well as the four nucleotide indices (e.g., A3), and ENc and eENc with the difference between the two calculated as a percentage of eENc were calculated (Supplementary Table 4). Large differences between ENc and eENc implicate selection for codon usage affecting sequences. For *rbcS* sequences, (ENc – eENc)/eENc values ranged between 0.06 – 0.36 (Supplementary Table 4).

To further examine the level of selection on codon bias, an ML based codon substitution model, FMutSel, that models selection on synonymous codons was supported and the null model rejected with a p value < 0.0001 for both the HKY and GTR nucleotide substitutions models and dN/dS rate category models M0 (1 rate) and M3 (3 rates) (Table 4). Parameter estimates for proportion of negatively selected sites with selection coefficient greater than 2 (P-_|S| > 2) and average selection coefficient for negatively selected sites (mean S-) are listed in Table 2. Estimates differ more by nucleotide substitution model (GTR vs HKY) than between the 1 dN/dS rate category model (M0) and the discrete rate model with three dN/dS estimates (M3).

For codon bias to drive the pattern of concerted evolution, different codons must be favored in separate lineages. To assess differences in codon usage bias, pairwise distances were calculated for codon usage in each sequence and clustered using the neighbor-joining method (Figure 4). Many sequences cluster by codon usage with other orthologous copies: sequences from locus 3 in *Petunia hybrida* cluster with a locus 3 sequence from *Nicotiana tabacum*, locus 3 sequences from *Solanum* species cluster together, locus 2 sequences form a cluster of sequences from *Nicotiana* and *Solanum* species and another cluster of sequences from *Solanum tuberosum* and *S. lycopersicum*, locus 1 sequences from three *Solanum* species cluster together. Notable paralogous groupings are also evident: *Jaltomata grandiflora* locus 3 sequences form a cluster sister to the locus 1 sequence (Jgrandiflora_5, identified using the transit peptide (Figure 2)),

locus 1 and locus 2 sequences from *Petunia hybrida* cluster and are sister to a cluster of locus 3 sequences from *Nicotiana* species, the transit peptide identified (Figure 2) locus 1 sequence (Sabutiloides_06.19) from *S. abutiloides* clusters with other *Solanum* locus 3 sequences, the locus 1 sequence from *Solanum phaseoloides* is sister to the locus 2 sequence from *S. abutiloides* and both are nested within a larger cluster of locus 2 sequences from *Solanum* and *Nicotiana* species, and locus 1 and locus 3 sequences from *Solanum dulcamara* cluster together.

Separation of synonymous and nonsynonymous substitutions

To investigate whether the unexpected similarity between *rbcS* sequences from *Solanum abutiloides* could have been affected by selection for the same amino acid sequence, synonymous (Fig. 4) and nonsynonymous (Fig. 5) neighbor joining trees were reconstructed separately. The synonymous substitution tree shows a pattern with orthologs generally clustering together. Locus 1 sequences from *Solanum* species, identified from the ML tree of the transit sequence (Fig. 1), form a monophyletic group (34%) nested within sequences of locus 3 from *Solanum*, *Jaltomata*, and *Capsicum*. Locus 3 sequences from *Petunia* form a monophyletic group (100%) and locus 3 sequences from both *Nicotiana* species form another monophyletic group (76%). Locus 2 sequences from *Solanum* and *Jaltomata* species are monophyletic (52%) and sister to the locus 1 and locus 3 sequences from *Solanum abutiloides* (52%). Locus 2 sequences from the two *Nicotiana* species are monophyletic (92%).

The neighbor joining tree of nonsynonymous substitutions shows a trend of paralogs clustering together (Fig. 6). Paralogs from *Solanum abutiloides* form a monophyletic group (98%) within a cluster of locus 2 sequences from other *Solanum* species. Paralogs from *S.*

muricatum form a monophyletic group (60%). Locus 1 and locus 3 sequences from *Jaltomata grandiflora* form a monophyletic group (46%). Both sequences from *Capsicum annuum* cluster together (100%). Paralogs from each *Nicotiana* species form monophyletic sister groups: *N. tabacum* sequences (18%), *N. benthamania* (17%).

Positive selection

Topological incongruence between the synonymous substitutions tree and the nonsynonymous substitutions prompted an examination for positive selection. A test for positive selection was performed using the ML tree from the mature peptide coding sequence (Figure 3). The nonsynonymous/synonymous substitution rate ratio (dN/dS) is used to indicate natural selection at the protein level when dN/dS > 1, neutral evolution when dN/dS = 1, and purifying selection when dN/dS < 1. The results of tests for positive selection under different models are summarized in Table 3. The Branch Site model A test implemented in PAML estimated a dN/dS value of 217.17 for the branch in Figure 3 shared by all paralogs in *Solanum abutiloides* ($p < 0.001$). High BEB support for positive selection implicated four codons from the region encoding the mature pepetide (Table 3): position 75 (0.763 BEB), position 80 (0.926 BEB), position 81 (0.999 BEB), position 118 (0.898 BEB).

Other evolutionary models, differing in parameterization for dN/dS rate variation, test for positive selection across all sites. Two comparisons indicate no improvement in fit when selection is estimated (Table 3). The first comparison was between the Nearly Neutral model (M1a), which models rates among sites as a portion under purifying selection ($0<dN/dS<1$) and under neutral evolution (dN/dS=1), and the Positive Selection model (M2a). The two additional parameters in M2a, which estimate a proportion of sites evolving under positive selection

(dN/dS>1), results in no improvement in fit (LRT: 0.00; p = 1.00). A second comparison of nested models evaluates a model incorporating dN/dS rates from a Beta distribution (M7). The additional parameters to model positive selection resulted in no better fit (LR: 0.007; p = 1.00).

Discussion

We have developed a framework to examine the mechanisms of concerted evolution by sequencing the three *rbcS* loci in species of increasing phylogenetic distance from tomato and potato. Our phylogenetic approach identified significant evidence of concerted evolution in *Solanum abutiloides*, a species of moderate genetic distance from tomato and potato, while copies in more distantly related species do not exhibit this pattern among all loci. We speculate that we have uncovered evidence of concerted evolution that is either specific to the *S. abutiloides* lineage or has simply progressed to a degree that allows identification.

The homogenization pattern among *rbcS* copies in *Solanum abutiloides* supports the selection hypothesis (Pichersky et al. 1986). Two results are important for this conclusion, 1) a test for positive selection identifies shared nonsynonymous codons in exon 3 of all three *S. abutiloides* loci that contribute to an estimated d_N/d_S ratio > 1 along the branch leading to these copies, and 2) separate phylogenetic analyses of synonymous and non-synonymous substitutions indicates the latter to be driving the similarity between the locus 2 sequence and its paralogs in *S. abutiloides*. Since non-synonymous substitutions encode the differences in amino acid sequence upon which selection acts, we infer that selection is the mechanism driving homogenization between copies in this lineage.

No support was found for the gene conversion hypothesis (Meagher et al. 1989) acting between paralogous loci, although, gene conversion is supported between tandemly repeated copies of *rbcS* at locus 3 in *Solanum lycopersicum*, *S. tuberosum*, and *S. abutiloides*. Scans of the alignment of *rbcS* sequences for significant clustering of silent substitutions failed to identify any regions of suspected gene conversion between sequences from different loci. Recombination break point analysis uncovered breaks between the transit and mature portions of the *rbcS* peptide. This breakpoint signal can reasonably be attributed to differences in selective constraints. There was no signal of recombination within the mature peptide coding region between loci.

Other potential mechanisms for homogenization were not supported. ENc indices and a ML codon analysis support a general codon bias within the coding sequence for the mature peptide, however, the bias appears to be more similar for orthologous loci than between paralogous loci as demonstrated by codon usage pairwise distances that generally resulted in clusters of orthologous clades. For example, codon usage for the locus 2 sequence in *Solanum abutiloides* (Sabutiloides_rbcS_2) is much more similar to locus 2 sequences from other *Solanum* species than to its own paralogs at other loci (Figure 6). If codon usage bias was driving concerted evolution then sequences from different loci should cluster together by species. No paralog clades were inferred from codon usage similarity analysis. Despite support for selection upon codon usage and strong departure from equal codon usage, similarity in codon usage is higher between sequences at the same locus in different lineages than among copies at different loci within a particular lineage. Selection does affect synonymous codons usage among *rbcS* sequences but results in biases that are more similar between orthologs and does not contribute to homogenization between paralogs in a species.

Recombination and codon bias are two mechanisms shown to be acting on *rbcS* sequences that can mislead subsequent analyses. Recombination can cause errors in analyses of detection for positive selection (Anisimova et al. 2003). For this *rbcS* dataset, evidence for recombination is limited to the tandemly repeated copies at locus 3. One breakpoint was inferred within the coding sequences between the transit and mature peptides and all subsequent analyses relied only on the coding region for the mature peptide. Furthermore, rates for synonymous substitutions can be reduced in highly expressed genes (Sharp and Li 1987) and *rbcS* exhibits a codon bias that constrains estimates for both of the rates in the dN/dS ratio. Comparisons between dN and dS are still valid when selection acts on synonymous substitutions, however, because the ratio is a contrast between the two rates before and after selection acts on the protein sequence (Yang and Nielsen 2008). Additionally, codon usage bias is not species specific, orthologous sequences are generally more similar in codon usage than paralogous sequences and a separate analysis of synonymous and nonsynonymous substitutions results in different topologies. The nonsynonymous substitutions support the grouping of paralogs in *Solanum abutiloides* and thus support the conclusion that selection on the amino acid sequence is the mechanism driving the pattern of concerted evolution.

Amino acids homogenized by positive selection among copies from *Solanum abutiloides* may be adaptive. The four residues with highest posterior probability all reside in exon 3 (Table 3). Of these four residues, the middle two are part of the **α**-helix B structure. The contribution of the small subunit **α**-helix B in the functioning of rubisco remains unknown, but the structure determines rubisco aggregation in the pyrenoids within algae (Meyer et al. 2012).

The transit portion and the mature peptide portion of the *rbcS* coding region provide contrasting evidence for the source of the unique, extra intron locus (locus 2) in Solanaceae and

may implicate ancestral recombination between loci. The transit sequence for locus 2 is more closely related to locus 3 (the tandem repeat locus) than locus 1 (the singleton locus) (Figure 2). While substitution trends in the coding region of the mature peptide indicate a closer relationship between locus 1 and locus 3 copies (Figure 3). However, *Petunia* sequences display more similarity between locus 1 and locus 2 than either loci do to locus 3 copies (Figures 2 and 3). These observations could indicate a relaxation of purifying selection at locus 2 such that copies independently acquire more changes, a common indicator of subfunctionalization.

This study demonstrates an aspect of multigene family evolution that should worry researchers using gene copies to infer species phylogeny. As depicted in the tree figures, orthologous sequences from different loci can lead to inferences of differing species relationships due to the differing rates of concerted evolution between species. Gene tree, species tree discordance has been well described (Maddison 1997; Slowinski and Page 1999)(Maddison 1997), but among gene copies undergoing sporadic concerted evolution the propensity for discordant topologies is virtually guaranteed.

Selection is inferred to act upon *rbcS* sequences. Purifying or negative selection plays a role in maintaining sequence similarity, although it is unable to homogenize paralogs. Selection on codon usage must also be acting on *rbcS* gene sequences, but no evidence was found to suggest it as a force for homogenization. Positive selection is supported as a mechanism acting to homogenize the paralogs within *Solanum abutiloides* and may also be important for increasing similarities among paralogs in other related lineages. The differences between gene trees inferred using only synonymous substitutions and only nonsynonymous substitutions support the selection hypothesis since only the latter infer close relationships between *rbcS* paralogs within a species.

While gene conversion remains an appealing hypothesis to explain patterns of concerted evolution, the role of selection in homogenization should not be relegated to an afterthought. Strong negative selection combined with selection for codon usage may decrease the rate of synonymous substitution to such a degree that may obscure all but the strongest signals of positive selection.

Literature Cited

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. Genetics 164:1229–1236.

Benovoy, D., and G. Drouin. 2009. Ectopic gene conversions in the human genome. Genomics 93:27–32.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. GenBank. Nucleic Acids Res. gks1195.

Berglund, J., K. S. Pollard, and M. T. Webster. 2009. Hotspots of Biased Nucleotide Substitutions in Human Genes. PLoS Biol 7:e1000026.

Bombarely, A., N. Menda, I. Y. Tecle, R. M. Buels, S. Strickler, T. Fischer-York, A. Pujar, J. Leto, J. Gosselin, and L. A. Mueller. 2011. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res. 39:D1149–D1155.

Brown, D. D., P. C. Wensink, and E. Jordan. 1972. A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. J. Mol. Biol. 63:57–73.

Chua, N.-H., and G. W. Schmidt. 1978. Post-translational transport into intact chloroplasts of a precursor to the small subunit of ribulose-1,5-bisphosphate carboxylase. Proc. Natl. Acad. Sci. 75:6110–6114.

Clegg, M. T., M. P. Cummings, and M. L. Durbin. 1997. The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. 94:7791–7798.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772–772.

Davis, J. J., and G. J. Olsen. 2010. Modal Codon Usage: Assessing the Typical Codon Usage of a Genome. Mol. Biol. Evol. 27:800–810.

Dean, C., P. van den Elzen, S. Tamaki, M. Black, P. Dunsmuir, and J. Bedbrook. 1987. Molecular characterization of the rbcS multi-gene family of Petunia (Mitchell). Mol. Gen. Genet. MGG 206:465–474.

Dean, C., E. Pichersky, and P. Dunsmuir. 1989. Structure, Evolution, and Regulation of RbcS Genes in Higher Plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 40:415–439.

Doyle, J., and J. Doyle. 1987. Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem Bull 19:11–15.

Drouin, G. 2002. Characterization of the Gene Conversions Between the Multigene Family Members of the Yeast Genome. J. Mol. Evol. 55:14–23.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12:640–649.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. 96:4482–4487.

Eickbush, T. H., and D. G. Eickbush. 2007. Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. Genetics 175:477–485.

Eirín-López, J. M., A. M. González-Tizón, A. Martínez, and J. Méndez. 2004. Birth-and-Death Evolution with Strong Purifying Selection in the Histone H1 Multigene Family and the Origin of orphon H1 Genes. Mol. Biol. Evol. 21:1992–2003.

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Hershberg, R., and D. A. Petrov. 2008. Selection on Codon Bias. Annu. Rev. Genet. 42:287–299.

Hood, L., J. H. Campbell, and S. C. R. Elgin. 1975. The Organization, Expression, and Evolution of Antibody Genes and Other Multigene Families. Annu. Rev. Genet. 9:305–353.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13–34.

Ingvarsson, P. K. 2008. Molecular evolution of synonymous codon usage in Populus. BMC Evol. Biol. 8:307.

Kawabe, A., and N. T. Miyashita. 2003. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet. Syst. 78:343–352.

Lin, Y.-S., J. K. Byrnes, J.-K. Hwang, and W.-H. Li. 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc. Natl. Acad. Sci. 103:14412–14416.

Lu, P., X. Han, J. Qi, J. Yang, A. J. Wijeratne, T. Li, and H. Ma. 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. 22:508–518.

Maddison, W. P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Meagher, R. B., S. Berry-Lowe, and K. Rice. 1989. Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. Genetics 123:845–863.

Meyer, M. T., T. Genkov, J. N. Skepper, J. Jouhet, M. C. Mitchell, R. J. Spreitzer, and H. Griffiths. 2012. Rubisco small-subunit α-helices control pyrenoid formation in Chlamydomonas. Proc. Natl. Acad. Sci. 109:19474–19479.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst. Biol. 52:674–683.

Mishkind, M. L., S. R. Wessler, and G. W. Schmidt. 1985. Functional determinants in transit sequences: import and partial maturation by vascular plant chloroplasts of the ribulose-1,5-bisphosphate carboxylase small subunit of Chlamydomonas. J. Cell Biol. 100:226–234.

Mondragon-Palomino, M., and B. S. Gaut. 2005. Gene Conversion and the Evolution of Three Leucine-Rich Repeat Gene Families in Arabidopsis thaliana. Mol. Biol. Evol. 22:2444–2456.

Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. 94:7799–7806.

Nei, M., I. B. Rogozin, and H. Piontkivska. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc. Natl. Acad. Sci. 97:10866–10871.

O'Brien, J. D., V. N. Minin, and M. A. Suchard. 2009. Learning to Count: Robust Estimates for Labeled Distances between Molecular Sequences. Mol. Biol. Evol. 26:801–814.

Ohta, T. 1983. On the evolution of multigene families. Theor. Popul. Biol. 23:216–240.

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

Pâques, F., and J. E. Haber. 1999. Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. 63:349–404.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Peden, J. F. 2000. Analysis of codon usage. Citeseer.

Pichersky, E., R. Bernatzky, S. D. Tanksley, and A. R. Cashmore. 1986. Evidence for selection as a mechanism in the concerted evolution of Lycopersicon esculentum (tomato) genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. Proc. Natl. Acad. Sci. U. S. A. 83:3880–3884.

Pond, S. L. K., and S. D. W. Frost. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533.

Pond, S. L. K., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098.

Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. 98:13757–13762.

Qiu, S., R. Bergero, K. Zeng, and D. Charlesworth. 2011. Patterns of Codon Usage Bias in Silene latifolia. Mol. Biol. Evol. 28:771–780.

Rooney, A. P., H. Piontkivska, and M. Nei. 2002. Molecular Evolution of the Nontandemly Repeated Genes of the Histone 3 Multigene Family. Mol. Biol. Evol. 19:68–75.

Sawchuk, M. G., T. J. Donner, P. Head, and E. Scarpella. 2008. Unique and Overlapping Expression Patterns among Members of Photosynthesis-Associated Nuclear Gene Families in Arabidopsis. Plant Physiol. 148:1908–1924.

Sawyer, S. 1999. GENECONV: A computer package for the statistical detection of gene conversion. Distrib. Author Dep. Math. Wash. Univ. St Louis.

Schneider, G., Y. Lindqvist, and C. I. Branden. 1992. Rubisco: Structure and Mechanism. Annu. Rev. Biophys. Biomol. Struct. 21:119–143.

Sharp, P. M., and W. H. Li. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4:222–230.

Shimodaira, H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. 51:492–508.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Slowinski, J., and R. Page. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48:814–825.

Smith, G. P. 1976. Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535.

Spreitzer, R. J., and M. E. Salvucci. 2002. RUBISCO: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme. Annu. Rev. Plant Biol. 53:449–475.

Sugita, M., and W. Gruissem. 1987. Developmental, organ-specific, and light-dependent expression of the tomato ribulose-1,5-bisphosphate carboxylase small subunit gene family. Proc. Natl. Acad. Sci. 84:7104–7108.

Sugita, M., T. Manzara, E. Pichersky, A. Cashmore, and W. Gruissem. 1987. Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. Mol. Gen. Genet. MGG 209:247–256.

Sukumaran, J., and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Swofford, D. L. 2002. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland MA Sinauer Assoc.

Teshima, K. M., and H. Innan. 2004. The Effect of Gene Conversion on the Divergence Between Duplicated Genes. Genetics 166:1553–1560.

Turner, N. E., W. G. Clark, G. J. Tabor, C. M. Hironaka, R. T. Fraley, and D. M. Shah. 1986. The genes encoding the small subunit of ribulose-l,5-bisphosphate carboxylase are expressed differentially in petunia leaves. Nucleic Acids Res. 14:3325–3342.

Wang, X., H. Tang, J. E. Bowers, F. A. Feltus, and A. H. Paterson. 2007. Extensive Concerted Evolution of Rice Paralogs and the Road to Regaining Independence. Genetics 177:1753–1763.

Weese, T. L., and L. Bohs. 2007. A three-gene phylogeny of the genus Solanum (Solanaceae). Syst. Bot. 32:445–463.

Wright, F. 1990. The “effective number of codons” used in a gene. Gene 87:23–29.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z., and R. Nielsen. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. Mol. Biol. Evol. 25:568–579.

Zhang, G., M. Hubalewska, and Z. Ignatova. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. 16:274–280.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol. Biol. Evol. 22:2472–2479.

Zickler, D., and N. Kleckner. 1999. MEIOTIC CHROMOSOMES: Integrating Structure and Function. Annu. Rev. Genet. 33:603–754.

Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proc. Natl. Acad. Sci. 77:2158–2162.

Zwickl, D. 2006. GARLI—genetic algorithm for rapid likelihood inference. See Httpwww Bio Utexas Edufacultyantisensegarligarli Html.

Supplementary Table 1. Voucher information for study taxa.

Taxa	Authority	Plant info/Voucher
Solanum abutiloides	(Griseb.) Bitter & Lillo	R.G. Olmstead S-73 (WTU)
Solanum appendiculatum	Dunal	Mexico, G. Anderson 1401 (CONN)
Solanum dulcamara	L.	Ann Arbor, Michigan, USA, no voucher
Solanum herculeum	Bohs	Morocco, Jury 13742 (WTU)
Solanum jasminoides	Paxton	Bogota, Estrada, R.G.Olmstead S-86 (WTU)
Solanum lycopersicum	L.	Territorial Seed#TM784, Var=sweetie cherry tomato, Pat Reeves 23
Solanum muricatum	Aiton	G. Anderson 1461 (CONN)
Solanum phaseoloides	Pol.	donated by Lynn Bohs
Solanum tuberosum	L.	USW 1718 ?????
Jaltomata grandiflora	(Robinson and Greenmann) D'Arcy, Mione & Davis	Mexico, Michoacan, Davis 1114 (MO)

Supplementary Table 2. Primers designed for PCR amplification of *rbcS* copies.

Primer	Location	Locus	primer sequence, 5'-3'
1	exon 1	all	CAATGGCTTCCTCwrTnnTnTCCTC
2	exon 3b	all	GGCTTGTArGCrATGAAACTGATrC
3	exon 1	3	CCCTGTTTCAAGGAAGCAAAACC
4	exon 1	3	GGACTTrAGkCCAGTGAAGGG
5	exon 1	all	AACCTTGACATTACyTCCmTTGC
6	exon 3b	all	ATGAAACTGATrCACTGCACTTGACG
7	5' UTR	2	GATTAmyGAGGTGCTTACACG
8	exon 3b	2	CCCTTCTGGCTTGTAGGC
9	5' UTR	2	AATTGTATAATGTTATCAAGAACCAC
10	exon 3b/3' UTR	2	TCCTAATATGAAACTTAGTAkCCTTC

Supplementary Table 3. Collection numbers, primers and clones for each *rbcS* sequence.

Genus	species	locus	name	primer F	primer R	clones	Identical to	ref	synonyms
Solanum	*abutiloides*	1	06.19	9	10	D			
		2	2.1	13	16	6			
		3	3.1	1	2	23			
				9	10	D			
				5	6	1			
		3	3.2	1	2	14			
Solanum	*appendiculatum*	1	3	9	10	3			
		1	4	9	10	2			
		3	1	9	10	5			
				5	6	1			
		3	2	9	10	5			
Solanum	*dulcamara*	1	04.2	1	2	23			
				9	10	4			
		2	2.1	15	16	13			
		3	04.1	1	2	23			
Solanum	*herculeum*	1	00.08	9	10	6			
Solanum	*jasminoides*	2	1	13	16	6			
			2	13	16	7			
Solanum	*lycopersicum*	1	X05982					Sugita et al. 1987	
		2	X05983					Sugita et al. 1987	
		3a	M13544				X05984	Pichersky et al. 1986	
		3b	D11112				X05985	Sugita et al. 1987	
		3c	X05986					Sugita et al. 1987	
Solanum	*muricatum*	1	1	9	10	4			
		2	2	13	16	25			
Solanum	*phaseoloides*	1	00.20	9	10	5			
		2	2	13	14	23			
Solanum	*tuberosum*	1	X69763					Fritz et al. 1993	
		2	1	13	16	4			
		2	2	13	16	5	X69759	Fritz et al. 1993	
		3a	X69760					Fritz et al. 1993	
		3b	X69761					Fritz et al. 1993	
		3c	X69762					Fritz et al. 1993	
Jaltomata	*grandiflora*	1	5	9	10	4			
		2	5	9	10	6			
		3	1	9	10	5			
		3	2	9	10	18			
		3	3	9	10	9			
		3	4	9	10	2			
Capsicum	*annuum*	1	SGNU19 6104					Bombarely et al. 2011	

		3	SGNU196105	AF065615	Bombarely et al. 2011	
Nicotiana	*benthamania*	2	24803532	25197092, 25267005, 24894358		
		2	24797485	24997047		
		3	25212782			
Nicotiana	*tabaccum*	2	c5644			
		2	c133569			
		2	c21998	X02353		
		3	c103243			TSSU3-2
		3	c126816			
		3	c126805	c182925		
		3	c12773	X53426		NySS41
Petunia	*hybrida*	1	cd611			
		2	X03820		Tumer et al. 1986	ssu301, ssu8
		3	X03821		Tumer et al. 1986, Dean et al. 1987	ssu511, ssu11a
		3	cd231			
		3	cd112			
		3	cd911			
		3	cd491			

Supplementary Table 4. Codon composition at 3rd position and ENc values.
(separate excel doc)

Table 1. Alternative topologies and *P* values from the approximately unbiased test.

Tree	Comparison tree	P value
Transit peptide tree	Transit peptide tree	1.000
	Mature peptide tree	0.000
Mature peptide tree	Transit peptide tree	0.000
	Mature peptide tree	1.000

Table 2. Selection on synonymous substitutions among different models. HKY and GTR nucleotide substitution models with 1 dN/dS rate category (M0) and 3 dN/dS rate categories (M3). Proportion of negatively selected sites with selection coefficient > 2 (P-_|S| > 2) and average selection coefficient for negatively selected sites (mean S-).

Rate model	M	LR	p-value	dN/dS	P-_|S|>2	mean S-
hky	0	257.64	<<0.0001	0.14	0.24551	-1.78
hky	3	232.43	<<0.0001	p: 0.51 0.30 0.19 w: 0.00 0.12 0.74	0.20121	-1.74
gtr	0	267.60	<<0.0001	0.13	0.99797	-9.44
gtr	3	257.87	<<0.0001	p: 0.65 0.33 0.02 w: 0.01 0.35 2.17	0.99973	-11.58

Table 3. Estimates of positive selection using 5 PAML models.

Model	ts/tv	dN/dS	lnL	LR	p-value	Foreground dN/dS	Selected sites
M1a	1.633	p: 0.828 0.172 w: 0.051 1.000	-3586.37				
M2a	1.633	p: 0.828 0.061 0.111 w: 0.051 1.000 1.000	-3586.37	0	1.000		
M7 (beta)	1.370	p: 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 w: 0.000 0.00002 0.00039 0.00251 0.010 0.030 0.076 0.170 0.349 0.685	-3550.45				
M8 (beta&w>1)	1.370	p: 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.100 0.00001 w: 0.000 0.00002 0.00039 0.00251 0.010 0.030 0.076 0.170 0.349 0.685 1.000	-3550.45	0.007	1.000		
Branch sites, model A	1.630			10.88	0.001	217.17	75 F 0.763 80 A 0.926 81 T 0.999 118 Y 0.898

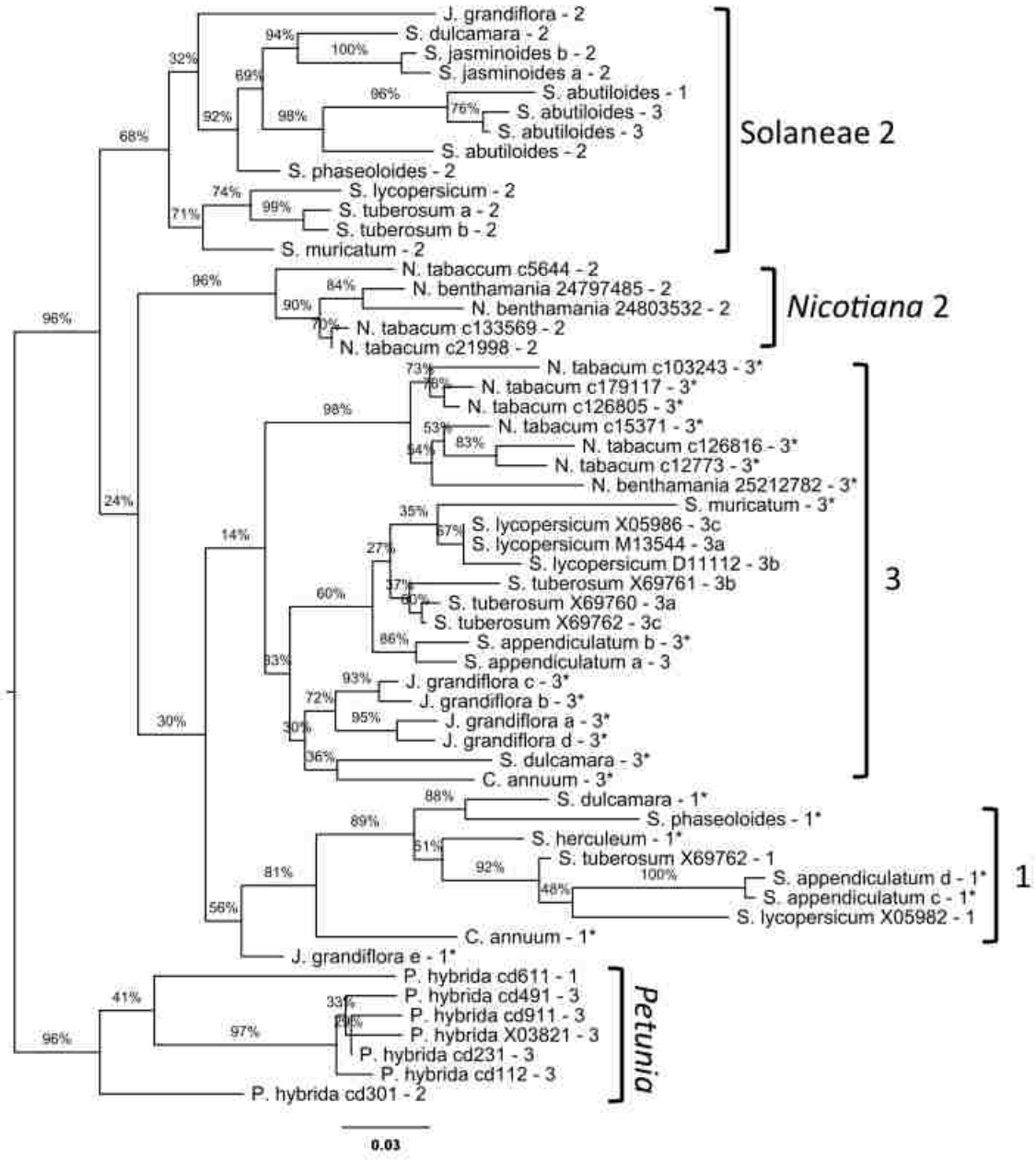


Figure 1. ML tree for the full coding sequence from *rbcS*. Bootstrap support values are shown as percentages above branches. Sequences are labeled with a single letter abbreviation for genus followed by species epithet, sequence identifier, and locus identity. '*' are used to indicate sequences that have been identified based on transit sequence similarity and have not been identified by intron structure or location with other tandem repeats.

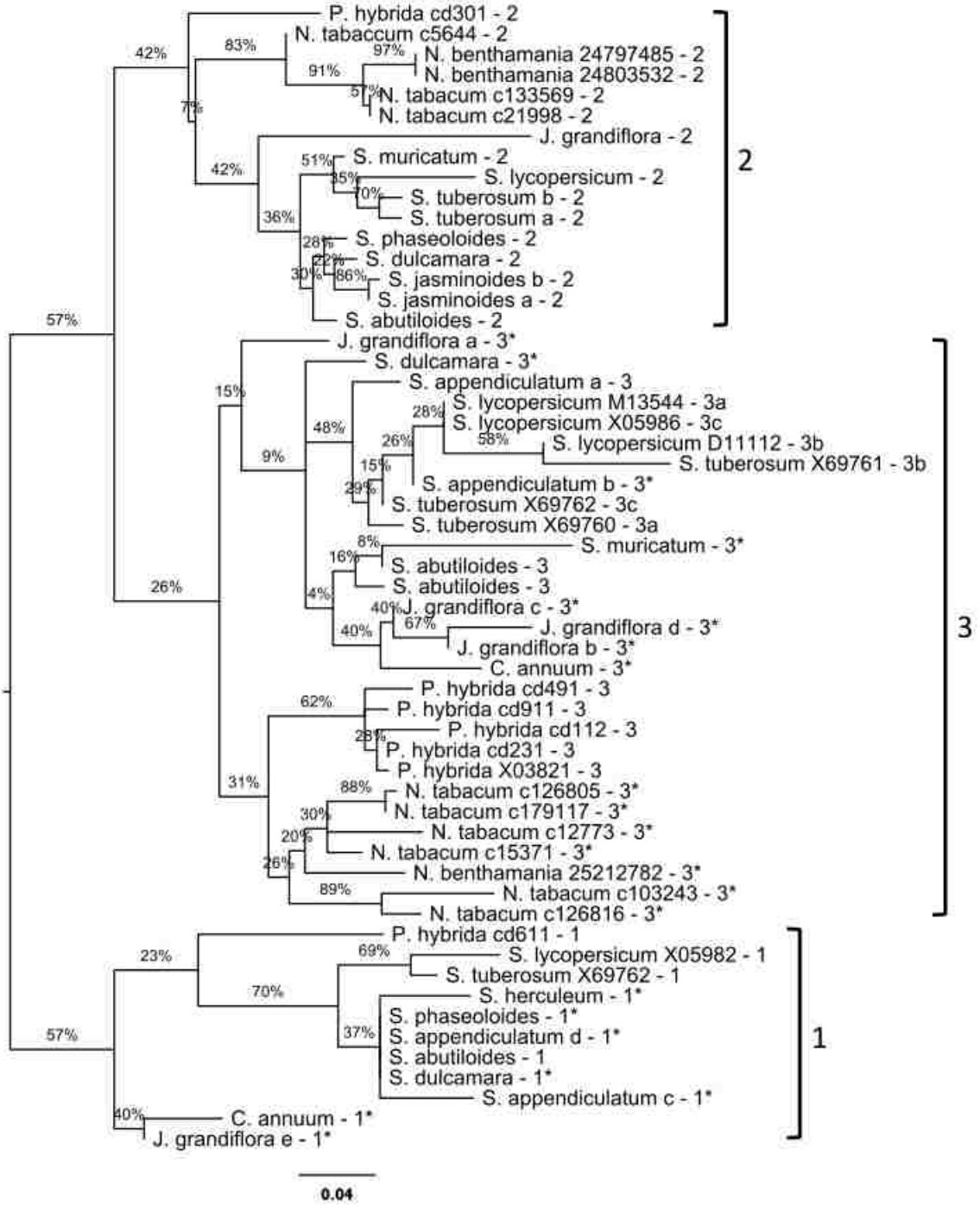


Figure 2. ML tree for the coding sequence of the *rbcS* transit peptide. Bootstrap support values are shown as percentages above branches.

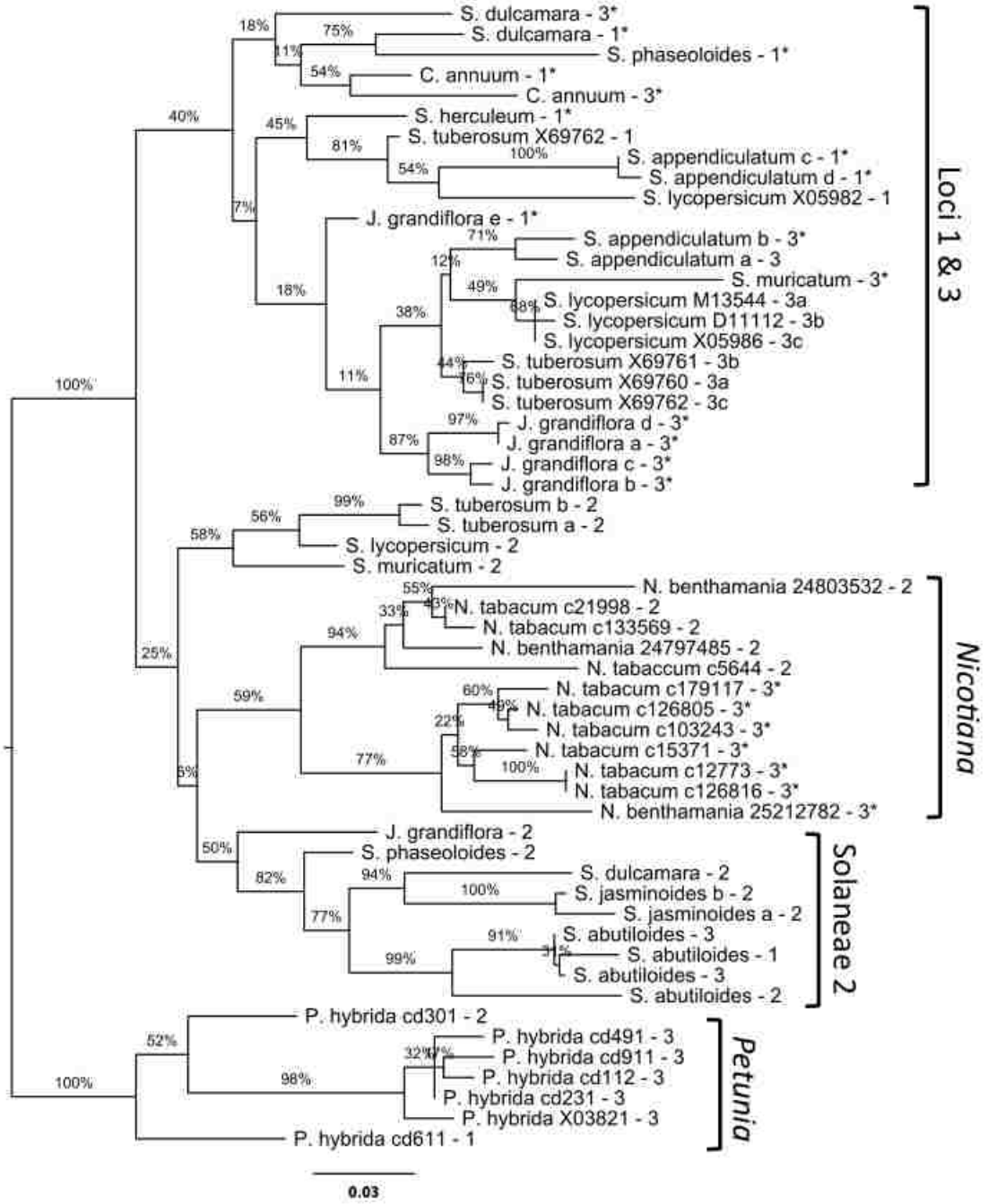


Figure 3. ML tree for the coding sequence of the *rbcS* mature peptide. Bootstrap support values are shown as percentages above branches.

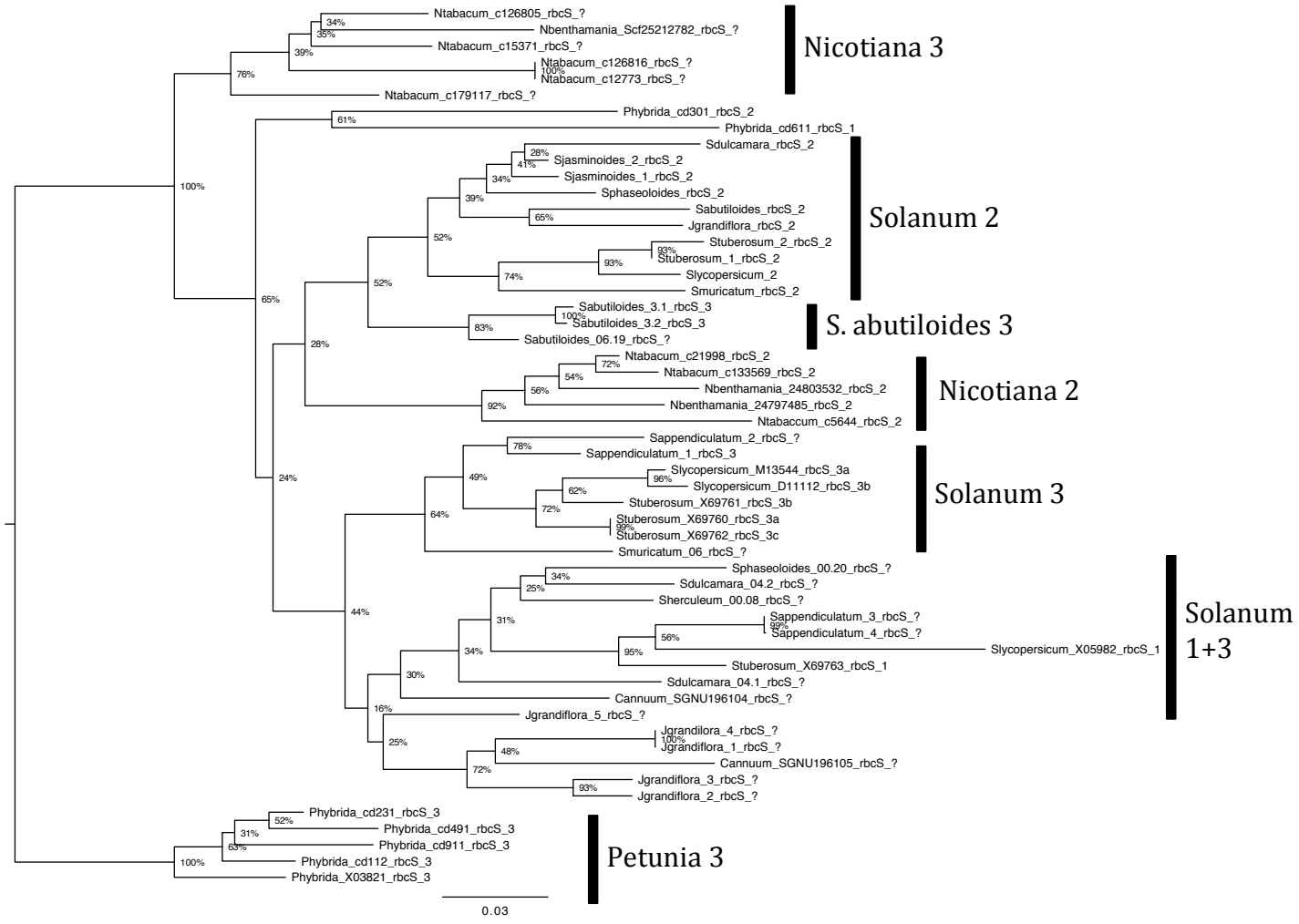


Figure 4. Neighbor joining tree of robust synonymous distances from the coding sequence of the *rbcS* mature peptide. Bootstrap support values are shown as percentages to the right of nodes.

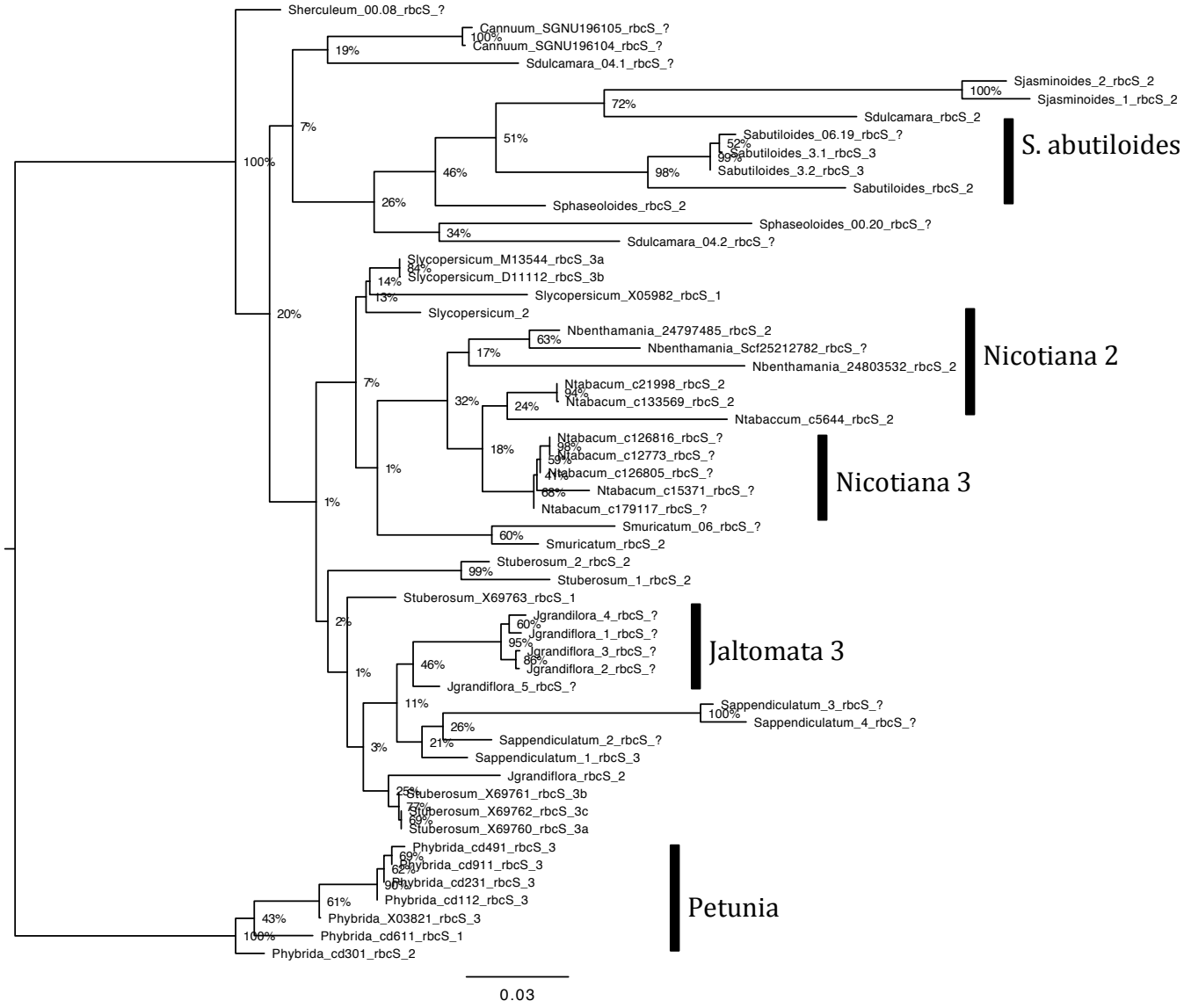


Figure 5. Neighbor joining tree of robust nonsynonymous distances from the coding sequence of the *rbcS* mature peptide. Bootstrap support values are shown as percentages to the right of nodes.

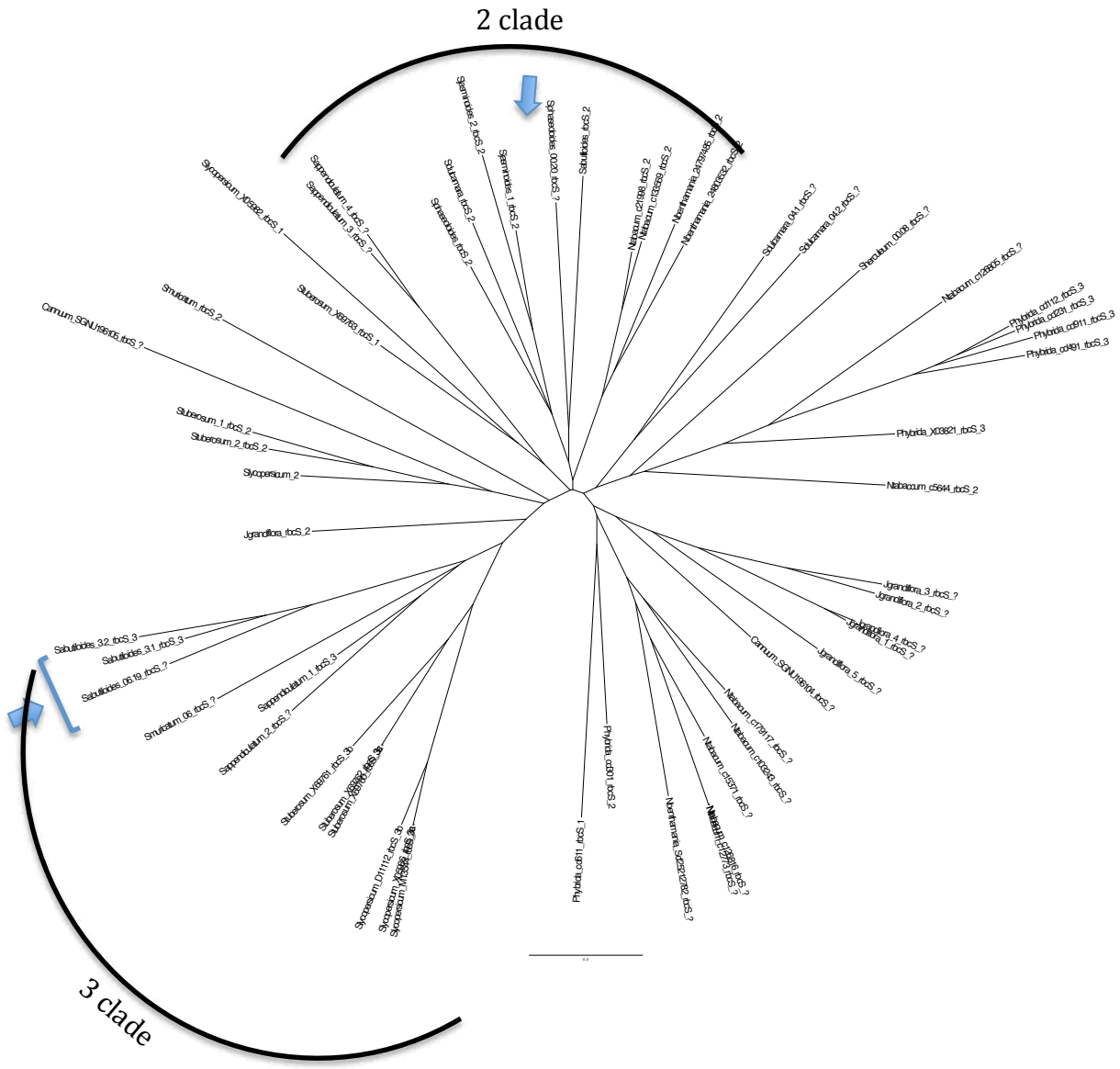


Figure 6. Neighbor joining tree of pairwise distances of codon usage. *Solanum abutiloides* copies are indicated by arrows.

Chapter 2: Evolution of a novel intron in Solanaceae

Introduction

Among land plants and green algae, the gene encoding the small subunit of Rubisco, *rbcS*, is encoded in the nucleus by 2-12 gene copies, with coding regions usually interspersed with two introns in homologous locations (Akazawa et al. 1984). Within Solanaceae, a unique locus of *rbcS* contains a novel, third intron. To explore the origin of this novel intron Solanaceae lineages were sampled for *rbcS* copies to better understand the phylogenetic distribution of species containing this locus. In addition, through comparisons of the novel intron from diverse Solanaceae species, evolutionary processes are identified that may have contributed to the origination of the intron.

Mechanisms of intron gain

Many mechanisms have been suggested to explain the formation of new introns (Yenerall and Zhou 2012). The best-supported mechanism for intron gain is through the tandem duplication of an exonic sequence containing a protosplice site, AGGT (Yenerall and Zhou 2012). In addition to indirect evidence from many different lineages, tandem duplication is the only mechanism that has been shown to generate a functional intron *in vivo* (Hellsten et al. 2011).

Less support exists for other proposed mechanisms of intron gain (Yenerall and Zhou 2012). Transposons may contribute to intron gain after insertion into a protosplice site if preexisting spice sites or ones originating during insertion allow spliceosome removal after transcription to

prevent coding sequence change (Giroux et al. 1994). Preexisting introns from another gene or from paralogous gene copies may also contribute to intron gain by inserting into a protosplice site. Preexisting introns are thought to move to a new gene by transposition of a spliced intron into an mRNA which is subsequently reverse transcribed and recombines with the original gene to cause insertion (Tarrío et al. 1998). Paralogous gene copies may provide another source for preexisting introns to move through gene conversion into a novel location (Hankeln et al. 1997). Introns are also throught to be created through repair of double-strand breaks. Non-homologous end joining may insert novel sequences into genes that by chance can be removed by the spliceosome (Li et al. 2009).

rbcS introns among land plants and green algae

Homologous locations for two introns are conserved for *rbcS* across land plants (Table 1). *rbcS* copies from a moss - *Physcomitrella patens*, lycophyte - *Selaginella moellendorfii* (Banks et al. 2011), fern - *Pteris vittata* (Hanania and Zilberstein 1994) conifer - *Larix laricina* (Hutchison et al. 1990), and numerous flowering plants (Dean et al. 1989) possess introns at shared locations. Within flowering plants, monocots exhibit differential loss of one intron. Within some grasses, the second intron is missing and all copies possess a single intron located at the same first intron position in other land plant *rbcS* genes (Sasanuma and Miyashita 1998). In the duckweed, *Lemna gibba*, *rbcS* copies lack the first intron and contain one intron at the identical position where other land plant *rbcS* copies contain the second intron (Silverthorne et al. 1990).

Among green algae, *rbcS* intron positions tentatively support a single shared splice site for the first intron position found in land plant *rbcS* (Yamazaki et al. 2005). A two-intron *rbcS* copy in *Nannochloris bacillaris* shares the first intron splice site with land plant *rbcS* genes (Yamazaki et al. 2005). Green algae also contain three-intron *rbcS* copies in *N. bacillaris*, *Dunaliella tertiolecta*, and *Chlamydomonas reinhardtii* (all Chlorophyceae, belonging to the most remote branch of green algal phylogeny from land plants; (Leliaert et al. 2012)) that sometimes differ by up to 15 base pairs in intron location; these are not homolous to any land plant intron positions . The absence of introns in cyanobacterial *rbcS* genes, combined with the evidence that some green algae have a third intron (although not in the same exact location as in Solanaceae) led to the proposal that exon shuffling followed by intron loss in some lineages (e.g., monocots) could support the possibility that a three-intron *rbcS* copy arose in the algae-land plant ancestor (Wolter et al. 1988). This scenario requires a shifting in splice sites over time, a homologous relationship between the three-intron algal copy and the one in Solanaceae, and independent loss of the three-intron *rbcS* copy in every other land plant lineage so far examined. Parallel intron gains at identical positions have been found in other genes and may be expected based on the nucleotide preferences for intron splice sites (Tarrío et al. 2003). Thus, a hypotheses of common ancestry for the third intron dating to the earliest ancestor of extant green algae and land plants, seems untenable.

During the evolution of green plants, *rbcS* was transferred from the chloroplast to the nucleus (Palmer 1985). Nuclear *rbcS* copies encode a protein consisting of a transit peptide region and the mature peptide for the small subunit of Rubsico. The transit peptide functions in import into

the chloroplast and is removed from the mature peptide during or after transport (Chua and Schmidt 1978). As with other transit peptides, the transit peptide sequences for *rbcS* vary substantially more than those encoding the mature peptide encoding region, although a nine amino acid segment near the middle is well conserved and is necessary for interaction with components of the chloroplast membrane (Mishkind et al. 1985). The mature peptide encoded by *rbcS* forms the small subunit of Rubisco. Eight small subunits interact with eight large subunits, encoded by the chloroplast gene *rbcL* in the chloroplast, to form the enzyme responsible for CO_2 fixation during photosynthesis, ribulose-1,5-bisphosphate carboxylase/oxygenase (Dean et al. 1989).

rbcS introns in Solanaceae

Within Solanaceae, three-intron *rbcS* copies have been identified from *Solanum*, *Nicotiana*, and *Petunia*. *Solanum* and *Nicotiana* fall within the well-supported "x=12" clade (Olmstead et al. 2008). *Petunia* (Petunieae), along with several other lineages (Cestroideae and Benthamielleae, Goetzeoideae and *Duckeodendron*, *Schizanthus*, Schwenckieae) form a poorly resolved basal grade relative to the X=12 clade (Olmstead et al. 2008; Särkinen et al. 2013).

To circumscribe the phylogenetic origin of the three-intron *rbcS* copy, Solanaceae species in these basal clades were examined for *rbcS* copies by PCR, cloning, and sequencing. DNA sequences from the novel intron in taxa containing the three-intron *rbcS* copy were examined for evidence of processes associated with intron gain.

Evidence for the absence of a specific genetic locus by PCR can be misleading due to divergence at the sites targeted by primers for amplification. However, *rbcS* sequences identified from many different taxa have been shown to undergo concerted evolution such that copies within a taxon are more similar to each other than to copies from other taxa (Pichersky et al. 1986; Meagher et al. 1989; Miller et al. in prep.). The homogenization of *rbcS* copies suggests that PCR primers at coding regions able to amplify one *rbcS* copy from a taxon should also successfully amplify other copies present in the taxon.

Materials and methods

Sequences sampled

Twelve species of Solanaceae were sampled for copies of *rbcS*: *Browallia speciosa*, *Cestrum nocturnum*, *Duckeodendron cestroides*, *Goetzea elegans*, *Jaltomata darcyana*, *Jaltomata paneroi*, *Jaltomata procumbens*, *Jaltomata ventricosa, Nicotiana plumbaginifolia, N. tabacum*, *Schizanthus pinnatus*, *Schwenckia glabrata* (Appendix 1). Available *rbcS* sequences from public databases were downloaded as follows: five genomic sequences from GenBank (www.ncbi.nlm.nih.gov/) (Benson et al. 2012) for both *Solanum lycopersicum* and *S. tuberosum*, and sequences from *Solanum abutiloides*, *S. appendiculatum*, *S. dulcamara*, *S. herculeum*, *S. jasminoides*, *S. muricatum*, *S. phaseoloides* and *Jaltomata grandiflora* (Miller et al. in prep.).

NCBI's BLAST (Altschul et al. 1990) was used with different species-specific cDNA sequences of *rbcS* (GenBank: KC176707, X01722) to identify genomic sequences in *Capsicum annuum* (Jo et al. 2011), *Nicotiana benthamiana* (GenBank: GCA000723945), *N. sylvestris* (GenBank: GCA000393655), and *N. tomentosiformis* (GenBank: ASAG01000000). Resulting sequences were compared to *rbcS* exons to identify intron structure and copies were compared pairwise to identify unique copies. Three *Capsicum annuum* Genome v0.3 scaffolds containing unique and complete copies of *rbcS* were retained and four *Nicotiana sylvestris* unplaced genomic scaffolds containing unique and complete 2-intron and 3-intron *rbcS* copies were included. One full length *Petunia hybrida* unigene from Sol Genomics Network (SGN) (www.solgenomics.net) (Bombarely et al. 2011) and five *Petunia hybrida* sequences from GenBank that were unique and full copies were included in further analysis. From the sister group of Solanaceae, Convolvulaceae, one *Ipomoea batatas* (CB329919) mRNA based sequence from GenBank was also included (Appendix 1).

DNA extraction, PCR, and sequencing

Total genomic DNA was extracted from field-collected, silica-gel dried tissue using the modified 2x CTAB method (Doyle and Doyle 1987) and purified using Wizard minicolumns (Promega, Madison, Wisconsin, USA). The polymerase chain reaction (PCR) was performed to amplify various regions of the nuclear gene for the small subunit of ribulose bisphosphate carboxylase oxygenase (*rbcS*). PCR was conducted in 25 μL volumes with annealing temperatures of 48–60°C using primers from Supplementary Table 1. Amplified *rbcS* products were either gel

isolated or cloned before direct sequencing. For gel isolation, bands were separated on a 1% agarose gel and purified using the Promega Wizard SV Gel and PCR Clean-up System (Promega, Madison, Wisconsin, USA). Otherwise, *rbcS* products were cleaned by precipitation from a 20% polyethylene glycol 8000/NaCl solution and washed with 70% EtOH prior to cloning. PCR products were cloned using the TOPO TA Cloning Kit for Sequencing (Invitrogen, Carlsbad, CA, USA) and 8-48 clones were PCR amplified with vector primers for direct sequencing. Both strands were sequenced using either DYEnamic ET Terminator (GE Healthcare, Piscataway, NJ, USA) or BigDye Terminator v3.1 (Applied Biosystems Inc, Foster City, CA, USA) cycle sequencing kits on ABI model 377 or 3100 automated DNA sequencers. Sequence data were proofed, edited, and contigs assembled using Sequencher v.4.7 (Gene Codes Corp., Ann Arbor, MI, USA). Sequences were grouped by similarity and consensus sequences were constructed from clones that differed by less than 1%.

PCR of mitochondrial sequences

Primers for amplification of the mitochondrial tRNA-Ile gene were designed from gene flanking regions based upon the *Nicotiana tabacum* mitochondrial genome, GenBank accession BA000042 (Supplementary Table 1) and used to amplify mitochondrial sequences from *Jaltomata procumbens* and *Solanum lycopersicum* (Appendix 1).

Phylogenetic analyses

Sequences were manually aligned using the program MacClade 4.08 OSX (Sinauer Associates, Inc, Sunderland, MA, USA). Introns could not be aligned across species nor between loci within the same species and were removed before further analysis. Phylogenetic analyses were performed using a maximum likelihood inference method. DT-Model Select (Minin et al. 2003) and jModelTest 2 (Darriba et al. 2012) were used to determine the best-fit model of sequence evolution. For the full coding region the SYM model (6 rates of nucleotide substitution, equal nucleotide frequencies) with discrete gamma distributed rate variation and a proportion of invariable sites was indicated as the best-fit to the data. The HKY model (2 rates of nucleotide substitution, estimated nucleotide frequencies) with discrete gamma distributed rate variation (Yang 1994) model of sequence evolution was indicated as the best-fit to the data for the coding region for the transit sequence and the TVMef model (variable transversion rates and equal transition rates, equal nucleotide frequencies) with discrete gamma distributed rate variation and a proportion of invariable sites was indicated as the best-fit for the coding region of the mature *rbcS* peptide by both the Akaike Information Criterion and Bayesian Information Criterion. Analyses with partitions of the transit and mature coding regions and without partitions were performed for the designated models and the GTR+I+G model. For partitioned analyses, subset specific rates were estimated (*e.g.*, GARLI setting: subsetspecificrates = 1). AIC scores from each model were compared and the GTR+I+G model, partitioned with subset specific rates, was indicated as the best-fit and used in subsequent analyses (Supplementary Table 2).

Maximum likelihood (ML) analyses were performed using the program GARLI version 2.0 (Zwickl 2006) on the CIPRES Science Gateway teragrid server (Miller et al. 2010). Forty

searches were performed under default parameters (including four categories of discrete approximation of gamma-distributed rate heterogeneity, starting trees created by stepwise-addition with 50 attachment branches evaluated for each taxon, branch-length optimization started at 0.5, reduced 20 times to a minimum of 0.01). To verify convergence, 20 searches were started from stepwise-addition trees and 20 from random trees. Free model parameters were estimated for each search replicate and automatically terminated after 20,000 generations without improvement in the topology score. Bootstrap analysis was conducted with 396 (Felsenstein 1985) replicates and were performed with the same parameters as above, but with two replicates per search and all starting trees created by stepwise-addition. Each pseudoreplicate was automatically stopped after 10,000 generations without improvement in the topology score and bootstrap proportions were calculated by computing a majority rule consensus tree with SumTrees (Sukumaran and Holder 2010).

Trees were rooted with sequences from Convolvulaceae, the sister group to Solanaceae (Olmstead et al. 2000; The Angiosperm Phylogeny Group 2009). Trees were visualized using the program FigTree (Rambaut 2009).

Identification of sequences within intron 3

rbcS intron 3 sequences were compared to intron sequences at other *rbcS* copies by BLAST search and were screened for repetitive elements in Repbase Update (http://www.girinst.org/repbase/) through the Censor web server (Jurka et al. 2005). Potential MITEs were identified using MUST v.1 with default settings except to include clusters with

similarity >0.5 (Chen et al. 2009). Previously identified MITES were found by BLAST search on the P-MITE database (http://pmite.hzau.edu.cn/django/mite/) (Chen et al. 2014).

Repeat sequences within intron 3

Intron 3 sequences were searched for tandem repeats using the Tandem Repeats Database (Gelfand et al. 2007). Repeats of similarity >60% between species were clustered using Tandem Repeat Finder with the CCA algorithm and Eucledian tables (Gelfand et al. 2007). The multi-sequence alignment of tandem repeats from *Jaltomata grandiflora* and *Solanum phaseoloides* was pruned to remove short repeat copies <70 bases in length. The region from *Jaltomata grandiflora* corresponding to the end of exon 3a and flanking intron sequence (residues: 1716-1863) was added to the alignment using MAFFT v.7 (Katoh and Frith 2012). Alignment was manually edited in SeaView v.4 (Gouy et al. 2010). A phylogenetic tree was inferred with the BIONJ algorithm and HKY correction with 200 bootstrap runs and rooted with the repeat added from *Jaltomata grandiflora* (*J. grandiflora* 1) (Gascuel 1997). Inverted repeats were identified within intron 3 using the Inverted Repeat Finder with default alignment parameters (match:2, mismatch: 3, indels: 5), a minimum score of 25, and options chosen for: a third alignment going inward and continuing to search for larger intervals at nearby centers (Gelfand et al. 2007).

Results

rbcS copies from land plants and green algae

Among land plants, *rbcS* introns occur in homologous locations (Table 1). Except for Solanaceae, every lineage includes only 2-intron copies (Table 1). Intron 1 occurs between the 2nd and 3rd amino acids of the mature peptide in phase 0 and intron 2 occurs between the 47th and 48th amino acids of the mature peptide in phase 0. In addition to 2-intron copies, *Solanum*, *Nicotiana*, and *Petunia* have 3-intron *rbcS* sequences. Intron 3 separates the nucleotides of the codon for the 65th amino acid of the mature peptide in phase 2 (i.e., the intron separates the first two bases from the last base of the codon) (Table 1). Previously published estimates for the number of *rbcS* copies within a species range from 2-22 (Table 1). The fern, *Pteris vittata*, contains at least four *rbcS* copies each with two introns (Hanania and Zilberstein 1994). Among monocots, *Lemna gibba* has five copies with one intron at the position homologous to other land plants intron 2 (Silverthorne et al. 1990). Lineages of grasses contain a single intron at a position homologous to other land plants intron 1: at least 12 copies in *Triticum aestivum* (Sasanuma and Miyashita 1998), five copies in *Oryza sativa* (Sakai et al. 2013), and at least 10 copies in *Zea mays* (Dean et al. 1989). Both introns have been identified in six *rbcS* copies from *Musa acuminata* (Thomas-Hall et al. 2007). Within eudicots, lineages within the rosids and asterids have *rbcS* copies with two introns: four copies in *Arabidopsis thaliana* (Lamesch et al. 2011), five copies in *Gossypium hisutum* (Paritosh et al. 2013), and four copies in *Pisum sativum* (Coruzzi et al. 1984). Within Solanaceae 2-intron and 3-intron *rbcS* copies have been identified: four 2-intron copies and one 3-intron copy in *Solanum lycoperiscum* (Sugita et al. 1987), four 2-

intron copies and two 3-intron copies in *Nicotiana sylvestris* (Sierro et al. 2013), and five 2-intron copies and one 3-intron copy in *Petunia hybrida* (Dean et al. 1987).

Green algae possess 2-intron and 3-intron *rbcS* sequences (Table 1). The 2-intron copies in *Nannochloris bacillaris* have a first intron after the 2nd amino acid of the mature peptide in a position homologous to intron 1 in land plants (Yamazaki et al. 2005). The 2-intron *N. bacillaris* copies possess a second intron that splits the codon for the 11th amino acid in phase 2 (Table 1). This intron occurs in a homologous site to the first intron among the green algae 3-intron copies (Yamazaki et al. 2005). The 3-intron copies from *N. bacillaris* and *Chlamydomonas reinhardtii* have intron 2 after the 33rd and in the 37th amino acids of the mature peptide, respectively (Yamazaki et al. 2005). *C. reinhardtii* intron 3 and *Dunaliella teriolecta* intron 2 both occur after the 66th amino acid; intron 3 in *N. bacillaris* occurs after the 69th amino acid. The *D. teriolecta* intron 3 occurs after the 108th amino acid.

rbcS 3-intron copies

Three-intron *rbcS* copies were isolated from *Cestrum nocturnum* with intron positions homologous to the three-intron copies in *Solanum*, *Nicotiana*, and *Petunia*. No other three-intron copies were identified in other species from the unresolved clade of "X=12," Petunieae, Schwenckieae, Benthamielleae, and Cestroideae.

Within Cestroideae, one 2-intron copy and three unique 3-intron copies were identified from *Cestrum nocturnum*. Two of the three 3-intron copies differ at 16 of 1,133 bases and may be alleles or a very recent duplication. In *Browallia speciosa*, a 2-intron copy and two copies lacking intron 1 were identified (Appendix 1). One of the two copies lacking intron 1 contains three stop codons within exon 3 and may represent a pseudogene that would produce a truncated peptide sequence if expressed (*B. speciosa* copy 2b). In Schwenckieae, three 2-intron copies and one copy lacking introns were identified from *Schwenckia glabrata* (Appendix 1). In more divergent lineages at the base of the unresolved clade, a single two-intron copy was found in *Duckeodendron cestroides*, five 2-intron copies were identified from *Goetzea elegans*, and three 2-intron copies identified from *Schizanthus pinnatus* (Appendix 1).

Introns from all *rbcS* copies possess the "GT-AG" consensus splicing signal. Introns 1 and 2 are both phase 0 with the intron positioned between codons. Intron 3 from *Solanum*, *Jaltomata*, *Nicotiana*, *Petunia*, and *Cestrum* are phase 2 with the intron positioned between the second and third positions of the codon.

Phylogenetic analyses of rbcS copies

A phylogenetic tree for the full coding sequence of *rbcS* was inferred by maximum likelihood (Figure 1). One clade of 3-intron *rbcS* copies from *Jaltomata*, *Nicotiana*, and *Solanum* was recovered (20%) containing subclades comprising sequences from each genus: *Jaltomata* (98%), *Nicotiana* (90%), *Solanum* (24%). Three 2-intron *rbcS* copies from *Solanum abutiloides* were

expected due to concerted evolution (Miller et al. in prep.). The 3-intron *rbcS* copies from *Petunia* and *Cestrum* both form clades with 2-intron copies of the same species. Species-specific clades were found for paralogous copies within six species. The three *Cestrum nocturnum* 3-intron copies form a clade (51%) included within a clade of all *Cestrum* copies (43%). *rbcS* copies form clades for *Petunia* (50%), *Browallia speciosa* (97%), and *Schwenckia glabrata* (66%). Copies from *Goetzea elegans* form a clade (88%) sister to the *Duckeodendron cestroides* copy (20%). A clade of copies from *Schizanthus pinnatus* (100%) is resolved at the base of the tree (Figure 1).

A phylogenetic tree of 3-intron *rbcS* sequences and 2-intron copies from species without a 3-intron copy in Solanaceae was inferred by maximum likelihood (Figure 2). A clade of copies from *Schizanthus pinnatus* (99%) is resolved as sister to the remaining copies in Solanaceae (37%). The *Duckeodendron cestroides* copy diverges next with weak support (14%) and is followed by the clade of all *Goetzea* copies (84%). Copies from the X=12 clade (including *Nicotiana*, *Capsicum*, *Jaltomata*, and *Solanum*) form a clade with *Schwenckia*, *Petunia*, *Cestrum*, and *Browallia* copies (28%). *Capsicum annuum* copies form a clade (79%) that diverges first from the remaining copies (13%). Copies from *Jaltomata*, *Solanum*, and *Nicotiana* form a clade (24%) composed of two main subclades: *Nicotiana* (93%) and Solaneae (66%). The Solaneae comprise two subclades: *Jaltomata* (99%) and *Solanum* (36%). *Schwenckia*, *Petunia*, *Cestrum*, and *Browallia* copies form a clade (5%) containing subclades for copies from each genus: *Schwenckia* (100%), *Cestrum* (74%), and *Browallia* (94%) (Figure 2).

A 2-intron copy from *Nicotiana tabacum* (E) is more similar to 3-intron copies than to other 2-intron copies (Figure 1-2). In addition to high coding sequence similarity, both introns in the 2-intron copy share high similarity to the first two introns from other *Nicotiana* 3-intron sequences. An alignment of *N. sylvestris* (KD950247) and *N. tabacum* (E) contained 17 mismatched bases and three gaps.

Phylogenetic analysis of the transit coding sequence, which have been shown previously to exhibit little or no concerted evolution (Miller et al., in prep), identified three primary clades of *rbcS* copies (Figure 3). Clade 1 sequences are all 2-intron copies and include sequences from *Solanum* and *Petunia* that were previously identified as *rbcS* locus 1 copies (Turner et al. 1986; Sugita et al. 1987). Clade 2 sequences are identifiable by the presence of most of the three-intron copies (Figure 3). Clade 3 sequences are mostly 2-intron copies and include sequences from *Solanum*, *Petunia*, and *Nicotiana* that were previously identified as copies from *rbcS* locus 3 (Turner et al. 1986; Sugita et al. 1987; Dean et al. 1989; Chapter 1).

rbcS copies cluster into three clades based on locus with three exceptions (Figure 3). *Goetzea* sequences form a grade of relationships at the base of the tree; the clade of 3-intron sequences (Clade 2) contains a 2-intron sequence from *Nicotiana tabacum* (E) (Figure 3); and, the clade of 2-intron copies (Clade 3) contains the 3-intron copies from *Cestrum* (Figure 3).

Elements identified within rbcS intron 3

No significant BLAST hits for intron 3 sequences and other *rbcS* introns were found. Sequence elements were identified within intron 3 including a full mitochondrial gene and partial copies of transposon sequences (Table 2). Three elements were identified with similarity to different super families of class II transposons (cut and paste transposons) (Feschotte and Pritham 2007): one CACTA-like element (Wicker et al. 2003), one *hAT*-like element (Rubin et al. 2001), and one previously identified MITE belonging to the MULE family (Chen et al. 2014) (Table 2). Four elements were identified with similarity to types of class I transposons (use an RNA intermediary) with long terminal repeats (LTR): three *Ty1-copia*-like elements and one *Ty3-gypsy*-like element (Wilhelm and Wilhelm 2001) (Table 2).

Repeats were identified as either inverted repeats (Table 3) or tandem repeats (Table 4). Three short inverted repeats of 18-24 bases were identified; one pair with 24 inverted bases in *S. muricatum* at position 1587-2219, and two nested pairs of 18 inverted bases from *S. tuberosum* at position 1735-1975 and 1845-1892 (Table 2). One long repeat of 116 bases was found in *Solanum lycopersicum* at position 1623-1869 (Table 3). Tandem repeats within *rbcS* intron 3, which score >100, cluster into two groups: one group occurs only within *Jaltomata* and one is present in both *Solanum* and *Jaltomata* (Table 4). The *Jaltomata* specific tandem repeat contains 2-4 copies of a 94-97 base repeat unit (Table 4). The second tandem repeat group contains a repeating tract of ~170bp that spans the last 23 bases of *rbcS* exon 3a and ~150 bases of flanking intron sequence. The repeat structure and nucleotide sequence are similar between *Solanum* and

Jaltomata, although, *J. grandiflora* contains 3.4 tandemly repeated copies and *S. phaeoloides* contains 2.4 tandemly repeated copies.

The *Solanum* and *Jaltomata* group of tandem repeats were numbered sequentially so that the most 5' unit, which spans the 23 bases at the end of *rbcS* exon 3a and 150 flanking intron bases, is labeled copy 1. Copies 2 and 3 are the tandem repeat units replicated at the 3' end of copy 1. The 2.4 copies in *S. phaseoloides* include one copy with nucleotides within *rbcS* exon 3a (copy 1) and one similar copy tandemly repeated within the intron; the 2.4 copies identified from *J. grandiflora* are all intron located such that the tandem repeat unit lacks the copy spanning part of exon 3a (*i.e.*, copies 2 and 3 were identified) (Table 4). Phylogenetic relationships were determined by neighbor joining for the two full-length copies from *S. phaseoloides* (copies 1-2) and the two full-length copies from *J. grandiflora* (copies 2-3) along with the additional *J. grandiflora* copy encompassing the end of exon 3a (copy 1) (Figure 4). Copies from *S. phaseoloides* form a clade with 85% support and copies spanning only intron sequence from *J. grandiflora* (copies 2 and 3) form a clade with 75% support (Figure 4). Copy 2 and 3 from *J. grandiflora* are more similar to copies from *S. phaseoloides* than they are to copy 1 from *J. grandiflora* (Figure 4).

A MITE search identified many elements in intron 3 overlapping with elements identified as insertion sequences, inverted repeats, or tandem repeats (Table 5). Transposon-like elements from *Jaltomata darcyana* and *Solanum abutiloides* (Table 2) overlap with regions identified as putative MITEs (Table 5). The mitochondrial gene copy insertion from *J. procumbens* at position

1985-2230 (Table 2) is flanked by two putative MITE elements at positions 1795-1961 and 2304-2412 (Table 5). Putative MITEs from *Solanum muricatum* and *S. tuberosum* (Table 5) include regions identified as inverted repeats (Table 3). Regions implicated as putative MITES (Table 5) were also identified as tandem repeats: one from *Jaltomata darcyana*, both repeat cluster types from *J. grandiflora*, one from *J. paneroi*, and one from *Solanum phaseoloides* (Table 4).

Discussion

Phylogenetic relationships among 3-intron rbcS copies

A 3-intron *rbcS* copy was identified in *Cestrum nocturnum*. Along with the copy already known from *Petunia*, these were the only 3-intron copies found in any of the basal Solanaceae lineages examined. The 3-intron copy has only been found in species within the unresolved clade containing "x=12," Petunieae, Schwenckieae, Benthamielleae, and Cestroideae, but three species within this clade appear to lack the 3-intron copy: *Capsicum annuum*, *Schwenckia glabrata*, and *Browallia speciosa*.

Lack of evidence for a 3-intron copy from *Schwenckia* and *Browallia* does not rule out the possibility one exists in these lineages and may be identified by future studies. However,

numerous *rbcS* copies were identified for each lineage and every copy was identified repeatedly using different PCR primers suggesting all copies were sampled (Chapter 3) unless 3-intron copies in these lineages have diverged substantially from the two-intron copies (Appendix 1). Published sequences from *Capsicum annuum* also indicate the absence of a 3-intron copy, which would suggest its loss in *Capsicum* and perhaps elsewhere in Solanaceae.

Despite possible independent loss of the 3-intron copy, the presence/absence of an *rbcS* three-intron copy may prove a useful marker to support relationships at the base of the Solanaceae tree. The 3-intron *rbcS* tree (Figure 2) lacks satisfactory support, but topology is consistent with previous studies indicating a close relationship between the X=12 clade and *Browallia*, *Cestrum*, *Petunia*, and *Schwenckia* to the exclusion of *Schizanthus*, *Goetzea*, and *Duckeodendron* (Olmstead et al. 2008; Särkinen et al. 2013).

The presence of so many identifiable sequence elements is somewhat surprising for an intron. Intron 3 lengths range from 73 bases in one *Cestrum nocturnum* copy (3i.a) and one *Nicotiana tabacum* (C) copy to 867 bases in *Jaltomata procumbens*. Long introns and sequences with many mutational triggers, such as repeats and inversions, are often targets for various mutational events that may lead to length and sequence changes (Kelchner 2000). Furthermore, insertions and excisions by transposable elements increase recombination through repair of double stranded breaks that can contribute to further length changes through repair by non-homologous end joining (Buchmann et al. 2012). Evidence for these processes in intron 3 potentially explain the high sequence divergence between closely related species (Table 2-5).

rbcS copies lacking introns

Although mechanisms for intron gain remain elusive, evidence for intron loss is dominated by evidence for a recombination mechanism involving cDNA reverse transcribed from the mRNA of the gene containing the intron (Roy and Gilbert 2005). The one-intron *rbcS* copies within *Browallia* (lacking intron 1), the *Schwenckia* sequence lacking all introns, and the 2-intron *Nicotiana tabacum* sequence nearly identical to 3-intron copies could all be evidence of intron loss at existing loci through recombination with a cDNA copy or, alternatively, may be evidence of new duplicated copies created by insertion after removal of one intron instead of recombination. The recombination mechanism is thought to more frequently remove 3' introns due to recombination with partially reverse transcribed mRNA molecules such that the cDNA would trend towards partial copies lacking 5' ends.

rbcS intron 3 sequence elements

Numerous types of transposable elements (TEs) were identified in *rbcS* intron 3 (Table 2-5) that may act to affect intron length and nucleotide sequence through transposition of gene sequences (Lisch 2009) or through the double strand breaks (DSBs) produced by their arrival or subsequent activity (Buchmann et al. 2012). Deletions within autonomous TEs produce truncated versions identified as miniature inverted-repeat transposable elements (MITEs) (Jiang et al. 2004). MITEs

were identified in most *rbcS* intron 3 sequences (Table 5). The frequency of these putative elements is surprising, but MITEs have been found to insert often and amplify in copy number through cross mobilization (Jiang et al. 2004; Zhang and Wessler 2004)(Zhang and Wessler 2004)

Inverted repeats can also initiate DSBs that initiate repair mechanism (Bi and Liu 1996; Stankiewicz and Lupski 2002). The inverted repeats in *rbcS* intron 3 may stimulate DSBs (Table 3). Repeat units as short as 13 bases with 8bp linkers can be targeted by recombinase which will result in either excisions or inversions based on inverted repeat sequence orientation (Turan and Bode 2011).

DSBs caused by repetitive DNA and/or TEs initiates recombination-based repair (Puchta 2005). Repair by a single-strand annealing mechanism will create deletions of a few bases in DNA and repair by a synthesis dependent strand annealing mechanism can insert filler DNA and produce duplications of a few – hundreds of bases (Puchta 2005; Buchmann et al. 2012). The filler DNA may be inserted from any genomic source by non-homologous end joining (Puchta 2005).

Origin of the novel intron

Among the mechanisms of proposed intron gain, the *rbcS* intron 3 exhibits a potential source for a protosplice site (AGGT) immediately 5' of the intron at the end of exon 3a, the same region tandemly repeated within *Solanum phaseoloides* and *Jaltomata grandiflora*, which may support the tandem duplication hypothesis (Yenerall and Zhou 2012).

Evidence for acquisition of a preexisting intron was not found through sequence similarity between *rbcS* intron 3 and other *rbcS* introns (Hankeln et al. 1997; Tarrío et al. 1998). The lack of evidence may suggest little, since there is also little similarity between intron 3 sequences among genera.

Evidence for TEs within *rbcS* intron 3 in many lineages supports the possibility of intron gain by co-option of a transposon (Giroux et al. 1994). Additionally, the presence of so many elements associated with the creation of DSBs and the presence of a full length mitochondrial gene insert from *Jaltomata procumbens* may support insertion of the original intron by non-homologous end joining (Li et al. 2009).

Evidence for independent and recent mechanisms acting within the intron is not evidence for the origin of the intron. Ongoing TE activity and recombination have most likely obliterated any evidence for a mechanism triggering the origination of the novel intron, however, it is possible the ongoing mechanisms producing the identifiable sequence elements have persisted since the intron originated.

The origin of the 3-intron locus may be independent from the origin of the extra intron; either a previously existing 2-intron locus gained an intron or the extra intron was inserted during a duplication event that simultaneously created a new locus with 3-introns. Phylogenetic relationships between copies of the transit coding sequence indicate 3-intron copies to share a more recent common ancestor with the 2-intron copies from locus 3 than from locus 1 (Figure 3).

Conclusion

Among land plants, *rbcS* copies contain two well-conserved intron positions (Table 1). Unique 3-intron *rbcS* copies are found in lineages at the base of the Solanaceae. The presence/absence of this novel intron may provide support for resolving relationships among lineages at the base of the Solanaceae. A number of independent losses and duplications of the 3-intron copies, which are indicated by phylogenetic analyses, may limit the utility of using this intron to support deep-branching relationships at the base of the Solanaceae, however, the rapid divergence in intron sequence between copies may prove useful for examining relationships within a genus.

Identification of possible TEs and a mitochondrial gene insertion within the novel intron suggest ongoing mechanisms of TE insertion and/or recombination repair leading to the insertion of filler DNA in many lineages. These processes may be linked such that TE insertion/excision increases the frequency of recombination repair, which subsequently provides opportunities for further DNA insertions and drives a fast rate of intron sequence divergence between genera.

Literature Cited

Akazawa, T., T. Takabe, and H. Kobayashi. 1984. Molecular evolution of ribulose-1,5-biphosphate carboxylase/oxygenase (RuBisCO). Trends Biochem. Sci. 9:380–383.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Banks, J. A., T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. dePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pils, M. Prigge, S. A. Rensing, D. M. Riano-Pachon, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. S?rensen, R. Sotooka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J.-K. Weng, W. W. G. T. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loque, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, and I. V. Grigoriev. 2011. The compact Selaginella genome identifies changes in gene content associated with the evolution of vascular plants. Science 332:960–963.

Bao, W., and J. Jurka. 2008. EnSpm-type DNA transposon from maize. Repbase Rep. 8:689–691.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. GenBank. Nucleic Acids Res. gks1195.

Bi, X., and L. F. Liu. 1996. DNA rearrangement mediated by inverted repeats. Proc. Natl. Acad. Sci. 93:819–823.

Bombarely, A., N. Menda, I. Y. Tecle, R. M. Buels, S. Strickler, T. Fischer-York, A. Pujar, J. Leto, J. Gosselin, and L. A. Mueller. 2011. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res. 39:D1149–D1155.

Buchmann, J. P., T. Matsumoto, N. Stein, B. Keller, and T. Wicker. 2012. Inter-species sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. Cell Mol. Biol. 71:550–563.

Chen, J., Q. Hu, Y. Zhang, C. Lu, and H. Kuang. 2014. P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. 42:D1176–D1181.

Chen, Y., F. Zhou, G. Li, and Y. Xu. 2009. MUST: A system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. Gene 436:1–7.

Chua, N.-H., and G. W. Schmidt. 1978. Post-translational transport into intact chloroplasts of a precursor to the small subunit of ribulose-1,5-bisphosphate carboxylase. Proc. Natl. Acad. Sci. 75:6110–6114.

Coruzzi, G., R. Broglie, C. Edwards, and N. H. Chua. 1984. Tissue-specific and light-regulated expression of a pea nuclear gene encoding the small subunit of ribulose-1,5-bisphosphate carboxylase. EMBO J. 3:1671–1679.

Daraselia, N. D., S. Tarchevskaya, and J. O. Narita. 1996. The Promoter for Tomato 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase Gene 2 Has Unusual Regulatory Elements That Direct High-Level Expression. Plant Physiol. 112:727–733.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772–772.

Dean, C., P. van den Elzen, S. Tamaki, M. Black, P. Dunsmuir, and J. Bedbrook. 1987. Molecular characterization of the rbcS multi-gene family of Petunia (Mitchell). Mol. Gen. Genet. MGG 206:465–474.

Dean, C., E. Pichersky, and P. Dunsmuir. 1989. Structure, Evolution, and Regulation of RbcS Genes in Higher Plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 40:415–439.

Doyle, J., and J. Doyle. 1987. Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem Bull 19:11–15.

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783.

Feschotte, C., and E. J. Pritham. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annu. Rev. Genet. 41:331–368.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

Gelfand, Y., A. Rodriguez, and G. Benson. 2007. TRDB—The Tandem Repeats Database. Nucleic Acids Res. 35:D80–D87.

Giroux, M. J., M. Clancy, J. Baier, L. Ingham, D. McCarty, and L. C. Hannah. 1994. De novo synthesis of an intron by the maize transposable element Dissociation. Proc. Natl. Acad. Sci. U. S. A. 91:12150–12154.

Gouy, M., S. Guindon, and O. Gascuel. 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol. 27:221–224.

Hanania, U., and A. Zilberstein. 1994. Ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit gene from the fern Pteris vittata. Plant Physiol. 106:1685–1686.

Hankeln, T., H. Friedl, I. Ebersberger, J. Martin, and E. R. Schmidt. 1997. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene 205:151–160.

Hellsten, U., J. L. Aspden, D. C. Rio, and D. S. Rokhsar. 2011. A segmental genomic duplication generates a functional intron. Nat. Commun. 2:454.

Hutchison, K. W., P. D. Harvie, P. B. Singer, A. F. Brunner, and M. S. Greenwood. 1990. Nucleotide sequence of the small subunit of ribulose-1,5-bisphosphate carboxylase from the conifer Larix laricina. Plant Mol. Biol. 14:281–284.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, P. Wincker, and French-Italian Public Consortium for Grapevine Genome Characterization. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.

Jiang, N., C. Feschotte, X. Zhang, and S. R. Wessler. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr. Opin. Plant Biol. 7:115–119.

Jo, Y. D., J. Park, J. Kim, W. Song, C.-G. Hur, Y.-H. Lee, and B.-C. Kang. 2011. Complete sequencing and comparative analyses of the pepper (Capsicum annuum L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. Plant Cell Rep. 30:217–229.

Jurka, J. 2010. LTR retrotransposons from the apple genome. Repbase Rep. 10:1668–1668.

Jurka, J. 2012. LTR retrotransposons from the date palm genome. Repbase Rep. 12:65.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Katoh, K., and M. C. Frith. 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 28:3144–3146.

Kelchner, S. A. 2000. The Evolution of Non-Coding Chloroplast DNA and Its Application in Plant Systematics. Ann. Mo. Bot. Gard. 87:482–498.

Kojima, K. K., and J. Jurka. 2012. DNA transposons from the Eutrema parvulum genome. Repbase Rep. 12:1455.

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. 2011. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. gkr1090.

Leliaert, F., D. R. Smith, H. Moreau, M. D. Herron, H. Verbruggen, C. F. Delwiche, and O. De Clerck. 2012. Phylogeny and Molecular Evolution of the Green Algae. Crit. Rev. Plant Sci. 31:1–46.

Li, W., A. E. Tucker, W. Sung, W. K. Thomas, and M. Lynch. 2009. Extensive, recent intron gains in Daphnia populations. Science 326:1260–1262.

Lisch, D. 2009. Epigenetic regulation of transposable elements in plants. Annu. Rev. Plant Biol. 60:43–66.

Meagher, R. B., S. Berry-Lowe, and K. Rice. 1989. Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. Genetics 123:845–863.

Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Pp. 1–8 *in* Gateway Computing Environments Workshop (GCE), 2010. IEEE.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst. Biol. 52:674–683.

Mishkind, M. L., S. R. Wessler, and G. W. Schmidt. 1985. Functional determinants in transit sequences: import and partial maturation by vascular plant chloroplasts of the ribulose-1,5-bisphosphate carboxylase small subunit of Chlamydomonas. J. Cell Biol. 100:226–234.

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

Olmstead, R. G., K.-J. Kim, R. K. Jansen, and S. J. Wagstaff. 2000. The Phylogeny of the Asteridae sensu lato Based on Chloroplast ndhF Gene Sequences. Mol. Phylogenet. Evol. 16:96–112.

Palmer, J. D. 1985. Comparative Organization of Chloroplast Genomes. Annu. Rev. Genet. 19:325–354.

Paritosh, K., D. Pental, and P. K. Burma. 2013. Structural and Transcriptional Characterization of rbcS Genes of Cotton (Gossypium hirsutum). Plant Mol. Biol. Report. 31:1176–1183.

Pichersky, E., R. Bernatzky, S. D. Tanksley, and A. R. Cashmore. 1986. Evidence for selection as a mechanism in the concerted evolution of Lycopersicon esculentum (tomato) genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. Proc. Natl. Acad. Sci. U. S. A. 83:3880–3884.

Puchta, H. 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J. Exp. Bot. 56:1–14.

Rambaut, A. 2009. FigTree v1. 3.1: Tree figure drawing tool. FigTree Website.

Roy, S. W., and W. Gilbert. 2005. The pattern of intron loss. Proc. Natl. Acad. Sci. U. S. A. 102:713–718.

Rubin, E., G. Lithwick, and A. A. Levy. 2001. Structure and evolution of the hAT transposon superfamily. Genetics 158:949–957.

Sakai, H., S. S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, H. Wakimoto, C. Yang, M. Iwamoto, T. Abe, Y. Yamada, A. Muto, H. Inokuchi, T. Ikemura, T. Matsumoto, T. Sasaki, and T. Itoh. 2013. Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. Plant Cell Physiol. 54:e6–e6.

Särkinen, T., L. Bohs, R. G. Olmstead, and S. Knapp. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol. Biol. 13:214.

Sasanuma, T., and N. T. Miyashita. 1998. Subfamily divergence in the multigene family of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcS*) in Triticeae and its relatives. Genes Genet. Syst. 73:297–309.

Sierro, N., J. N. Battey, S. Ouadi, L. Bovet, S. Goepfert, N. Bakaher, M. C. Peitsch, and N. V. Ivanov. 2013. Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. Genome Biol. 14:R60.

Silverthorne, J., C. F. Wimpee, T. Yamada, S. A. Rolfe, and E. M. Tobin. 1990. Differential expression of individual genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase in Lemna gibba. Plant Mol. Biol. 15:49–58.

Stankiewicz, P., and J. R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet. 18:74–82. You've got to bite it.

Sugita, M., T. Manzara, E. Pichersky, A. Cashmore, and W. Gruissem. 1987. Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. Mol. Gen. Genet. MGG 209:247–256.

Sukumaran, J., and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala. 2003. A new Drosophila spliceosomal intron position is common in plants. Proc. Natl. Acad. Sci. 100:6580–6583.

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala. 1998. New Drosophila introns originate by duplication. Proc. Natl. Acad. Sci. 95:1658–1662.

The Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. 161:105–121.

Thomas-Hall, S., P. R. Campbell, K. Carlens, E. Kawanishi, R. Swennen, L. Sági, and P. M. Schenk. 2007. Phylogenetic and molecular analysis of the ribulose-1,5-bisphosphate carboxylase small subunit gene family in banana. J. Exp. Bot. 58:2685–2697.

Turan, S., and J. Bode. 2011. Site-specific recombinases: from tag-and-target- to tag-and-exchange-based genomic modifications. FASEB J. 25:4088–4107.

Turner, N. E., W. G. Clark, G. J. Tabor, C. M. Hironaka, R. T. Fraley, and D. M. Shah. 1986. The genes encoding the small subunit of ribulose-l,5-bisphosphate carboxylase are expressed differentially in petunia leaves. Nucleic Acids Res. 14:3325–3342.

Wicker, T., R. Guyot, N. Yahiaoui, and B. Keller. 2003. CACTA Transposons in Triticeae. A Diverse Family of High-Copy Repetitive Elements. Plant Physiol. 132:52–63.

Wilhelm, M., and F.-X. Wilhelm. 2001. Reverse transcription of retroviruses and LTR retrotransposons. Cell. Mol. Life Sci. CMLS 58:1246–1262.

Wolter, F. P., C. C. Fritz, L. Willmitzer, J. Schell, and P. H. Schreier. 1988. rbcS genes in Solanum tuberosum: conservation of transit peptide and exon shuffling during evolution. Proc. Natl. Acad. Sci. 85:846–850.

Yamazaki, T., M. Yamamoto, W. Sakamoto, and S. Kawano. 2005. Isolation and molecular characterization of rbcS in the unicellular green alga Nannochloris bacillaris (Chlorophyta, Trebouxiophyceae). Phycol. Res. 53:67–76.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

Yenerall, P., and L. Zhou. 2012. Identifying the mechanisms of intron gain: progress and trends. Biol. Direct 7:29.

Zhang, X., and S. R. Wessler. 2004. Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. Proc. Natl. Acad. Sci. U. S. A. 101:5589–5594.

Zwickl, D. 2006. GARLI—genetic algorithm for rapid likelihood inference. See Httpwww Bio Utexas Edufacultyantisensegarligarli Html.

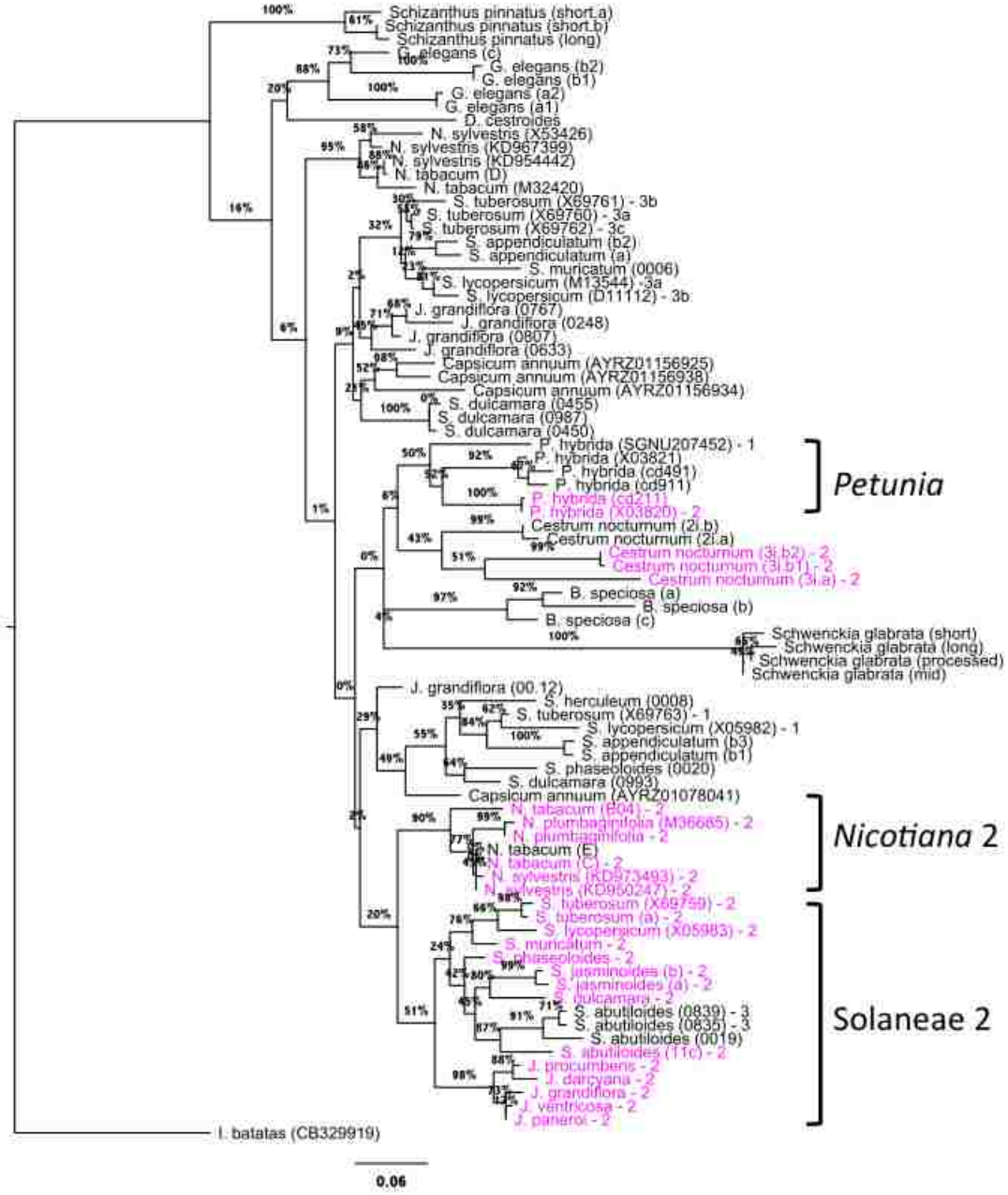


Figure 1. ML tree of Solanaceae *rbcS* sequences. Copies have two introns (black) and three introns (pink). Bootstrap support values are shown as percentages above branches.

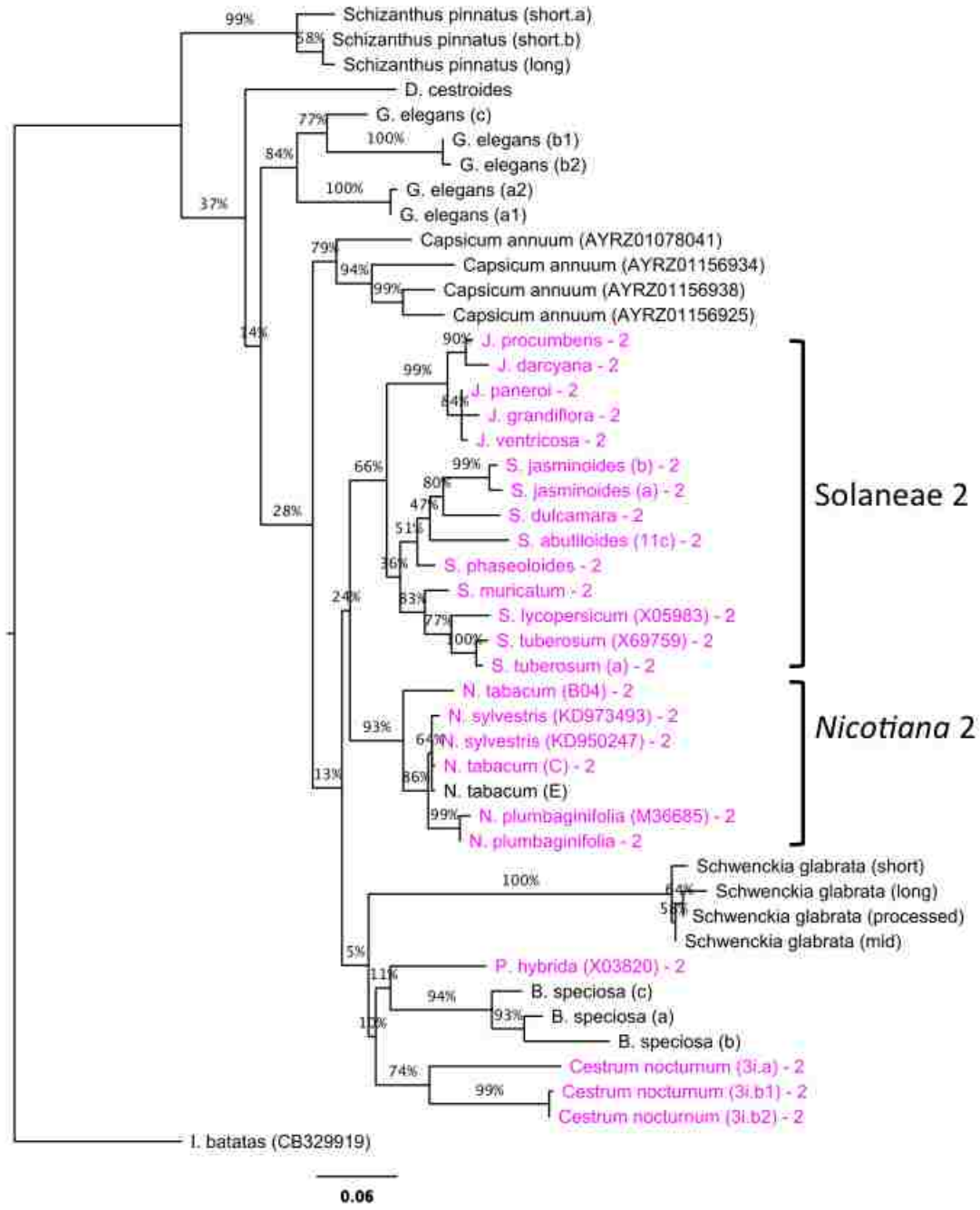


Figure 2. ML tree of Solanaceae 3-intron *rbcS* sequences (*rbcS* locus 2). Species lacking 3-intron copies are represented by all 2-intron copies available. Copies have two introns (black) and three introns (pink). Bootstrap support values are shown as percentages above branches.

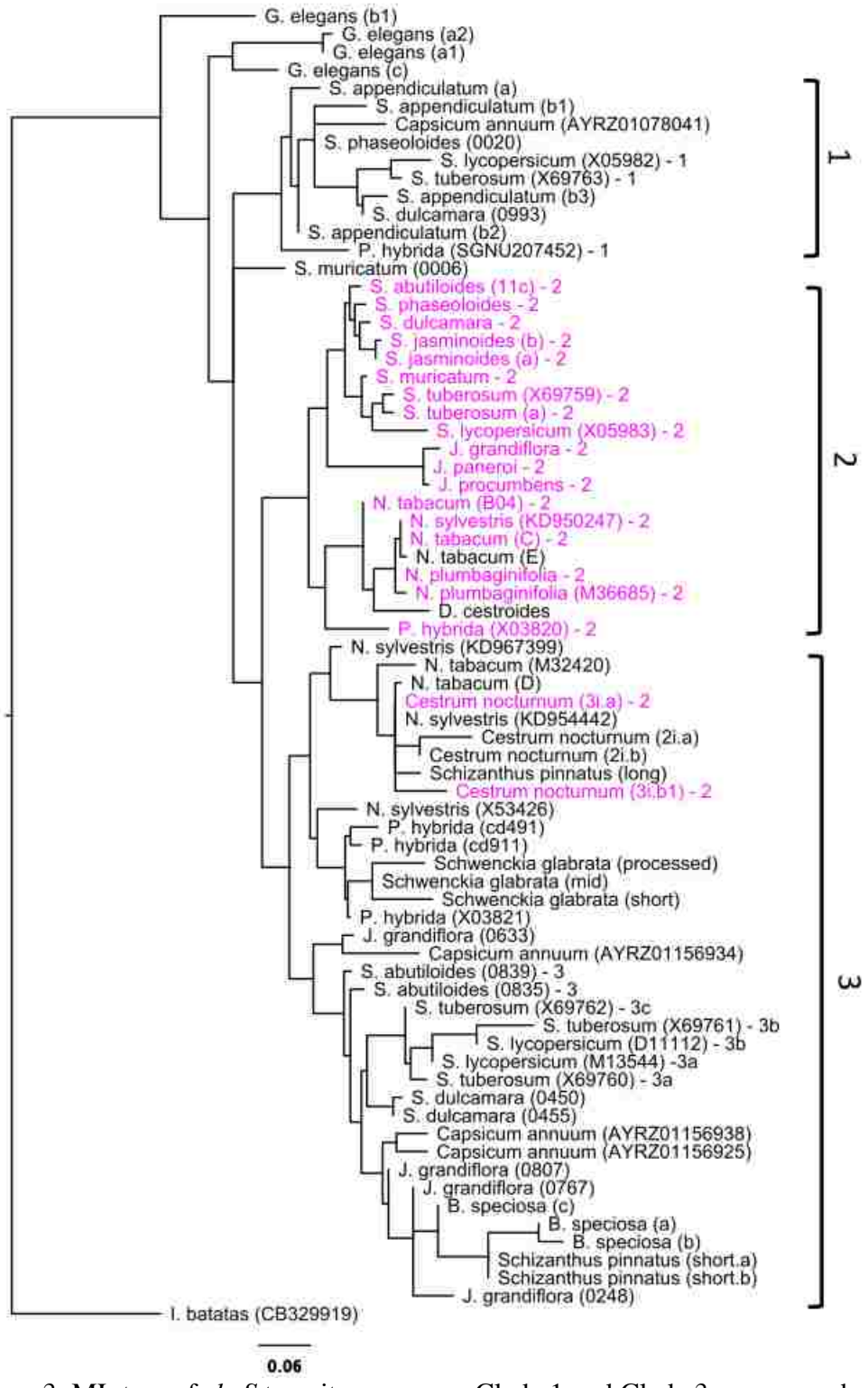


Figure 3. ML tree of *rbcS* transit sequences. Clade 1 and Clade 3 sequences have 2 introns (black); Clade 2 sequences have three introns (pink).

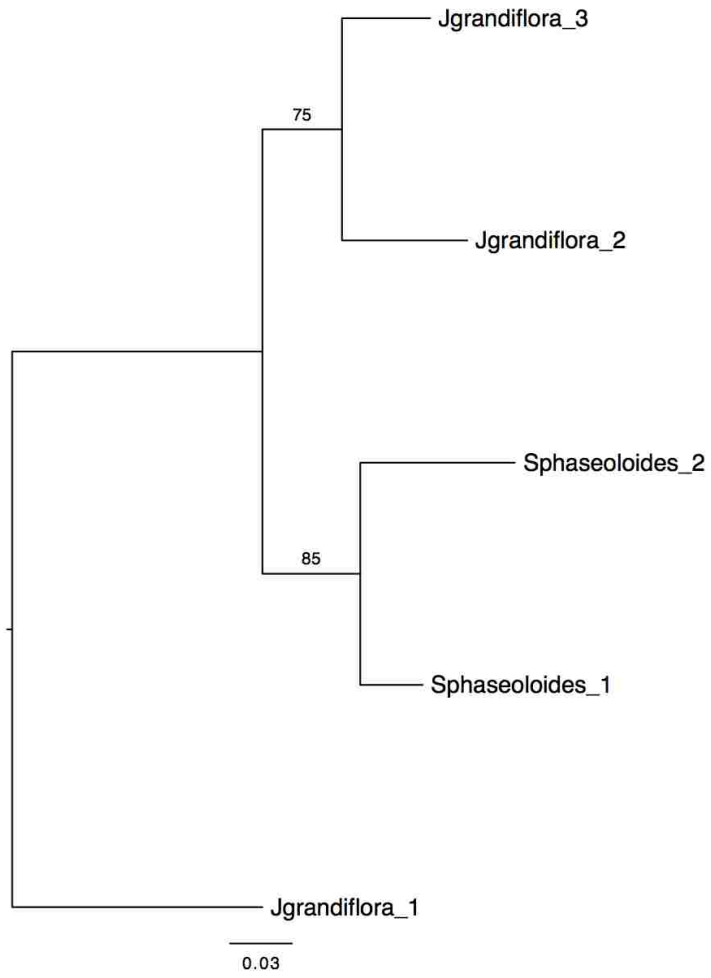


Figure 4. NJ tree of tandem repeats identified from *Jaltomata grandiflora* and *Solanum phaseoloides rbcS* intron 3. The tandem repeat unit includes 23 bases from the end of exon 3a and over 150 intron bases. Tandem repeats are numbered sequentially with 1: exon 3a and bases from flanking intron, 2: repeat immediately 3' of copy 1, 3: repeat immediately 3' of copy 2. Bootstrap support values are shown as percentages above branches.

Table 1. Intron number and positions for *rbcS* among land plants and green algae. Among land plants, introns occur in homologous locations: intron 1 occurs between the 2nd and 3rd amino acids of the mature peptide in phase 0, intron 2 occurs between the 47th and 48th amino acids of the mature peptide in phase 0, and intron 3 separates the nucleotides of the codon for the 65th amino acid of the mature peptide in phase 2.

			------introns-----				
Lineage	**Species**	**copies**	**no.**	**1**	**2**	**3**	**GenBank accession:**
moss	*Physcomitrella patens*	-	2	+	+		ABEU01009194, ABEU01009192, ABEU01010557
lycophyte	*Selaginella moellendorfii*	-	2	+	+		ADFJ01003544, ADFJ01001130
fern	*Pteris vittata*	4+	2	+	+		X77813, X98414
conifer	*Larix laricina*	-	2	+	+		X16039
monocots							
Alismatales	*Lemna gibba*	5	1		+		X17230, X17231, X17232, X17233, X17234
commelinids							
Poales	*Triticum aestivum*	12+	1	+			AB020941, AB020957, AB020958, AB042064-8
	Oryza sativa	5	1	+			Os04g0658300, Os11g0707000, Os12g0274700, Os12g0291100, Os12g0292400,
	Zea mays	10+	1	+			Y00322
Zingiberales	*Musa acuminata*	6	2	+	+		DQ088090-9
eudicots							
rosids							
Brassicales	*Arabidopsis thaliana*	4	2	+	+		NM105379, NM123204, NM123203, NM123202
Malvales	*Gossypium hirsutum*	5	2	+	+		JN608783, JN608788, JN608790, JN608791, JN608792
Fabales	*Pisum sativum*	4	2	+	+		X00806, X04333, X04334
asterids							
Asterales	*Helianthus annuus*	-	2	+	+		Y00431
Lamiales	*Erythranthe guttata*	-	2	+	+		APLE01009767, APLE01007204
Gentianales	*Coffea arabica*	-	2	+	+		AJ419827
Solanales	*Solanum lycopersicum*	4	2	+	+		X05982, X05986, D11112, M13544
	Solanum lycopersicum	1	3	+	+	+	X05983
	Nicotiana sylvestris	4	2	+	+		X53426, KD954442, KD959040, KD967399, KD975324
	Nicotiana sylvestris	2	3	+	+	+	KD950247, KD973493
	Petunia hybrida	5	2	+	+		X03821, X12987
	Petunia hybrida	1	3	+	+	+	X03820
green algae (Chlorophyta)							
Chlorellales	*Nannochloris*	2	2	+			AB125312-3

	bacillaris			
	Nannochloris bacillaris	1	3	AB125314
Chlamydomonadales	*Dunaliella teriolecta*	2	3	AY530155-6
	Chlamydomonas reinhardtii	2	3	X04471-2

Table 2. Sequences elements from *rbcS* intron 3 in *Jaltomata* and *Solanum* identified by similarity. Nucleotide position of element starting position within full-length *rbcS* sequence (From) and ending position (To) indicated with length of sequence element (Size). Sequence element identified from Repbase Update library (Sequence) with starting position (From) and ending position (To) identified. Class of element listed according to Repbase Update (Class), Class I transposons use an RNA intermediate and Class II transposons are cut and paste DNA elements. Similarity calculated between species sequence and sequence element (Sim), a similarity score over the ratio of mismatches to transitions (S/Mm:Ts) and BLAST score (Score) are listed for each species. The sequence from *J. procumbens* was identified by BLAST search and the sequence from *S. lycoperiscum* identified by BLAST search of the P-MITE database.

Species	From	To	size	Sequence	From	To	Class	Sim	S/Mm:Ts	Score	
J. darcyana	1069	1160	91	ENSPM-6_ZM	11350	11437	II: CACTA	0.78	2.6	272	(Bao and Jurka 2008)
J. darcyana	1405	1509	104	Gypsy-23_VV-I	1581	1689	I: LTR/Gypsy	0.72	1.2	269	(Jaillon et al. 2007)
J. grandiflora	1953	1998	45	Copia-65_Mad-I	49	100	I: LTR/Copia	0.81	1.2	232	(Jurka 2010)Jurka 2010
J. procumbens	1767	1790	23	EnSpm1_HV	9339	9363	II: CACTA	1.00	99.0	234	(Bao and Jurka 2008)
J. procumbens	1985	2230	245	mt_tRNA-Ile	120950	121195	*Nicotiana tabacum* complete mitochondrial genome (BA000042)	0.99		438	
S. abutiloides	1856	1939	83	TORTL1	9538	9620	I: LTR/Copia	0.80	2.5	358	(Daraselia et al. 1996)
S. abutiloides	1986	2020	34	Copia-8_PD-I	1691	1729	I: LTR/Copia	0.89	1.0	212	(Jurka 2012)Jurka 2012
S. dulcamara	2064	2109	45	hAT-4_EPa	905	953	II: hAT	0.81	2.0	208	(Kojima and Jurka 2012)
S. lycoperisum	1913	2047	279	SQ372025426	1	279	II: MITE/Mutator-like	1.00		456	(Chen et al. 2014)

Table 3. Inverted repeats identified from intron 3 in *Solanum*. Sequence position listed for first (Left) and second (Right) repeat units.

Sequence	Left		Length	Right		Length	Loop size	%Match	%Indel	Score
S. muricatum	1585	1608	24	2197	2220	24	588	87.5	0.0	33
S. tuberosum	1735	1752	18	1958	1975	18	205	88.9	0.0	26
S. tuberosum	1845	1862	18	1875	1892	18	12	100.0	0.0	36
S. lycopersicum	1624	1742	119	1755	1870	116	12	78.2	2.5	102

Table 4. Tandem repeats identified from intron 3. Each repeat is summarized by size in nucleotides (Length), number of repeats (Copies), percent match (%Match), percent indels (%Indel), BLAST Score, and the position (Start, End) and length (Array) of all identified copies. Among the elements with score >100, two clusters are found: a *Jaltomata* specific cluster of repeats of ~95 bases and a longer repeat over 180 bases long within *Solanum phaseoloides* and *Jaltomata grandiflora*.

Sequence	Length	Copies	%Match	%Indel	Score	Start	End	Array
S. phaseoloides	182	2.4	93	2	601	2291	2713	423
J. grandiflora	184	2.4	90	4	509	1805	2237	433
J. darcyana	94	3.0	92	2	388	1241	1524	284
J. ventricosa	97	3.0	92	4	394	1306	1595	290
J. paneroi	95	2.9	92	3	358	1766	2041	276
J. grandiflora	94	2.8	93	2	382	2244	2505	262
S. muricatum	17	2.1	91	2	45	2180	2214	35

Table 5. MITEs identified in *rbcS* intron 3 from *Solanum*, *Jaltomata*, *Nicotiana*, *Petunia*, and *Cestrum*. Elements from each species are summarized by position in sequence (Start, End) and length (Length), with direct repeat size (DR), size (TIR) and percent match (%Match) of terminal inverted repeat listed.

Sequence	Start	End	Length	DR	TIR	%Match
S. tuberosum b - 2	1737	1973	237	2	13	92.31
S. tuberosum a - 2	1737	1973	237	2	13	92.31
S. lycopersicum X05983 - 2	1580	1858	279	5	10	90.00
S. jasminoides b - 2	1475	1587	113	2	9	88.89
S. phaseoloides - 2	2436	2584	149	30	9	88.89
S. phaseoloides - 2	2624	2744	121	6	8	87.50
S. muricatum - 2	1600	1994	395	5	9	88.89
S. dulcamara - 2	1690	1886	197	2	11	90.91
S. dulcamara - 2	1913	2047	135	19	9	88.89
S. abutiloides - 2	1853	2033	181	2	10	90.00
J. grandiflora - 2	1758	1952	195	2	10	90.00
J. grandiflora - 2	2285	2393	109	2	12	91.67
J. paneroi - 2	1293	1456	164	2	11	90.91
J. paneroi - 2	1833	1997	165	30	10	90.00
J. ventricosa - 2	849	1012	164	2	11	90.91
J. ventricosa - 2	1039	1180	142	16	9	88.89
J. ventricosa - 2	1209	1389	181	5	10	90.00
J. ventricosa - 2	1472	1580	109	5	11	90.91
J. darcyana - 2	919	1114	196	2	10	90.00
J. darcyana - 2	1303	1466	164	29	9	88.89
J. procumbens - 2	1795	1961	167	5	10	90.00
J. procumbens - 2	2304	2412	109	6	11	90.91
N. tabacum X02353 - 2	1662	1802	141	4	10	90.00
N. sylvestris KD950247 - 2	727	867	141	4	9	88.89
N. sylvestris KD973493 - 2	843	982	140	4	9	88.89

N. tabacum C - 2	1557	1697	141	4	10	90.00
N. plumbaginifolia M36685 - 2	1673	1800	128	2	8	87.50
N. plumbaginifolia - 2	1532	1660	129	2	8	87.50
P. hybrida X03820 - 2	1590	1698	109	3	8	87.50
C. nocturnum 3a - 2	460	566	107	2	11	63.64
C. nocturnum 3b1 - 2	913	1019	107	11	10	60.00
C. nocturnum 3b2 - 2	902	1015	114	11	10	60.00

Supplementary Table 1. Primers used to amplify *rbcS* copies within Solanaceae.

Primer	Location	Locus	primer sequence, 5'-3'
1	exon 1	all	CAATGGCTTCCTCwrTnnTnTCCTC
2	exon 3b	all	GGCTTGTArGCrATGAAACTGATrC
3	exon 1	3	CCCTGTTTCAAGGAAGCAAAACC
4	exon 1	3	GGACTTrAGkCCAGTGAAGGG
5	exon 1	all	AACCTTGACATTACyTCCmTTGC
6	exon 3b	all	ATGAAACTGATrCACTGCACTTGACG
7	5' UTR	2	GATTAmyGAGGTGCTTACACG
8	exon 3b	2	CCCTTCTGGCTTGTAGGC
9	5' UTR	2	AATTGTATAATGTTATCAAGAACCAC
10	exon 3b/3' UTR	2	TCCTAATATGAAACTTAGTAkCCTTC

Supplementary Table 2. Phylogenetic model compared by AIC score. GTR+I+G model, partitioned with subset specific rates was indicated as the best-fit and used in subsequent analyses. Column labels are abbreviated as partition of data as 0 (none) and separate transit and mature coding regions (1): p, link models parameter in Garli: l; subset specific rates are estimated (=1) or not (=0): s; substitution model (2nd model below 1st): Model; number of parameters: #P; subset rate multiplier for each partition: S1, S2; substitution rates: AC, AG, AT, CG, CT, (GT = 1); nucleotide frequencies: A, C, G, T; alpha shape parameter for discrete gamma rate heterogeneity distribution: a; proportion of invariable sites: I.

p	l	s	Model	#P	lnL	AIC	S1	S2	AC	AG	AT	CG	CT	A	C	G	T	a	I
1	1	1	SYM+I+G	8	-8232.466367	16480.9	1.21	0.90	2.38	4.64	2.46	1.45	5.93	0.25	0.25	0.25	0.25	0.84	0.23
0	-	-	SYM+I+G	7	-8232.000311	16478.0			2.39	4.44	2.43	1.42	5.95	0.25	0.25	0.25	0.25	0.85	0.24
1	1	1	GTR+I+G	11	-8231.663331	16485.3	1.12	0.94	2.11	3.91	2.11	1.30	5.61	0.27	0.25	0.24	0.24	0.85	0.24
0	-	-	GTR+I+G	10	-8230.919246	16481.8			2.21	4.05	2.24	1.43	6.01	0.27	0.24	0.25	0.24	0.85	0.24
1	0	0	HKY+G	11	-8227.189436	16476.4	-	-	1.00	3.19	1.00	1.00	3.19	0.27	0.30	0.15	0.28	0.45	-
			TVMef+I+G				-	-	2.45	4.79	2.21	1.43	4.79	0.25	0.25	0.25	0.25	0.99	0.27
1	0	1	HKY+G	12	-8221.165544	16466.3	1.29	0.86	1.00	3.27	1.00	1.00	3.27	0.28	0.30	0.15	0.28	0.48	-
			TVMef+I+G						2.41	4.80	2.22	1.52	4.80	0.25	0.25	0.25	0.25	0.99	0.30
1	0	0	GTR+I+G	20	-8219.08944	16478.2	-	-	1.66	4.04	2.22	1.12	5.68	0.26	0.29	0.20	0.26	0.55	0.11
			GTR+I+G						2.49	4.02	2.17	1.63	6.00	0.28	0.22	0.27	0.24	1.03	0.28
1	0	1	GTR+G	20	-8213.173434	16466.3	1.30	0.86	1.55	4.43	2.19	1.10	5.21	0.25	0.31	0.18	0.27	0.46	
			GTR+I+G						2.36	3.88	1.99	1.55	5.64	0.28	0.22	0.27	0.23	1.00	0.30
1	0	1	GTR+I+G	21	-8211.601096	16465.2	1.29	0.86	1.49	4.20	2.15	1.08	5.10	0.25	0.31	0.18	0.27	0.47	0.00
			GTR+I+G						2.40	3.82	2.04	1.52	5.69	0.28	0.22	0.27	0.23	1.04	0.30

Appendix

Taxa and vouchers for species sampled. Species, geographic origin (specific to collection if known or general for species), collector and collection number (herbarium), DNA sequence identifier (no. clones identified for each unique copy).

***Browallia speciosa* (Hook.),** South America, *Olmstead* S.0416 (BIRM), a (8), b (6), c (11); ***Cestrum nocturnum* (Hook.),** South America, *Olmstead* S.0416 (WTU), 2i.a (12), 3i.a (9), 2i.b (4), 3i.b1 (7), 3i.b2 (16); ***Duckeodendron cestroides* (Kuhlm.),** Brazil, *Ribeiro* 1189 (K), (23); ***Goetzea elegans* (Wydl.),** Puerto Rico, *Olmstead* (WTU), a1 (4), a2 (5), b1 (3), b2 (6), c (3); ***Jaltomata darcyana* (Mione),** Costa Rica, *T. Mione & L. Yacher* 694 (NY, CR, MO), ; ***Jaltomata paneroi* (Mione & S. Leiva),** Peru, Cajamarca, *Mione et al*. 705 (CCSU); ***Jaltomata procumbens* (Cav.) J.L. Gentry ,** Costa Rica, *T. Mione & L. Yacher* 692 (CCSU); ***Jaltomata ventricosa* (Baker) Mione,** Peru, La Libertad, *T. Mione, S. Leiva G. & L. Yacher* 712 (DNA only); ***Jaltomata grandiflora* (B.L. Rob. & Greenm.) D'Arcy, Mione, & T. Davis,** Mexico, grown from Davis 1114 (MO), *Mione* 454 (COLO, CONN, MEXU, VT), 2, 00.12, 0633, 0767, 0807, 0248; ***Nicotiana plumbaginifolia* (Viv.),** 0, *Olmstead* S-54 (WTU); ***Nicotiana tabacum* (L.),** in cult. , *no voucher* 0 (0), B04, E, F, B05, C, D; ***Solanum abutiloides* (Griseb.) Bitter & Lillo,** 0, *Olmstead* S-73 (WTU), 11c (5), 0019 (1), 0621 (2), 0624 (8), 0821 (4), 0835 (3), 0839 (5), down3 (1), up3 (1); ***Solanum appendiculatum* (Dunal),** 0, *Greg Anderson* 1401 (CONN), b1 (19), b2 (3), b3 (2), 0013 (1), down3 (1), up3 (1), a (18); ***Schizanthus pinnatus* (Ruiz & Pav.),** Chile, *Olmstead* S-72 (WTU), long (3), short.a (5), short.b (14); ***Schwenckia glabrata* (Kunth),** Venezuela , *C. Benitez de Rojas* 3992 (MO), long (11), processed (3), mid (16), short

(8); ***Solanum dulcamara* (L.),** U.S.A., *no voucher*, 0005, 0450, 0455, 0987, 0993, 2, 0008 (1); ***Solanum jasminoides* (Paxton),** Bogota, *Olmstead* S-86 (WTU), 2a, 2b; ***Solanum muricatum* (Aiton),** 0, *Greg Anderson* 1461 (CONN), 0006, 2; ***Solanum phaseoloides* (Pol.),** 0, *Lynn Bohs* 0 (0), 0020, 2; ***Solanum tuberosum* (L.),** 0, *Olmstead* 1610 (WIS), 2a, 2b.

Chapter 3: How many is enough? A simple method to statistically determine when to stop sequencing PCR clones when the goal is to obtain all unique gene copies

Introduction

The evolutionary importance of gene duplication has long been recognized (Ohno 1970), and recent genome studies have illustrated a high rate of duplication and deletion occurring within populations (Kidd et al. 2008; Conrad et al. 2010). Among humans, 74% of genomic variation at the nucleotide level is attributed to copied or deleted genome sequences (Levy et al. 2007). Copy numbers for many genomic regions vary to such a large degree that they may impact gene function more than point mutations (Korbel et al. 2008; Schlattl et al. 2011). The fact that duplication and deletion determine so much of the functional variation within populations may be a recent insight, and the importance these processes contribute to the differences between species has never been better understood.

Between closely related plant species, investigations into copy number variation parallel the findings within humans. Among 80 strains of *Arabidopsis thaliana*, thousands of differences in copy number exist between coding regions (Cao et al. 2011). In maize, hundreds of genes vary in copy number and thousands are found to be absent in one inbred line compared to another (Springer et al. 2009; Swanson-Wagner et al. 2010). From these types of studies, it has become clear that even between closely related species the difference in number of gene copies for most genes is unknown.

Despite awareness of the high rate of duplication and deletion within genomes, researchers examining gene copies generally fail to ascertain any sampling confidence for the number of gene copies identified. The most common sampling strategy begins with polymerase chain reaction (PCR) to amplify members of a gene family, and then cloning to isolate unique

sequences (Howarth and Donoghue 2009; Schenk et al. 2009; Inda et al. 2010). A convenient number of the resulting colonies are then sequenced to identify different gene sequences. The number of screened colonies is rarely reported and when it is, there is often no apparent attempt to determine statistically whether the sampling scheme was adequate ((Howarth and Donoghue 2009; Schenk et al. 2009; Inda et al. 2010). We have followed this same strategy and ignored the issue a few times in our own research (Chapters 1-2). To find a better method that acknowledges the importance of copy number variation, we propose a simple method that identifies a suitable sampling strategy. We begin by summarizing the hazard for any analysis performed without thorough sampling, and then briefly summarize the literature concerning sampling, before demonstrating a simple strategy that that allows for some simple guidelines for researchers to determine when to stop searching for new sequences.

A classic problem associated with gene families is identifying gene orthology between species. Correct orthology determination is critical to accurately infer gene trees and the subsequent identification of species trees (Maddison 1997). Repeated rounds of gene duplication and loss will confound identification, but the first difficulty is in obtaining all of the copies in each genome. If some copies remain unobserved, comparing sequence similarity between sets of gene copies may identify orthology incorrectly.

Well-developed statistical estimators are rarely employed to estimate the number of gene copies present in a species. Estimates of the size of the very large gene family for olfactory receptors are a notable exception ((Glusman et al. 2000; Steiger et al. 2008). Variant sequences not yet identified in the human genome have also been estimated (Ionita-Laza et al. 2009). Many

software tools are available to estimate unseen numbers when sampling is extensive (Chao and Shen 2010; Wang 2011; Bunge et al. 2012).

For genes families with fewer copies there are few statistical tools available. A Next-Gen sequencing method (Galan et al. 2010) included a means to determine genotype confidence levels by incorporating the probability of finding artifactual sequence variants and estimating the minimum number of total sequence variants that need to be found to ensure a 0.001 probability of missing only one variant. Another study (Reeves and Olmstead 2003) identified a correlation between the total number of sequences amplified by degenerate primers and the number of unique sequences isolated. Huang and Weir (Huang and Weir 2001) applied three estimators to calculate the number of unsampled alleles in a population. All of these studies estimate the number of unseen copies from large sample sizes. This study uses computer simulation on very small sample sizes to inform decisions on when to stop searching for the sequences that have not yet been seen.

Adequate sampling in a search for an unknown total is a question applicable to many different disciplines. For DNA samples, techniques for determining the number of alleles in a population (Huang and Weir 2001) and the number of copies impacting dosage effects of human disease (Fernandez-Jimenez et al. 2011) addresses a similar question. However, no method examines this question in terms of unique copies of a genetic locus that have duplicated and undergone divergence. Distinguishing divergence between alleles at one locus and gene copies at divergent loci is difficult to quantify and further complicated by the potential for PCR- and sequencing-based differences. In this study, both alleles and PCR/sequencing based differences are assumed

to produce differences that do not exceed more than 1% difference between sequences and create unique sequences that are similar enough to each other to be recognized (Box 1).

Determining the number of unique copies of a particular gene that are present in an organism is important for understanding gene diversity and evolution, however, doing so is not a trivial matter. Since R. A. Fisher estimated the number of species in a population (Fisher et al. 1943), assessing strategies for estimating species richness has remained an active research question in statistics and ecology.

Numerous methods have been developed to estimate species numbers and they can be divided into at least five main categories: extrapolation from abundance curves and parametric and nonparametric approaches under both frequentist and Bayesian frameworks. These methods all extrapolate from the observations in a sample to estimate the number of unobserved individual types in a population with an unknown number of types. A common weakness for most methods is in underestimating the true number of unique species but most of the estimation procedures approach the true number as sample size increases.

These methods can be applied to gene families to estimate the total number of unique gene copies based upon the number found in an initial sample. Subtracting the number of copies found from this estimate would thereby provide an indirect means for guiding the molecular biologist wishing to know when to stop sampling. More direct methods exist that can also be grouped into each of the same five main categories, that focus, not on estimating species numbers, but on estimating how many new species will be observed after a second sample is taken. These resampling methods can be applied to gene families to estimate the probability of finding any

previously unobserved copies in an additional sample. Below, we review these five categories by focusing on the underlying principle used by many of the more successful nonparametric methods of sample coverage.

Species accumulation curves are a plot of species number versus a measurement of sampling, such as area, biomass, time, or individuals. This technique has been used to extrapolate estimates for species numbers for entire populations (Smith and Grassle 1977; Colwell and Coddington 1994). The species accumulation curve is not based on any statistical theory and the results are sensitive to whichever model is used to fit the curve (Smith et al. 2009). Since many models fit the curve equally well it is unclear which should be used (Smith et al. 2009).

Parametric methods, first developed by Fisher (Fisher et al. 1943), most often use the frequentist framework and commonly employ maximum likelihood estimation (MLE) with a statistical distribution to model the probabilities of observing different species. Fisher (Fisher et al. 1943) used a gamma distribution to model the Poisson parameter for distributions of observed types of butterfly species, but many other statistical distributions have been employed (Bulmer 1974; Ord and Whitmore 1986; Sichel 1986; Lloyd and Yip 1991; Coull and Agresti 1999). The same types of distributions can be used as priors under the Bayesian framework (Lewins and Joanes 1984; Rodrigues et al. 2001; Barger and Bunge 2008). However, whichever statistical framework is used, parametric approaches suffer from two difficulties: model inadequacy and parameters in the model that need to be estimated along with the estimate for total species number. No parametric models have been found that have broad applicability, nor good behavior as sample size decreases (Bunge and Fitzpatrick 1993).

Nonparametric approaches better meet the criteria of good predictors in that as sample size increases the estimator approaches the true value and has low bias (Boneh et al. 1998). Many nonparametric methods exist (Good and Toulmin 1956; Efron and Thisted 1976; Burnham and Overton 1978; Chao 1984; Agresti 1994; Boneh et al. 1998; Solow and Polasky 1999; Chao and Shen 2004). Common to many of these methods (Good and Toulmin 1956; Chao 1984; Chao and Shen 2004) is a reliance on the frequency counts of rare species. Rare species, especially those found only once, sometimes called the frequency of singletons, is an important aspect of sample coverage that is statistically better understood than species estimation.

Sample coverage is an estimate of the sum of frequencies of the species observed in the sample. For example, sample coverage close to 1.0 indicates that all species with appreciable frequencies have already been sampled. The key component to sample coverage methods is the premise that if sampling has been thorough then most gene copies should have been discovered more than once. When sampling has not been thorough and many gene copies are sampled only once, then these estimators will predict that many undiscovered copies remain.

The first coverage estimator used the number of species found once, f_1 (the frequency of singletons), and the number of samples, n, to estimate the coverage, C (Good 1953):

$$\hat{C} = 1 - f_1/n.$$

This estimate of sample coverage ($\hat{C}$) is very efficient even when compared to more complex nonparametric estimators (Esty 1986). In general, coverage estimators have the double benefit of being easy to calculate and have been shown to have better estimation properties than MLE of

the species number (Darroch and Ratcliff 1980). Additionally, Esty (Esty 1986) found coverage estimators to be less sensitive to nonrandom sampling and concluded they should be preferred to MLE of the species number.

Other successful nonparametric estimators rely on sample coverage. An estimator developed by Chao (1984) has been shown to perform at least as well as other approaches and is applicable to a wide variety of situations (Bunge et al. 1995; Walther and Moore 2005). This estimator is based on the frequency of singletons and the number of species found twice (frequency of double) to predict species number (Chao 1984). Estimators like the Chao 1984 are most efficient for larger sample sizes and behave unpredictably with sample sizes less than 100.

A PCR/cloning simulation was used to explore the utility of these estimators when sample sizes are less than 50. The lack of a satisfactory method that functions well with such small sample sizes led to the development of a set of guidelines for researchers to determine when sampling can be expected to have identified all unique gene copies.

Methods

A computer simulation was used to generate data similar to that produced by a PCR/cloning sampling strategy. For a given number of unique gene copies, N, and a given sample size, s, samples were simulated with individuals within a sample determined from a vector of probabilities, v. The vector of probabilities were either all equal, such that for $N = 4$, v = [0.25 0.25 0.25 0.25] or the vector of probabilities was randomly produced from two symmetric

Dirichlet distributions to incorporate potential PCR/cloning fluctuations resulting in different probabilities for finding each copy. The first distribution with concentration parameter 10 assigned more uniform probabilities while still allowing for minimal fluctuations. The second distribution with a concentration parameter of 5 assigned values to v with greater differences between individual probabilities.

For each simulation, the number of unique copies found was recorded as well as the number of times a unique sequence was found just once, f_1.

To find a minimum sample size that identified every copy for a given N, simulations were started and sample size increased by one until a sample size was achieved such that 99% of simulations with that sample size produced the number of unique copies equal to the true copy number N. For each N, we used 100,000 simulations to determine whether a particular sample size achieves the desired 99% threshold.

The same procedure was used with 100 non uniform probabilities drawn from the more uniform Dirichlet distribution and another 100 drawn from the less uniform Dirichlet distribution.

A parametric bootstrap-like test was developed using f_1 as a test statistic. We start with data obtained with one PCR experiment and record the number of unique copies (N_{obs}) and f_1 for this experiment. This number of unique copies found is used as the starting point for successive iterations of a bootstrap procedure. We simulate B realizations of the PCR experiment with equal

probabilities. For each bootstrap iteration, we record the number of unique copies found and f_1. The bootstrap distribution of the number of unique copies found and f_1 is then compared to the starting PCR simulation to determine whether the existence of more copies is likely. The bootstrap procedure is repeated with the true number of copies set to N_{obs}, $_{Nobs+1}$, …, N_{obs+k} for some small number k (*e.g.*, k=3).

Results

The sample size necessary to find every copy increases as the number of unique gene copies rises and as differences in probabilities of finding copies increase When each copy is sampled with equal probability, there is a 99% probability that a sample size of 16 clones will contain all three unique sequences within a gene family (Figure 1). Sampling 28 clones will insure identification for five unique loci and searching through 56 clones will find every copy in a ten member gene family 99% of the time (Figure 1).

When each unique locus is not sampled with uniform probability, the number of clone samples needed to find every copy increases (Figure 1). Under nearly uniform probabilities that each unique copy is represented in a sample, an average of 20 clones should be screened to find all three unique loci, 36 samples will contain all five unique copies, and 76 samples will contain each of ten copies 99% of the time (Figure 1a).

As sampling probabilities become less uniform, the average number of copies needed to find each locus remains the same but the variance increases (Figure 1b).

For a given sample size, a parametric bootstrap simulating 1,000 samples (B) for progressively higher numbers of unique copies produced a distribution of found copies and f_1. Comparisons between these results indicate whether further sampling is necessary (Figure 2 and 3). Two examples illustrate the procedure to determine whether more copies likely remain unidentified.

A PCR simulation was used to represent possible results from a search of 16 clones. The simulation produced four unique copies with one copy found only once ($f_1 = 1$). Sample coverage is estimated as $\hat{C} = 1 - 1/16 = 0.9375$.

To determine whether every unique copy was found, these results were compared to distributions from a bootstrap test. A parametric bootstrap with 1,000 iterations, simulated with $n = 16$ and four true unique copies ($N = N_{obs}$), produced a distribution of found copies and f_1. When the true number of unique copies is four, each unique copy is identified with high frequency and $f_1 = 1$ in 200 simulations (Figure 2a). The bootstrap distribution is consistent with the original sample, so finding four unique copies when four exist is expected.

However, in 1,000 bootstrap iterations, when the true number of unique copies is five ($N = N_{obs+1}$), each unique copy is identified with high frequency and $f_1 = 1$ in nearly 400 simulations,

but four unique copies are identified in 10% of the simulations (Figure 2b). The bootstrap distribution is consistent with a potential fifth unique copy missing from the original sample.

When the number of unique copies is six in the bootstrap test ($N = N_{obs+2}$), each unique copy is identified with high frequency, but only five copies are found in 300 simulations, and only four copies are occasionally found (Figure 2c). f_1 is frequently 0-2 and three singletons are observed with low frequency (Figure 2c). The bootstrap distribution is not consistent with a potential sixth copy missing from the original sample, although, the results in the original sample are seen when six copies exist in approximately 25 (2.5%) of the 1,000 bootstrap simulations (Figure 2c).

When the number of unique copies is seven ($N = N_{obs+3}$), each unique copy is identified with high frequency. However, fewer copies are also frequently found and f_1 varies between 0-4 (Figure 2d). The bootstrap test for seven copies resulted in fewer than 10 simulations (<1%) where four unique copies were found and indicates that the original sample is unlikely to have missed three additional copies.

The bootstrap test (k=3) indicates the four unique copies identified in the original sample may represent all unique copies, but 100 of 1,000 simulations had one additional unique copy (Figure 2b) and fewer than 10 of another 1,000 simulations had two additional unique copies. More samples should be collected to rule out the potential presence of additional unique copies.

The same parametric bootstrap procedure was performed for a situation when 16 clones were sampled and five unique copies were found (Figure 3). Comparisons to the bootstrap distributions indicate that the original sample may have missed identifying up to three unique copies and more clones should be sampled.

Discussion

In general, studies reporting gene copy numbers lack a statistical framework to report confidence in sampling strategy. The dearth of methods inspired this investigation into a statistical method to determine sampling confidence for gene family studies. PCR simulations were performed to identify the number of samples necessary to find every copy with 99% confidence. A researcher now can easily identify the number of samples necessary to confidently sample all of the copies in a small to moderately sized gene family (Figure 1).

A parametric bootstrap procedure was performed to determine whether each unique copy was identified after a search for gene copies is completed. Researchers can compare the number of unique copies found and the number of singletons (f_1) for the number of samples screened to distributions of these values under the possibility that more copies remain unidentified. In the example where four copies are identified, it is possible that another unseen copy exists (Figure 2b), although probably not more than five (Figure 2c-d), and more samples should be screened.

In the example where five copies are identified in 16 clones, it is probable that more unseen copies exist since five copies are often identified when the true number of unique copies is six or seven (Figure 3).

Estimators of coverage based upon values such as f_1 are unreliable for determining coverage reliably when sample sizes are small since a range of values are found at high frequencies. However, the value of f_1 combined with the number of copies found can contribute to informing a researcher whether more sampling is necessary.

It is our hope that researchers will report the number of clones sampled and the frequency of copies found only once when publishing results from studies on sequences of gene families. These values can be compared to distributions from the parametric bootstrap test to determine whether further sampling is warranted.

The guidelines reported from the PCR/cloning simulations should be viewed conservatively. When PCR/cloning reactions are performed in the lab numerous challenges may artificially raise or lower the number of unique copies present in the sample (Box 1). These simulations attempt to incorporate some of the variability inherent in PCR and cloning by allowing non-uniform probabilities for each locus to be sampled, but researchers should incorporate strategies like pooling separate PCRs, limiting template concentrations, and keeping cycle numbers low (Box 1). The results highlight how even moderate differences in the probability each locus is represented in a sample can drastically increase the number of sample clones that must be screened for unique sequences.

Bibliography

Acinas, S. G., R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz. 2005. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Appl. Environ. Microbiol. 71:8966–8969.

Agresti, A. 1994. Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort. Biometrics 50:494.

Barger, K., and J. Bunge. 2008. Bayesian Estimation of the Number of Species using Noninformative Priors. Biom. J. 50:1064–1076.

Beerenwinkel, N., H. F. Gunthard, V. Roth, and K. J. Metzner. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3.

Boneh, S., A. Boneh, and R. J. Caron. 1998. Estimating the Prediction Function and the Number of Unseen Species in Sampling with Replacement. J. Am. Stat. Assoc. 93:372–379.

Borriello, F., and K. S. Krauter. 1990. Reactive site polymorphism in the murine protease inhibitor gene family is delineated using a modification of the PCR reaction (PCR + 1). Nucleic Acids Res. 18:5481–5487.

Bradley, R. D., and D. M. Hillis. 1997. Recombinant DNA sequences generated by PCR amplification. Mol. Biol. Evol. 14:592–593.

Bulmer, M. G. 1974. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. Biometrics 30:101.

Bunge, J., and M. Fitzpatrick. 1993. Estimating the Number of Species: A Review. J. Am. Stat. Assoc. 88:364–373.

Bunge, J., M. Fitzpatrick, and J. Handley. 1995. Comparison of three estimators of the number of species. J. Appl. Stat. 22:45–59.

Bunge, J., L. Woodard, D. Böhning, J. A. Foster, S. Connolly, and H. K. Allen. 2012. Estimating population diversity with CatchAll. Bioinformatics 28:1045–1047.

Burnham, K. P., and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:625–633.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Müller, C. Alonso-Blanco, K. Borgwardt, K. J. Schmid, and D. Weigel. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43:956–963.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scand. J. Stat. 11:265–270.

Chao, A., and T. Shen. 2010. SPADE: Species Prediction And Diversity Estimation. Program and user's guide.

Chao, A., and T.-J. Shen. 2004. Nonparametric prediction in species sampling. J. Agric. Biol. Environ. Stat. 9:253–269.

Colwell, R. K., and J. A. Coddington. 1994. Estimating Terrestrial Biodiversity through Extrapolation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 345:101–118.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.

Coull, B. A., and A. Agresti. 1999. The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies. Biometrics 55:294–301.

Cronn, R., B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99:291–311.

Darroch, J. N., and D. Ratcliff. 1980. A Note on Capture-Recapture Estimation. Biometrics 36:149.

Efron, B., and R. Thisted. 1976. Estimating the number of unseen species: How many words did Shakespeare know? Biometrika 63:435–447.

Esty, W. W. 1986. The Efficiency of Good's Nonparametric Coverage Estimator. Ann. Stat. 14:1257–1260.

Fernandez-Jimenez, N., A. Castellanos-Rubio, L. Plaza-Izurieta, G. Gutierrez, I. Irastorza, L. Castaño, J. C. Vitoria, and J. R. Bilbao. 2011. Accuracy in Copy Number Calling by qPCR and PRT: A Matter of DNA. PLoS ONE 6:e28910.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J. Anim. Ecol. 12:42.

Galan, M., E. Guivier, G. Caraux, N. Charbonnel, and J.-F. Cosson. 2010. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. BMC Genomics 11:296.

Glusman, G., A. Bahar, D. Sharon, Y. Pilpel, J. White, and D. Lancet. 2000. The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm. Genome 11:1016–1023.

Good, I. J. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika 40:237–264.

Good, I. J., and G. H. Toulmin. 1956. The Number of New Species, and the Increase in Population Coverage, When a Sample Is Increased. Biometrika 43:45–63.

Howarth, D. G., and M. J. Donoghue. 2009. Duplications and Expression of DIVARICATA-Like Genes in Dipsacales. Mol. Biol. Evol. 26:1245–1258.

Huang, S.-P., and B. S. Weir. 2001. Estimating the Total Number of Alleles Using a Sample Coverage Method. Genetics 159:1365–1373.

Huber, T., G. Faulkner, and P. Hugenholtz. 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20:2317–2319.

Inda, L. A., M. Pimentel, and M. W. Chase. 2010. Chalcone synthase variation and phylogenetic relationships in Dactylorhiza (Orchidaceae). Bot. J. Linn. Soc. 163:155–165.

Ionita-Laza, I., C. Lange, and N. M. Laird. 2009. Estimating the number of unseen variants in the human genome. Proc. Natl. Acad. Sci. 106:5008–5013.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tüzün, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, and E. E. Eichler. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64.

Kobayashi, N., K. Tamura, and T. Aotsuka. 1999. PCR Error and Molecular Population Genetics. Biochem. Genet. 37:317–321.

Korbel, J. O., P. M. Kim, X. Chen, A. E. Urban, S. Weissman, M. Snyder, and M. B. Gerstein. 2008. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr. Opin. Struct. Biol. 18:366–374.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. C. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y.-H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter. 2007. The Diploid Genome Sequence of an Individual Human. PLoS Biol 5:e254.

Lewins, W. A., and D. N. Joanes. 1984. Bayesian Estimation of the Number of Species. Biometrics 40:323.

Lloyd, C. J., and P. Yip. 1991. A unification of inference from capture-recapture studies through martingale estimating functions. Estim. Equ. 65–88.

Maddison, W. P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Ohno, S. 1970. Evolution by gene duplication. xv + 160 pp.

Ord, J. K., and G. A. Whitmore. 1986. The poisson-inverse gaussian disiribuiion as a model for species abundance. Commun. Stat. - Theory Methods 15:853–871.

Qiu, X., L. Wu, H. Huang, P. E. McDonel, A. V. Palumbo, J. M. Tiedje, and J. Zhou. 2001. Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning. Appl. Environ. Microbiol. 67:880–887.

Reeves, P. A., and R. G. Olmstead. 2003. Evolution of the TCP Gene Family in Asteridae: Cladistic and Network Approaches to Understanding Regulatory Gene Family Diversification and Its Impact on Morphological Evolution. Mol. Biol. Evol. 20:1997–2009.

Rodrigues, J., L. A. Milan, and J. G. Leite. 2001. Hierarchical Bayesian Estimation for the Number of Species. Biom. J. 43:737–746.

Saitoh, K., and W.-J. Chen. 2008. Reducing cloning artifacts for recovery of allelic sequences by T7 endonuclease I cleavage and single re-extension of PCR products — A benchmark. Gene 423:92–95.

Scharf, S. J., C. M. Long, and H. A. Erlich. 1988. Sequence analysis of the HLA-DRß and HLA-DQß loci from three Pemphigus vulgaris patients. Hum. Immunol. 22:61–69.

Schenk, M. F., J. H. Cordewener, A. H. America, W. P. van't Westende, M. J. Smulders, and L. J. Gilissen. 2009. Characterization of PR-10 genes from eight Betula species and detection of Bet v 1 isoforms in birch pollen. BMC Plant Biol. 9:24.

Schlattl, A., S. Anders, S. M. Waszak, W. Huber, and J. O. Korbel. 2011. Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. Genome Res. 21:2004–2013.

Shafikhani, S. 2002. Factors affecting PCR-mediated recombination. Environ. Microbiol. 4:482–486.

Sichel, H. . 1986. Parameter estimation for a word frequency distribution based on occupancy theory. Commun. Stat. - Theory Methods 15:935–949.

Smith, M. A., J. Fernandez-Triana, R. Roughley, and P. D. N. Hebert. 2009. DNA barcode accumulation curves for understudied taxa and areas. Mol. Ecol. Resour. 9:208–216.

Smith, W., and J. F. Grassle. 1977. Sampling Properties of a Family of Diversity Measures. Biometrics 33:283.

Solow, A. R., and S. Polasky. 1999. A QUICK ESTIMATOR FOR TAXONOMIC SURVEYS. Ecology 80:2799–2803.

Springer, N. M., K. Ying, Y. Fu, T. Ji, C.-T. Yeh, Y. Jia, W. Wu, T. Richmond, J. Kitzman, H. Rosenbaum, A. L. Iniguez, W. B. Barbazuk, J. A. Jeddeloh, D. Nettleton, and P. S. Schnable. 2009. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5:e1000734.

Steiger, S. S., A. E. Fidler, M. Valcu, and B. Kempenaers. 2008. Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? Proc. R. Soc. B Biol. Sci. 275:2309–2317.

Swanson-Wagner, R. A., S. R. Eichten, S. Kumari, P. Tiffin, J. C. Stein, D. Ware, and N. M. Springer. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20:1689–1699.

Thompson, J. R., L. A. Marcelino, and M. F. Polz. 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR." Nucleic Acids Res. 30:2083–2088.

Wagner, A., N. Blackstone, P. Cartwright, M. Dick, B. Misof, P. Snow, G. P. Wagner, J. Bartels, M. Murtha, and J. Pendleton. 1994. Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift. Syst. Biol. 43:250.

Walther, B. A., and J. L. Moore. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815–829.

Wang, G. C., and Y. Wang. 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl. Environ. Microbiol. 63:4645–4650.

Wang, J.-P. 2011. SPECIES: An R Package for Species Richness Estimation. J. Stat. Softw. 40:1–15.

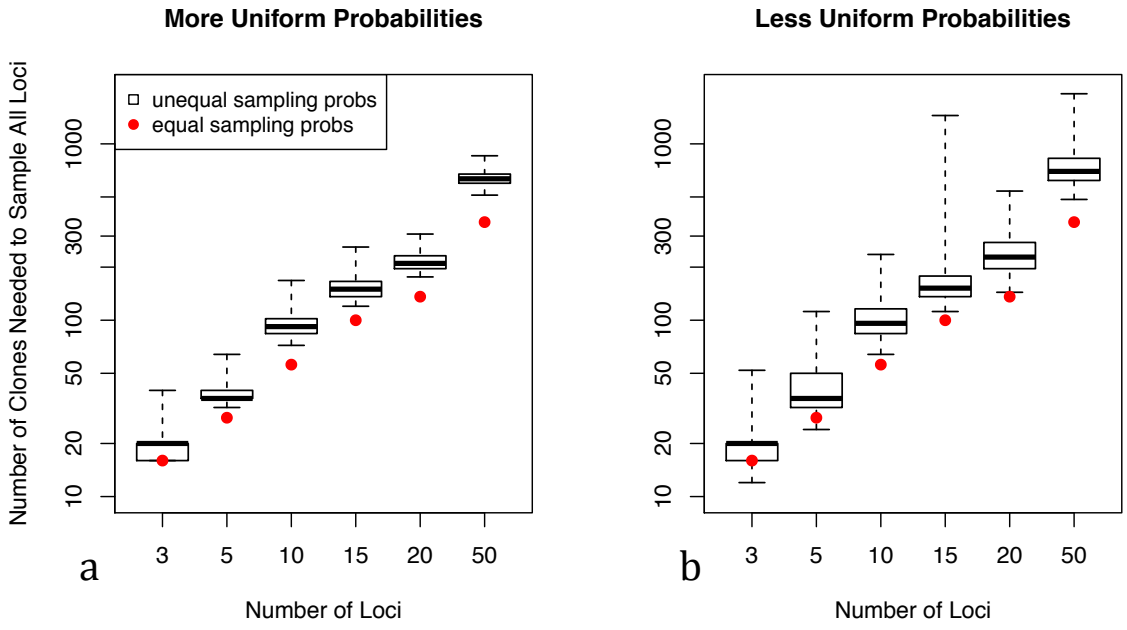


Figure 1. Sample size needed to reach 99% probability that all unique copies are identified in 100,000 simulations. For each number of unique loci, the sample size of clones needed to find all unique copies is depicted on a logarithm scale. Each unique copy is sampled with equal probability (red dots) or non uniform probabilities (boxplots). Each boxplot reflects 100 draws from two Dirichlet distributions, a: more uniform probabilities for sampling each unique copy, and b: less uniform.

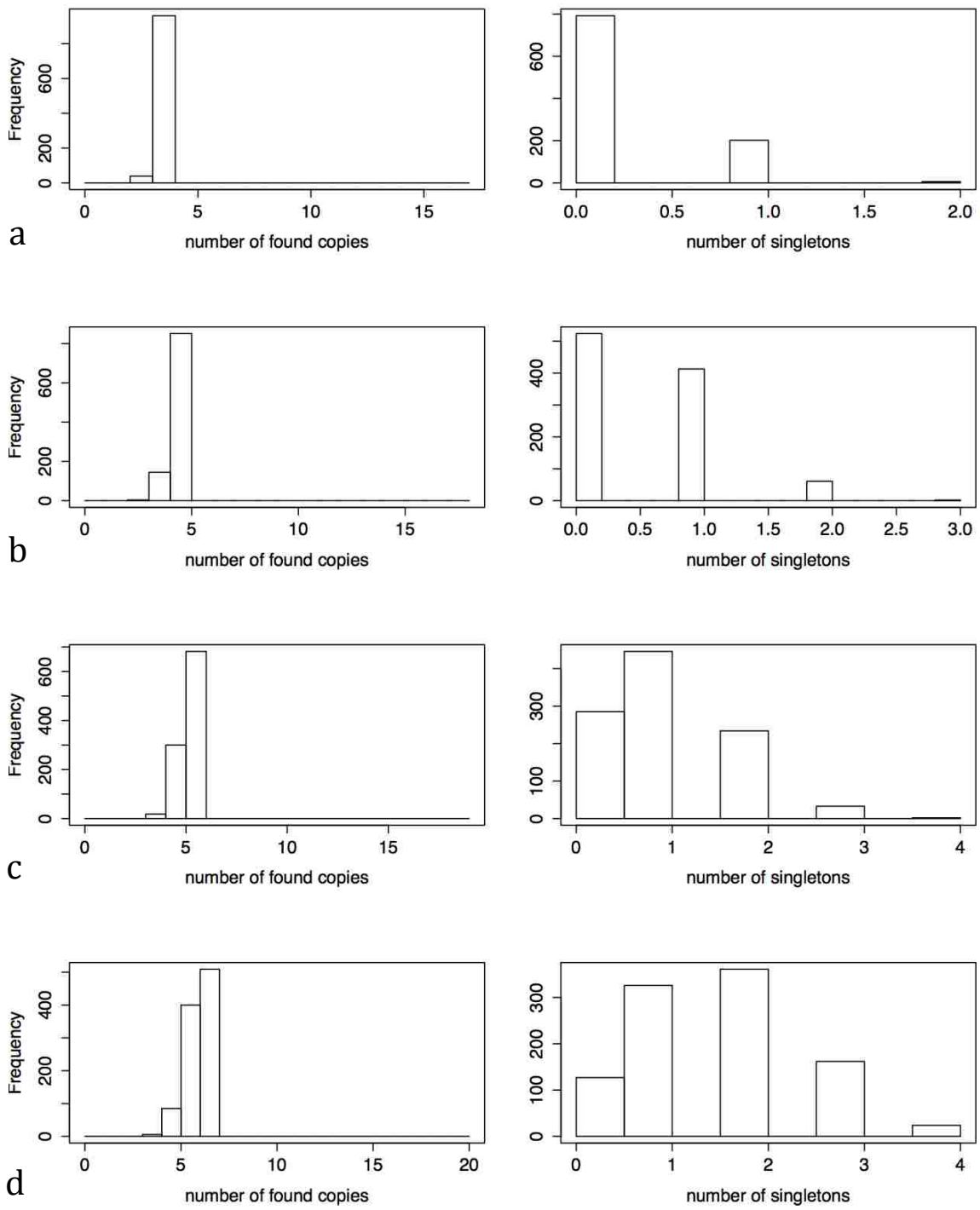


Figure 2. Parametric bootstrap test of 1,000 simulations showing the frequency of found copies and number of singletons where the true number of unique copies equals a: 4, b: 5, c: 6, d: 7.

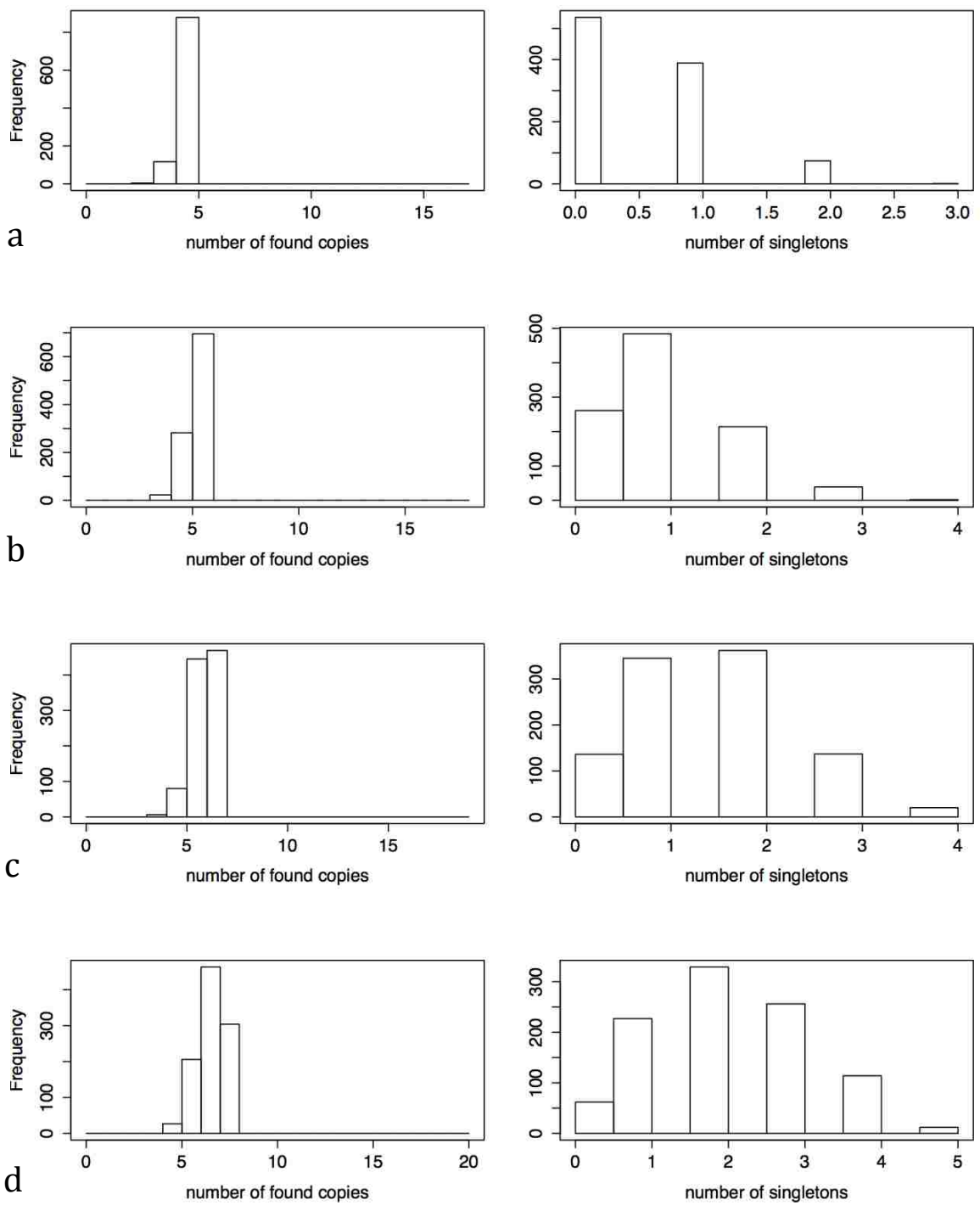


Figure 3. Parametric bootstrap test of 1,000 simulations showing the frequency of found copies and number of singletons where the true number of unique copies equals a: 5, b: 6, c: 7, d: 8.

Box 1. Summary of challenges and methods to improve the PCR/cloning sampling strategy.

Polymerase Chain Reaction (PCR) is commonly used to obtain members of a gene family. New methods utilizing Next-Generation Sequencing can more thoroughly sample members of large gene families (Cronn et al. 2012) but separating actual variants present in the sample from variants created during sequencing remains a challenge (Beerenwinkel et al. 2012). For low-copy number genes, PCR sampling is still the most common means to examine gene copies. The strategy most often employed utilizes PCR to amplify multiple templates simultaneously using the same primer pair. PCR products are then cloned to separate them and then sequenced individually. Two well-documented issues arise when multiple templates are amplified together in PCR and affect the number of unique PCR products differently: sequence artifacts and PCR bias (Thompson et al. 2002).

Sequence artifacts

Sequence artifacts are novel sequence types produced during PCR that do not occur in the original template. Only 60% of *Taq* amplified products retain their starting DNA sequence (Kobayashi et al. 1999). New sequence types can be created through misincorporation of nucleotides by DNA polymerase or through PCR. Simple methods have been shown to reduce these types of sequence artifacts. Enzymes with proof reading ability, such as *Pfu* polymerase, can be used with *Taq* polymerase (Kobayashi et al. 1999) or alone to reduce the number of misincorporated nucleotides (Shafikhani 2002).

Two types of recombinant sequence types can form during PCR. Chimera molecules are formed through *in vitro* recombination when DNA polymerase falls off the template before completely

extending the new DNA molecule in a cycle of PCR. The partially extended molecule can then anneal to a heterologous template molecule in the next cycle and be extended to form a single stranded chimera molecule that represents a new product not present in the template genome (Scharf et al. 1988; Bradley and Hillis 1997; Wang and Wang 1997). The percentage of chimeric molecules increases with template number (Wang and Wang 1997). Lengthening extension times, reducing PCR cycles, and insuring high primer concentrations relative to product have been shown to reduce the formation of chimera molecules (Wang and Wang 1997; Qiu et al. 2001; Thompson et al. 2002). A second type of recombinant sequence may form during PCR from heteroduplex products. Heteroduplex molecules are formed during the later cycles of PCR when two different single stranded DNA molecules anneal to produce a hybrid or heteroduplex DNA molecule (Borriello and Krauter 1990; Thompson et al. 2002). If the heteroduplex DNA is cloned into a host with mismatch repair (such as the commonly used DH5α strain of *E. coli*) then mosaic sequences can be formed through *in vivo* recombination artificially elevating the number of unique sequence types in the clone library (Shafikhani 2002). Heteroduplex molecules can be separated through a re-extension method before cloning (Saitoh and Chen 2008). Re-extension is a simple protocol modification that eliminates heteroduplex molecules through a single cycle of PCR employing a 30 minute extension step (Saitoh and Chen 2008).

PCR bias

Two general types of PCR bias can occur to decrease the number of observable gene copies: PCR selection and PCR drift. Intrinsic differences between templates cause PCR selection by preferentially amplifying one or more copies over others. Selection for certain copies over others is evident when multiple reactions are run in parallel. Reducing the concentration of starting genomic DNA and/or reducing the number of PCR cycles can minimize the effects of PCR selection (Wagner et al. 1994). PCR drift occurs when stochastic processes in early PCR cycles lead to differences in product concentrations between identical reactions. Running several independent reactions and pooling the products can reduce the variation caused by PCR drift (Wagner et al. 1994).

Cloning

PCR products containing more than one sequence type must be cloned in order to determine the nucleotide sequence. Choosing the number of clones to sequence is a trade-off between satisfactory sampling and cost or time. Unfortunately, cloning is also a stochastic process where template length and composition can affect cloning efficiency. Differing lengths and nucleotide structures may limit incorporation into the cloning vector and affect the final distribution in the sample of sequences represented in the bacterial colonies (Invitrogen product information: http://tools.invitrogen.com/content/sfs/appendix/Cloning_Trans/PCR Cloning Considerations.pdf).

DNA sequencing and copy identification

Determining the DNA sequence of the cloned fragments first requires amplification by either bacterial miniprep or by PCR. Both methods amplify a single cloned product for direct sequencing. The initial template may contain previously incorporated errors, however, this last amplification process uses only one template (assuming the colony grew from a bacterium without a heteroduplex insert) and does not suffer from most of the challenges discussed above.

Unique sequence types are identified by alignment after DNA sequencing. Typically, a 1% cut-off is used to cluster sequences by similarity so that misincorporated nucleotides do not falsely increase the count of unique sequences (Acinas et al. 2005). For environmental sampling, chimeras can be identified using software programs like Bellerophon (Huber et al. 2004), that rely on partial tree analysis to identify sequence changes that substantially alter sequence relationships. However, this type of analysis has the potential to falsely implicate actual gene copies present in the genome that have recently undergone gene conversion.

These challenges to PCR surveys of gene copies are ignored at the researcher's peril. Some processes, like PCR bias, can prevent the amplification of copies and lead to underestimation of diversity in sequence types. Other processes leading to chimeras or heterduplexes may artificially increase the number of unique sequence types among the PCR products and obscure the true number of unique genomic copies. All of these well-described challenges need to be addressed by researchers amplifying numerous gene copies.