DOI: 10.1534/genetics.106.062455

Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade

Feinan Wu,[*,†] Lukas A. Mueller,[*,†] Dominique Crouzillat,[‡] Vincent Pétiard[‡] and Steven D. Tanksley[*,†,1]

[*]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, [†]Department of Plant Biology, Cornell University, Ithaca, New York 14853 and [‡]Nestlé Research Center, 37097 Tours Cedex 2, France

Manuscript received June 23, 2006
Accepted for publication August 8, 2006

ABSTRACT

We report herein the application of a set of algorithms to identify a large number (2869) of single-copy orthologs (COSII), which are shared by most, if not all, euasterid plant species as well as the model species Arabidopsis. Alignments of the orthologous sequences across multiple species enabled the design of "universal PCR primers," which can be used to amplify the corresponding orthologs from a broad range of taxa, including those lacking any sequence databases. Functional annotation revealed that these conserved, single-copy orthologs encode a higher-than-expected frequency of proteins transported and utilized in organelles and a paucity of proteins associated with cell walls, protein kinases, transcription factors, and signal transduction. The enabling power of this new ortholog resource was demonstrated in phylogenetic studies, as well as in comparative mapping across the plant families tomato (family Solanaceae) and coffee (family Rubiaceae). The combined results of these studies provide compelling evidence that (1) the ancestral species that gave rise to the core euasterid families Solanaceae and Rubiaceae had a basic chromosome number of $x = 11$ or 12.2) No whole-genome duplication event (*i.e.*, polyploidization) occurred immediately prior to or after the radiation of either Solanaceae or Rubiaceae as has been recently suggested.

ORTHOLOGS are defined as genes sharing a common ancestor by speciation. By contrast, paralogs are duplicated copies within a genome and arise through such phenomena as polyploidization or tandem duplications (Gogarten and Olendzenski 1999; Sonnhammer and Koonin 2002). Establishing orthology in divergent species, while theoretically possible, is not a trivial exercise. In plants, identifying true orthologs is further complicated by the fact that most plants are paleopolyploids and extensive gene duplication events have occurred during their evolution (Ku *et al.* 2000; Fulton *et al.* 2002; Blanc and Wolfe 2004; O'Brien *et al.* 2005). Yet many aspects of comparative genomics, systematics, and evolutionary biology, including comparative genome mapping (Paterson *et al.* 2000), reconstruction of ancestral genomes (Blanchette *et al.* 2004), phylogenetic studies (Rokas *et al.* 2003), deciphering patterns of natural selection on coding regions (Bustamante *et al.* 2005), and predictions of common gene function across species (Eisen 1998; Doganlar *et al.* 2002), depend on the availability of validated sets of orthologous genes.

The recent development of sequence data sets for many species raises the possibility of using computational methods to assist in identifying sets of putative orthologs across multiple species (Fulton *et al.* 2002). However, most ortholog-finding algorithms are designed for the use of complete genome sequences for the two or more species being compared (Tatusov *et al.* 2000; Remm *et al.* 2001; Lee *et al.* 2002; Li *et al.* 2003). Less well developed are the methods that can be used to identify and validate orthologs using multiple, incomplete sequence databases (*e.g.*, EST databases) (Fulton *et al.* 2002; Lee *et al.* 2002). Most sequence databases are currently incomplete, necessitating the development of methods for identifying and validating orthologs in such data sets.

Using multiple, incomplete EST databases for species in the euasterid clade of flowering plants, we have endeavored to apply a combined set of computational and phylogenetic algorithms to identify, verify, and annotate a large set (2869) of conserved, single-copy, putatively orthologous genes. Moreover, we demonstrate the use of this new ortholog resource to shed light on issues related to comparative genomics, molecular systematics, and gene evolution studies in the euasterid clade.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ422968–DQ423202.

[1] *Corresponding author:* Department of Plant Breeding and Genetics, Department of Plant Biology, 248 Emerson Hall, Cornell University, Ithaca, NY 14853. E-mail: sdt4@cornell.edu

Euasterid species were chosen for this study because they represent the largest clade of flowering plants, encompassing >75,000 or one-quarter of the estimated 300,000 flowering plant species that currently occupy the earth. Included in this clade are many important and diverse crop species, including tomato, potato, eggplant, pepper, petunia, tobacco, coffee, sweet potato, olives, mints, sesame, sunflower, and lettuce, as well as model species for evolutionary biology and floral development such as Mimulus, Antirrhinum, and Petunia (Figure 1A).

MATERIALS AND METHODS

EST-derived unigene data sets and Arabidopsis genomic data sets: The Arabidopsis genomic data sets [including gene sequences, translated peptide sequences, and coding sequences (CDS)] and euasterid unigene data sets used in this research are listed in Table 1. Detailed information about the cDNA libraries, EST collection, and unigene assembly can be obtained from the SOL Genomics Network (SGN) (http://www.sgn.cornell.edu/content/sgn_data.pl) and The Institute for Genomic Research (TIGR) (http://www.tigr.org/tdb/tgi/index.shtml).

BLAST searches: BLAST search was implemented by the BLASTALL program (ftp://ftp.ncbi.nih.gov/blast/executables/release/2. 2. 6/).

Selecting single-copy genes: To establish a useful criterion for declaring a gene as single copy, each of the five data sets was blasted against itself using BLASTN. At a relaxed criterion (*e*-value ≥1E-10), 49% (14,847) of the tomato genes, 52% (13,024) of the potato genes, 74% (7052) of the pepper genes, 65% (8600) of the coffee genes, and 43% (12,265) of the Arabidopsis genes were classified as single copy. The lowest percentage (43%) for the Arabidopsis genome probably reflects the fact that Arabidopsis is the only fully sequenced genome, a percentage that is similar to that previously reported for single-copy genes (Arabidopsis Genome Initiative 2000). Therefore, for this study, a gene was regarded as single copy if the expect value of its most similar copy in self-BLASTN was ≥1E-10. Additionally, the copy number of a gene was defined as the number of blast hits (including itself) with an expect value ≤1E-10 in self-BLASTN.

Use of phylogenetics to test orthologies of COSII genes: For each COSII group, the most suitable DNA substitution model was chosen by MODELTEST (Posada and Crandall 1998) via a likelihood-ratio test between a null model (*i.e.*, equal base frequency, equal transition rates, and equal transversion rates, rates equal among sites and no invariable sites) and each of the other 55 complex models. Subsequently, the gene tree was reconstructed using the maximum-likelihood method in PAUP*4.0 (Swofford 2003) and the above DNA substitution model based on the overlapping region shared by all members in the multiple species alignment. The tree was then rooted using Arabidopsis. Bootstrapping was done with 500 replications.

Multiple species alignments: To obtain the most accurate alignments for a COSII group, the following protocol was followed:

1. Each euasterid I member was translated into peptide sequence in the frame that yielded the highest BLASTX score against the Arabidopsis member in the group.
2. Peptide sequences of all members were aligned by T_COFFEE (Notredame *et al.* 2000) and the corresponding DNA sequence alignment was produced accordingly.
3. 5′- and 3′-UTRs were trimmed according to the alignment with the Arabidopsis CDS sequence.
4. Each euasterid I member was further subjected to visual error correction (*e.g.*, frameshift caused by sequencing errors) on the basis of the alignment.

Design of universal primers: On the basis of the above multiple sequence alignments, universal primers for a COSII group were designed to meet the following criteria:

1. The 3′-end of each primer had to match a contiguous stretch of nucleotides (at least eight nucleotides) shared by all the euasterid I species in the alignment. The remaining 5′ portion of the primer was composed of a tomato-specific sequence (or potato, if no tomato was in the group, or pepper, if neither tomato nor potato were available) because the tomato and potato sequences tended to be longer and of better quality.
2. Total primer length had to be 20–30 nucleotides with melting temperature (T_m) = 55°–65°.
3. Primer dimer and self-complementary primers were excluded. OLIGOS 8.62 (Kalendar 2001) was used to calculate T_m and to detect primer dimmers and self-complementary primers.

Mapping of COSII genes in tomato as PCR-based markers: PCR by universal primers and cleaved amplified polymorphic sequence assays were applied on an F_2 population of 80 individuals derived from the interspecific cross *Solanaceae lycopersicum* × *S. pennellii* (Fulton *et al.* 2002; Frary *et al.* 2005).

Mapping of COSII genes in coffee: A diploid coffee population, derived from a cross between a Congolese robusta clone (BP409) and a hybrid type between Congolese and Guinean (Q121) and consisting of 93 individual genotypes, was used to generate a consensus genetic map. JoinMap software version 3.0 (Stam 1993; Van Ooijen 2001) was used for linkage analysis and map calculations.

Data availability: A COSII gene member list, multiple sequence alignments, gene trees, details of the mapping experiments, and the map positions can be found in the SGN (http://www.sgn.cornell.edu/markers/cosii_markers.pl).

RESULTS AND DISCUSSION

The first part of this project focused on computational analysis of partial sequence data sets (EST-derived unigenes) from species in the euasterid I clade of eudicot species (which include such cultivated species as coffee, tomato, potato, and pepper), using Arabidopsis as an outgroup (Figure 1, Table 1). As a result, 2869 putative conserved ortholog groups (referred to as COSII genes) were identified, validated, and annotated. The challenges of identifying orthologous gene sets in such partial data sets and the methodology that we have implemented to overcome these problems are described below.

Screening for putative orthologs using the reciprocal best match method: Identification of orthologs *in silico* is based on the assumption that orthologous genes begin diverging after speciation (Figure 2A). Thus, in a comparison of the complete sequence of two genomes, a pair of orthologous genes will, in most cases, be *r*eciprocal *b*est *m*atches (RBM) in BLAST comparisons (Li *et al.* 2003). Likewise, if the genome sequence of

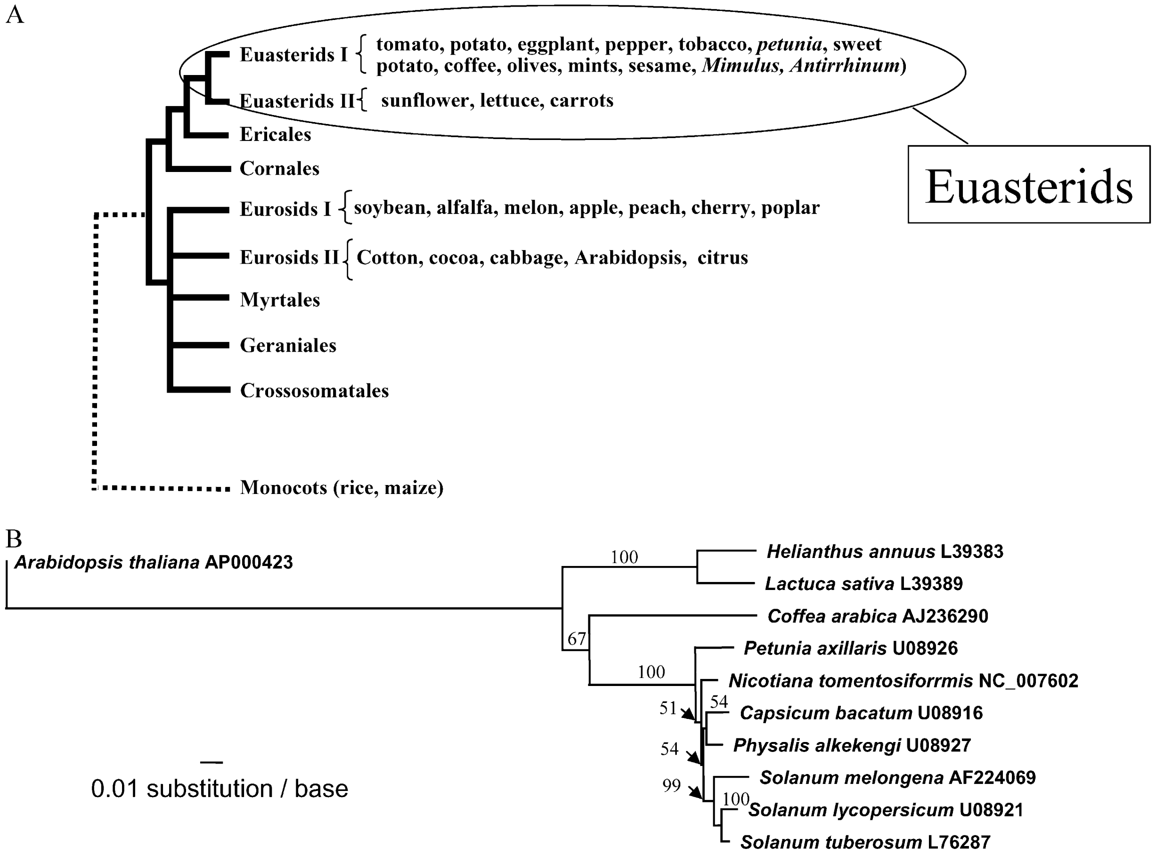
A
Euasterids I { tomato, potato, eggplant, pepper, tobacco, *petunia*, sweet potato, coffee, olives, mints, sesame, *Mimulus, Antirrhinum*)
Euasterids II{ sunflower, lettuce, carrots
Ericales
Cornales
Eurosids I { soybean, alfalfa, melon, apple, peach, cherry, poplar
Eurosids II { Cotton, cocoa, cabbage, Arabidopsis, citrus
Myrtales
Geraniales
Crossosomatales
Monocots (rice, maize)
Euasterids
B
Arabidopsis thaliana AP000423
100
Helianthus annuus L39383
Lactuca sativa L39389
Coffea arabica AJ236290
67
Petunia axillaris U08926
100
Nicotiana tomentosiforrmis NC_007602
51
54
Capsicum bacatum U08916
54
Physalis alkekengi U08927
99
Solanum melongena AF224069
100
Solanum lycopersicum U08921
Solanum tuberosum L76287
0.01 substitution / base

FIGURE 1.—(A) Phylogenetic tree showing relationship placement of euasterids to other eudicot plant species on the basis of APG II 2003 (BREMER *et al.* 2003). (B) Phylogenetic relationships of the plant species included in this study. The maximum-likelihood tree was reconstructed using published chloroplast *ndhF* sequences (each species name is followed by its GenBank accession number). Bootstrap values are placed on the branches. The tree is consistent with previous reports (CHASE *et al.* 1993; OLMSTEAD 1999).

a third species is added, all pairwise comparisons of genomes will yield RBM triangles (Figure 2, A–D). This RBM method has been effective in computational screening for orthologs among fully sequenced genomes (TATUSOV *et al.* 2000; REMM *et al.* 2001; LEE *et al.* 2002; LI *et al.* 2003).

Complete genome sequence is available for only a few plants species (*e.g.*, Arabidopsis and rice) while the vast majority of plant sequence data sets (*e.g.*, EST data sets) contain only a portion of the gene repertoires. When the RBM method is applied to two incomplete genome data sets, there is a significant probability that the RBM method will mistakenly pair up paralogs. For example, if gene duplication preceded speciation, both resulting species would carry paralogous copies of each gene (Figure 2C). If the data sets for any two species being compared are incomplete, and each is missing one of the gene copies, the possibility exists that the RBM method would identify paralogs (*e.g.*, A2 and B1 in Figure 2E). A similar false pairing of paralogs could also result from extinct lineages and/or gene extinction (Figure 2, E and F). One way to mitigate this problem is to include data sets for three or more species. If a third species were added in which both copies were present, the resulting RBM triangle would be internally inconsistent (Figure 2E); in other words, RBM pairs (A2, B1), (A2, C2), and (B1, C1) could not form a closed triangle, thus revealing that A2 and B1 are paralogs rather than orthologs. Nonetheless, if all of the data sets are incomplete with regards to the genes being compared, the RBM triangulation method can still lead to false pairing of paralogs (Figure 2F). Resolving this problem requires the phylogenetic analysis described later.

Application of the RBM triangulation method to data sets from Arabidopsis and euasterid I species: The euasterid I clade of eudicot species is ideally suited for testing the RBM triangulation method for identifying orthologs from partial-sequence data sets as EST-derived unigene data sets currently exist for many euasterid I species (Table 1). We have thus endeavored to use the RBM triangulation method, combined with phylogenetic analysis, to identify *c*onserved *o*rthologs *s*et (COS) genes in species from the euasterid I clade of eudicots. We set the further requirement that the identified orthologs must have a single gene match in Arabidopsis, which belongs to a separate and divergent clade of

TABLE 1

Species and sequence data sets used for computational screening of COSII genes

Common name	Species	Resource	Statistics	URL
Arabidopsis	*Arabidopsis thaliana*	TAIR	28,581 sequences; from 60 to 15468 bp; 1274 bp on average	ftp://tairpub:tairpub@ftp.Arabidopsis.org/home/tair/Sequences/blast_datasets/
Tomato	*S. lycopersicum* and *S. pennellii*	SGN	30576 sequences; from 89 to 4127 bp; 773 bp on average	ftp://ftp.sgn.cornell.edu/unigene_builds/new-tomato.seq
Potato	*S. tuberosum*	SGN	24,931 sequences; from 151 to 4200 bp; 740 bp on average	ftp.sgn.cornell.edu/unigene_builds/Solanum_tuberosum.seq
Pepper	*Capsicum annuum*	SGN	9554 sequences; from 150 to 3182 bp; 556 bp on average	ftp.sgn.cornell.edu/unigene_builds/Capsicum_combined.seq
Coffee	*Coffea canephora var. robusta*	CGN	13,175 sequences; from 150 to 2714 bp; 677 bp on average	http://coffee.pgn.cornell.edu
Sunflower	*Helianthus annuus*	TIGR	20,520 sequences; from 100 to 4587 bp; 478 bp on average	ftp://ftp.tigr.org/pub/data/tgi/elianthus_annuus
Lettuce	*Lactuca sativa*	TIGR	22,185 sequences; from 100 to 5544 bp; 632 bp on average	ftp://ftp.tigr.org/pub/data/tgi/Lactuca_sativa

eudicots (eurosid) (Figure 1A). The requirement that the identified orthologs have one and only one gene match in Arabidopsis creates a gene-for-gene link with this important model plant and potentially allows extension of these COS genes into other more distantly related plant species (Figure 1A).

The first step toward identifying COS genes in euasterid I species was to apply BLASTN for pairwise comparisons of the four largest euasterid I unigene data sets (tomato, potato, pepper, and coffee) (Table 1). Comparisons were made at the nucleotide level, not at the translated peptide level. The reason is that ESTs are single-pass sequenced; thus sequencing errors are common and can result in frameshifts that would invalidate peptide comparisons. Pairwise comparisons were also applied to each euasterid I unigene data set and the Arabidopsis translated peptide data set using TBLASTN/BLASTX because of high divergence in nucleotide sequence between euasterid I species and Arabidopsis. RBM triangles were then identified among Arabidopsis and each of two euasterid I species, which resulted in six types of RBM triangles (Table S1 at http://www.genetics.org/supplemental/). Subsequent combining of RBM triangles with a common Arabidopsis

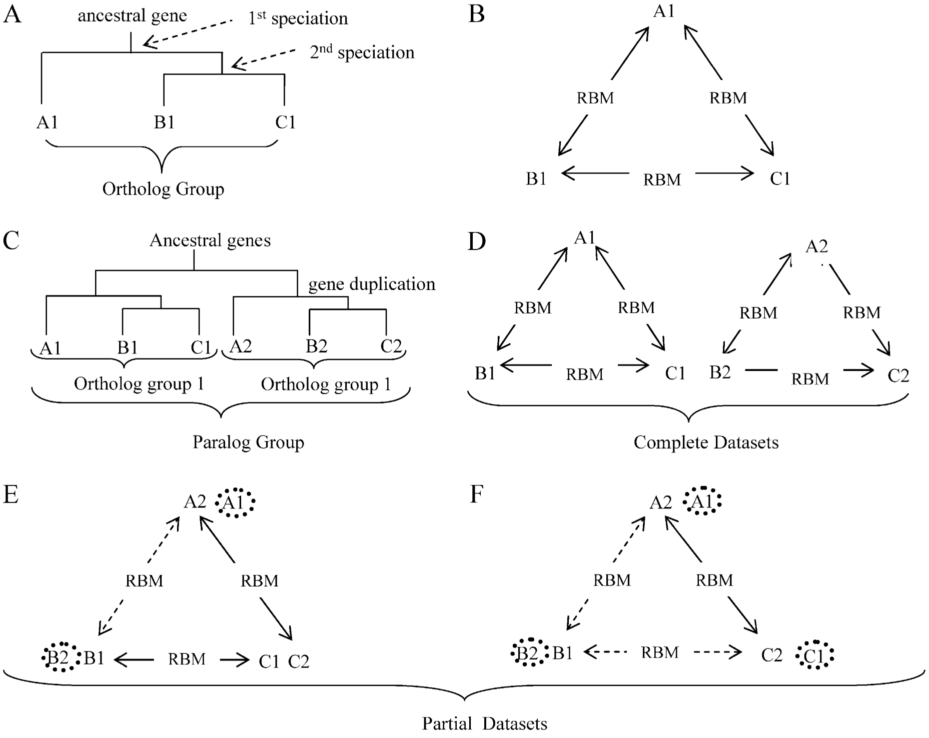

Figure 2.—(A) Evolution of an ancestral single-copy gene into three single copy orthologs (A1, B1, C1), one in each of the three related species. (B) RBM relationships of single-copy orthologs (A1, B1, C1) from pairwise comparisons of three fully sequenced genomes. (C) Evolution of paralogs, created by ancestral gene duplication, in the genomes of three related species. (D) Application of RBM triangulation method to distinguish orthologs from paralogs from pairwise comparisons of three fully sequenced genomes. (E and F) Application of RBM triangulation in genomes of three related species that have incomplete sequence data sets. Dotted circles indicate paralogous genes missing from the data set. Dashed arrowed lines connect two paralogs that form an erroneous RBM pair.

gene yielded 6415 putative ortholog groups. Each group contained from one to six RBM triangles, every one of which should include an Arabidopsis gene. In total, these corresponded to 6415 Arabidopsis genes, 5512 tomato unigenes, 5156 potato unigenes, 2997 pepper unigenes, and 3282 coffee unigenes (Tables S1 and S2 at http://www.genetics.org/supplemental/).

Selecting a subset of single-copy orthologous genes (COSII): Considering that a gene duplication event may occur independently in only some, but not all, lineages subsequent to speciation, and that genes in pepper and coffee unigene sets more likely appear to be "single copy" due to small data sets, a putative ortholog group was regarded as a single-copy ortholog group only if it met the following criteria: (1) the Arabidopsis gene member had to be single copy; (2) at least two of the included euasterid I gene members had to be single copy; and (3) if tomato and/or potato and pepper and/or coffee were included in the group, tomato or potato or both had to be single copy. This last criterion was included as an attempt to minimize the problem of the smaller pepper and coffee data sets. As a result, 2869 (45%) of the 6415 original putative ortholog groups were classified as single copy. Included in these groups are 2869 Arabidopsis genes (10% of the gene repertoire of this species), 2527 tomato unigenes, 2398 potato unigenes, 1368 pepper unigenes, and 1506 coffee unigenes (Table S2 at http://www.genetics.org/supplemental/). Hereafter, this subset is referred to as the COSII genes (or COSII groups), to differentiate them from the COS markers originally reported by Fulton *et al.* (2002). Of the 2869 COSII genes, 328 are identical to COS markers published by Fulton *et al.* (2002). For practical purposes, each COSII gene is referred to as the corresponding Arabidopsis gene locus (*e.g.*, C2_At1g44575).

Use of phylogenetic analyses to verify orthology for a subset of COSII genes: Phylogenetics can be used as an independent method for validating orthology (Dehal and Boore 2005; de la Torre *et al.* 2006). If a group of genes are truly orthologous, the gene tree and species trees should be in concordance (Figure 2A). Phylogenetic trees for the species included in this study have been previously reported (Chase *et al.* 1993; Olmstead 1999). Moreover, on the basis of concatenated COSII sequences we have also generated a high-confidence phylogeny for these same species (described later).

In an effort to determine what proportion of the COSII genes might erroneously contain one or more paralogs, gene trees for a subset of 401 COSII genes containing at least four members were reconstructed and 76% (304) of them were in concordance with the species tree (Figure S1 at http://www.genetics.org/supplemental/). Possible explanations for the incongruent cases include: (1) paralogs inadvertently included in a COSII group; (2) horizontal transfer (including introgressive hybridization); (3) gene duplication and extinction (Maddison 1997); (4) long-branch attraction and heterotachy (Felsenstein 1978; Philippe *et al.* 2005); and (5) insufficient sequence length sampled from orthologous genes failing to reconstruct an accurate gene tree. It was possible to test the impact of this fifth variable. As gene sequences are more complete, permitting greater sequence overlaps, the frequency of gene tree and species tree discrepancies should be reduced. The hypothesis was supported by a strong negative correlation ($r = -0.912$, P-value = 0.002) between average overlap of sequence length of COSII groups and the corresponding percentage of the incongruence between gene trees and species trees (Figure S1 at http://www.genetics.org/supplemental/). To be considered congruent, a COSII gene tree and species tree had to have exactly the same branch structure (Figure S1 at http://www.genetics.org/supplemental/). To further test this possibility, two COSII groups, with minimal sequence overlap in multiple alignments that produced gene trees incongruent with the species trees, were subjected to further sequencing to increase the overlap used for gene tree reconstruction. In both cases, gene trees reconstructed with the longer overlapping sequences turned into concordance with the species tree, indicating that both are valid ortholog sets (Figure S1 at http://www.genetics.org/supplemental/). These results therefore supported the notion that a significant portion of the species tree and the gene tree incongruence are due to a lack of full-length sequences (and thus to insufficient sequence overlaps) but not to a lack of orthology. On the basis of these cumulative data, we thus conclude that a majority of the 2869 COSII genes reported herein constitute valid ortholog sets.

Annotation of COSII genes: The fact that all COSII genes have a single homologous match in Arabidopsis indicates that this subset of genes has likely been under selection pressure to retain detectable homology (within euasterid species) and/or to remain (or become) single copy after the divergence of Solanaceae from Arabidopsis, which is estimated to have occurred 94–125 MYA (Gandolfo *et al.* 1998; Crepet *et al.* 2004). Each of the 2869 COSII genes was assigned a functional annotation on the basis of an matching Arabidopsis gene member. The 2869 COSII genes were assigned 38 role categories (Berardini *et al.* 2004) (Figure 3). Z-test was used to compare the relative proportion of COSII genes in each role category with that of the entire Arabidopsis gene repertoire. The results of these analyses indicate that genes encoding proteins targeted to mitochondria and plastids/chloroplasts occur in a significantly higher frequency in COSII genes ($P < 0.001$, Figure 3). Nuclear-encoded proteins targeted to organelles either physically interact with organelle-encoded proteins or play roles in common metabolic pathways. It is well known that chloroplast- and mitochondrion-encoded proteins evolve at a slower rate at nonsynonymous sites than most

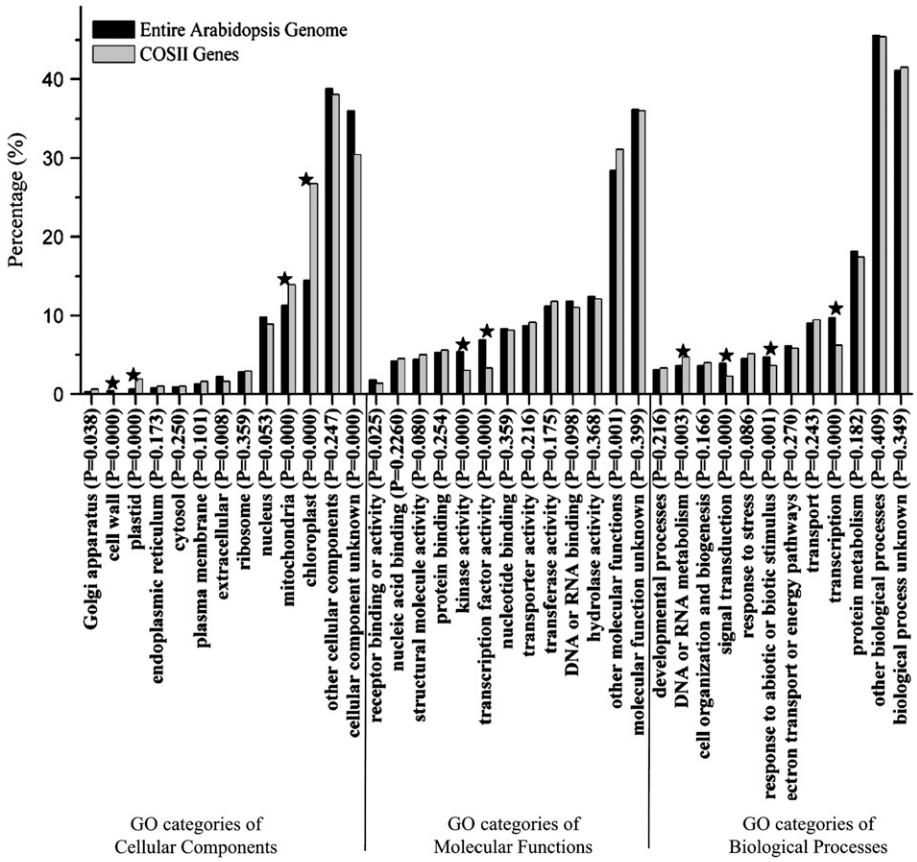

FIGURE 3.—Comparisons of role categorization between the entire Arabidopsis genome and COSII genes based on gene ontology (GO) annotation of the Arabidopsis genes. Stars represent categories showing significant differences (individual $P <$ 0.001 for an overall significance level of 0.05) between COSII genes and the entire Arabidopsis repertoire.

nuclear-encoded genes (WOLFE 1987). Hence, proteins encoded in the nucleus, but functioning in an organelle, may evolve at a rate comparable with their organelle-encoded counterparts or be under selection to maintain proper balance of organellar and nuclear gene counterparts. Also overrepresented in COSII genes are those encoding proteins involved in DNA or RNA metabolism (Figure 3).

Underrepresented in the COSII genes are those encoding proteins localized in cell walls, proteins kinases, transcription factors and those involved in signal transduction ($P < 0.001$, Figure 3). These categories of proteins may thus be evolving at a higher rate or be prone to gene family expansions/retentions—attributes that would disqualify them as COSII genes and hence account for their paucity in this subset of genes (MAERE *et al.* 2005). It is worth noting that previous studies have shown that gene duplication/retention plays a key role in the evolution of transcription factors in Arabidopsis where transcription factor gene families have much higher copy numbers than in other organisms (RIECHMANN and RATCLIFFE 2000).

Alignments and design of euasterid I universal primers for COSII genes: The COSII genes reported herein represent the largest set of putatively orthologous, single-copy genes in eudicot plant species. This unique resource has the potential to significantly enable a number of scientific endeavors, including: (1) phylogenetic studies in plant taxonomy/evolution and (2) use as orthologous markers for development of synteny maps for euasterid I and other related eudicot species. These topics will be discussed further in the following sections. However, for these applications to become a reality, it is necessary not only that the majority of COSII be *bona fide* orthologs, but also that they be easily assayed from many plant species—most of which have no sequence databases. Thus, a utilitarian goal of this project was to attempt to design a set of prototypical "universal" consensus PCR primers for each COSII gene with which one can amplify the corresponding ortholog from related plant species.

Design of COSII universal primers: A visual examination of the Arabidopsis: euasterid I COSII alignments quickly revealed that DNA sequence divergence between Arabidopsis and these other species is, in most instances, too great to allow design of consensus primers. However, examination of the tomato, potato, pepper, and coffee sequences revealed that, in many cases, universal primers could be developed to amplify, via PCR, the corresponding ortholog from each of the member species and therefore would be more likely to amplify the same ortholog from euasterid I species for which no sequence data sets exist (Figures 4 and 5). We refer to such primer sets as universal primers for euasterid I (UPA) to denote that they were designed to maximize the chances of amplification in species of

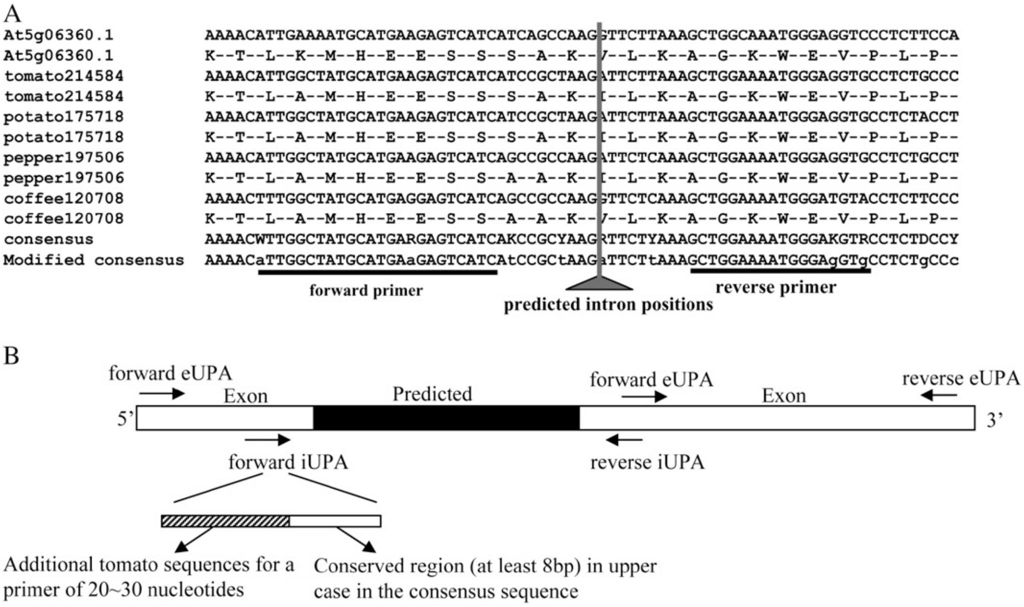
A
At5g06360.1 AAAACATTGAAAATGCATGAAGAGTCATCATCAGCCAAGGTTCTTAAAGCTGGCAAATGGGAGGTCCCTCTTCCA
At5g06360.1 K--T--L--K--M--H--E--E--S--S--S--A--K--V--L--K--A--G--K--W--E--V--P--L--P--
tomato214584 AAAACATTGGCTATGCATGAAGAGTCATCATCCGCTAAGATTCTTAAAGCTGGAAAATGGGAGGTGCCTCTGCCC
tomato214584 K--T--L--A--M--H--E--E--S--S--S--A--K--I--L--K--A--G--K--W--E--V--P--L--P--
potato175718 AAAACATTGGCTATGCATGAAGAGTCATCATCCGCTAAGATTCTTAAAGCTGGAAAATGGGAGGTGCCTCTACCT
potato175718 K--T--L--A--M--H--E--E--S--S--S--A--K--I--L--K--A--G--K--W--E--V--P--L--P--
pepper197506 AAAACATTGGCTATGCATGAAGAGTCATCAGCCGCCAAGATTCTCAAAGCTGGAAAATGGGAGGTGCCTCTGCCT
pepper197506 K--T--L--A--M--H--E--E--S--S--A--A--K--I--L--K--A--G--K--W--E--V--P--L--P--
coffee120708 AAAACTTTGGCTATGCATGAGGAGTCATCAGCCGCCAAGGTTCTCAAAGCTGGAAAATGGGATGTACCTCTTCCC
coffee120708 K--T--L--A--M--H--E--E--S--S--A--A--K--V--L--K--A--G--K--W--E--V--P--L--P--
consensus AAAACWTTGGCTATGCATGARGAGTCATCAKCCGCYAAGRTTCTYAAAGCTGGAAAATGGGAKGTRCCTCTDCCY
Modified consensus AAAACaTTGGCTATGCATGAaGAGTCATCAtCCGCtAAGaTTCTtAAAGCTGGAAAATGGGAgGTgCCTCTgCCc
forward primer
predicted intron positions
reverse primer
B
forward eUPA
Exon
Predicted
forward eUPA
Exon
reverse eUPA
5'
3'
forward iUPA
reverse iUPA
Additional tomato sequences for a primer of 20~30 nucleotides
Conserved region (at least 8bp) in upper case in the consensus sequence

FIGURE 4.—Design of universal primers for euasterid I species (UPA) in a COSII group. (A) Multiple alignment of euasterid I species and the corresponding Arabidopsis ortholog. Intron positions of euasterid I species were predicted on the basis of that of the Arabidopsis ortholog. UPAs were designed in conserved portions of exons. (B) iUPAs amplify mostly intronic sequences including <400 bp of the flanking exons, while eUPAs amplify at least 400-bp exonic sequences with or without the intervening intron(s).

this eudicot clade (Figure 1). Previous research has shown that most homologous tomato and Arabidopsis genes share common intron positions (KU *et al.* 2000). Hence, intron positions of the euasterid I members for the COSII genes with alignments were predicted on the basis of the Arabidopsis gene model. The reason for identifying the intron positions was that UPAs could be designed to amplify either intronic regions

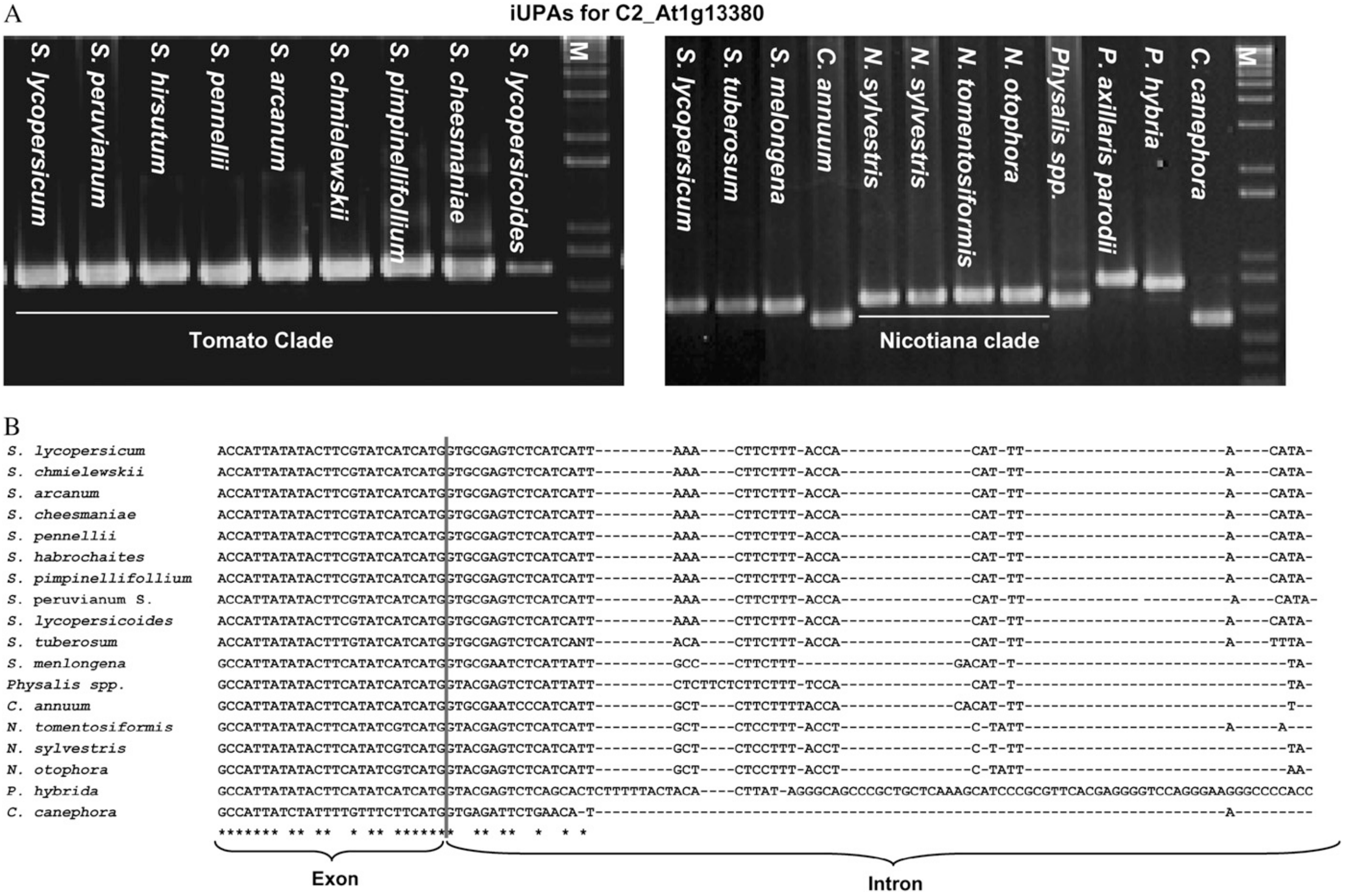
A
iUPAs for C2_At1g13380
S. lycopersicum
S. peruvianum
S. hirsutum
S. pennellii
S. arcanum
S. chmielewskii
S. pimpinellifollium
S. cheesmaniae
S. lycopersicoides
M
Tomato Clade
S. lycopersicum
S. tuberosum
S. melongena
C. annuum
N. sylvestris
N. sylvestris
N. tomentosiformis
N. otophora
Physalis spp.
P. axillaris parodii
P. hybria
C. canephora
M
Nicotiana clade
B
S. lycopersicum ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. chmielewskii ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. arcanum ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. cheesmaniae ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. pennellii ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. habrochaites ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. pimpinellifollium ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. peruvianum S. ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT------------ ----------A----CATA-
S. lycopersicoides ACCATTATATACTTCGTATCATCATGGTGCGAGTCTCATCATT---------AAA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----CATA-
S. tuberosum ACCATTATATACTTTGTATCATCATGGTGCGAGTCTCATCANT---------ACA----CTTCTTT-ACCA---------------CAT-TT-----------------------A----TTTA-
S. menlongena GCCATTATATACTTCATATCATCATGGTGCGAATCTCATTATT---------GCC----CTTCTTT-----------------GACAT-T------------------------------TA-
Physalis spp. GCCATTATATACTTCATATCATCATGGTACGAGTCTCATTATT---------CTCTTCTCTTCTTT-TCCA---------------CAT-T------------------------------TA-
C. annuum GCCATTATATACTTCATATCATCATGGTGCGAATCCCATCATT---------GCT----CTTCTTTTACCA-------------CACAT-TT-----------------------------T--
N. tomentosiformis GCCATTATATACTTCATATCGTCATGGTACGAGTCTCATCATT---------GCT----CTCCTTT-ACCT---------------C-TATT-----------------------A-----A---
N. sylvestris GCCATTATATACTTCATATCGTCATGGTACGAGTCTCATCATT---------GCT----CTCCTTT-ACCT---------------C-T-TT-----------------------------TA-
N. otophora GCCATTATATACTTCATATCGTCATGGTACGAGTCTCATCATT---------GCT----CTCCTTT-ACCT---------------C-TATT-----------------------------AA-
P. hybrida GCCATTATATACTTCATATCATCATGGTACGAGTCTCAGCACTCTTTTACTACA----CTTAT-AGGGCAGCCCGCTGCTCAAAGCATCCCGCGTTCACGAGGGGTCCAGGGAAGGGCCCCACC
C. canephora GCCATTATCTATTTTGTTTCTTCATGGTGAGATTCTGAACA-T---A---------
******* ** ** * ** ******* ** ** * * *
Exon
Intron

FIGURE 5.—Use of universal primers (UPAs) to amplify orthologous counterparts from genomic DNA of different solanaceous species and coffee. (A) Amplification by iUPAs for C2_At1g13380. (B) Part of the sequence alignment of amplified sequences by iUPAs for C2_At1g13380. Asterisks indicate identical sites in the multiple sequence alignment.

(iUPA primers) or exonic regions (eUPA primers) for each COSII gene. Both types of primers were designed on the conserved exonic region; however, eUPA primers were designed to amplify >400-bp exonic regions with or without intervening intron(s), while iUPA primers were designed to amplify <400-bp exonic regions to increase the portion of intron in the amplicons (Figure 4).

Testing COSII UPAs in euasterid I species: A set of 548 COSII iUPAs was tested in three tomato species: *S. lycopersicum*, *S. pimpinellifolium*, and *S. pennellii.* The percentage of single-band amplification rate was quite similar among the three species, ranging from 89% in *S. pennellii* to 92% in *S. lycopersicum* (Table S3 at http://www.genetics.org/supplemental/). Failure in PCR amplification may be due to nonspecific primers with high mismatches and/or difficulty in amplifying large introns of tomato. The first possibility is not likely since, of the 29 failures, 28 contained a tomato sequence in the multiple alignment used to design the universal primers. With regard to the second possibility, a significant positive correlation was observed between tomato intron length and the corresponding Arabidopsis intron ($r = 0.321$, $P < 0.001$). Moreover, a paired T-test showed that tomato introns are significantly longer (2× average) than their Arabidopsis counterparts ($P < 0.001$). Of the 29 failed cases, the average Arabidopsis intron length was nearly twice as long as that of the 548 cases in total. These results are all consistent with the notion that most of the 29 failed amplifications were attributed to exceptionally long tomato intron lengths.

A smaller subset of ∼100 iUPAs of the above 548 cases was further tested for PCR amplification on genomic DNA from a cross section of solanaceous (potato, eggplant, pepper, Physalis, Nicotiana, and Petunia) and rubiaceous (coffee) species. To reduce possibility of amplifying multiple bands due to allelic polymorphism and/or polyploidy, most species used are inbred diploid lines or dihaploid lines (*S. tuberosum*) except for *Physalis* spp. and *Coffea canephora* var. robusta. Single-band amplification rates ranged from 89% in *S. lycopersicum* to 40% in *Physalis* spp., and 66 (67%) of the iUPAs amplified a single band from at least four tested species (an example is shown in Figure 5). These results not only confirmed the use of UPAs but also suggested that the intron positions have remained conserved among these species as well as in the distant Arabidopsis, a question further addressed by sequencing the amplicons (described later). The observance of multiple bands in some species but not in others is consistent with the assumption made in selection of COSII genes that gene duplication events subsequent to speciation may occur in only one or a few but not all species, although another possibility is that the lines used for amplification are heterozygous at these loci and that alleles varied in intron size (due to indels).

Use of COSII genes in species outside the euasterid I clade: While the COSII genes and universal primers reported herein have focused on species in the euasterid I clade, they may still have use in more divergent taxa. The fact that each COSII group was required to have a single-copy Arabidopsis member presents the possibility that these COSII genes may be conserved in other eudicot taxa (Figure 1). As described earlier, the sequence divergence between Solanaceae/Rubiaceae and Arabidopsis COSII genes is too great to allow designing universal primers that can amplify the COSII orthologs from any plant species. However, a similar ortholog set can be built up for the other eudicot clade on the basis of the comparison against these current COSII genes, and universal primers can be designed using the same strategy as well.

To investigate the utility of such COSII sequence searching outside of the euasterid I clade, two other plant data sets—EST-derived unigene data sets for lettuce and sunflower (Table 1), members of the Asteraceae family in the euasterid II clade—were blasted against the five data sets used for screening COSII genes, *i.e.*, tomato, potato, pepper, coffee (by BLASTN), and Arabidopsis (by BLASTX). To qualify as a new member for a COSII group, a gene in the new query data set had to meet the following criteria: (1) the gene should form a RBM pair with an Arabidopsis COSII member and (2) the gene must be a single-copy gene using the same cut-off value (1E-10) as described in MATERIALS AND METHODS. As a result, 190 COSII groups with a sunflower member, 229 COSII groups with a lettuce member, and 26 groups with both members were identified. Gene trees, based on concatenated COSII sequences in lettuce and/or sunflower, Arabidopsis, and euasterid I species were generated. In all cases, the concatenated COSII gene trees matched the known species tree for all euasterid taxa tested, indicating the validity of the COSII members identified in lettuce and sunflower (Figure 1B; Figure S2 at http://www.genetics.org/supplemental/).

Distribution of COSII genes in Arabidopsis and tomato genomes: As stated earlier, one of the goals of identifying COSII genes was to provide a set of conserved single-copy orthologous genes for comparative mapping across wide taxonomic distances—a necessity for understanding both genome and chromosome evolution and for applying genome sequence data from fully sequenced genomes to the species that lack such information. In this regard, one of the first applications of COSII genes may be to develop common synteny maps for many euasterid species, such that all can benefit from sequencing of the tomato genome, an international project that is now underway (http://www.sgn.cornell.edu/help/about/tomato_sequencing.html). Furthermore, these comparative maps across wide phylogenetic distances should further inform scientists about the nature of genome evolution in this clade of plants.

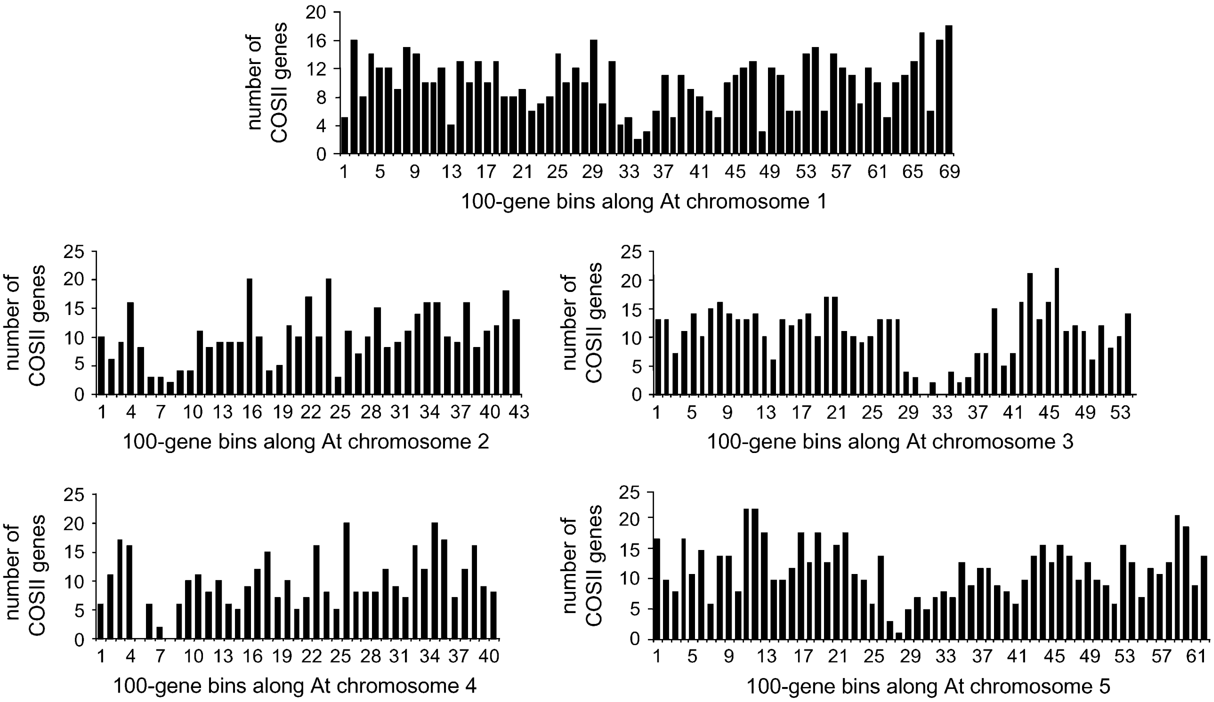

Figure 6.—Distribution of COSII genes (Arabidopsis orthologs) across the Arabidopsis genome. Each Arabidopsis chromosome was divided into sequential bins of 100 genes each (including all the predicted nuclear genes in TAIR, Table 1). The number of observed COSII genes per bin was then plotted for all five Arabidopsis chromosomes.

Moreover, since each COSII gene is anchored to a single Arabidopsis gene, the possibility exists that these COSII markers can help unite the genetics/genomics of an even wider taxonomic group of plant species.

To evaluate and facilitate the potential of the COSII genes to meet the objectives outlined above, three experiments were performed: (1) the identification of the positions of all COSII genes in the Arabidopsis genome to determine how well the entire genome is covered; (2) the mapping of a subset of the COSII genes mapped in tomato to further determine genome coverage; and (3) the mapping of a subset of the COSII mapped on tomato in coffee to determine whether COSII markers could be used to detect synteny in such interfamilial comparisons (*e.g.*, Solanaceae *vs.* Rubiaceae). The results of the above experiments are described below.

Distribution of COSII genes in Arabidopsis genome: Each Arabidopsis chromosome was divided into sequential bins of 100 genes each [including all the predicted nuclear-encoded genes in the Arabidopsis Information Resource (Table 1)]. The number of observed COSII genes per bin was then plotted for all five Arabidopsis chromosomes (Figure 6). At first glance, on a gene basis, all Arabidopsis chromosomal regions appear to be covered by COSII genes; in other words, COSII genes appear to be a random sample from the entire Arabidopsis gene repertoire. However, a chi-square goodness-of-fit test (for the number of COSII markers per bin) for each chromosome revealed significant deviations ($P < 0.05$) from binomial distribution for all chromosomes except chromosome 1 ($P = 0.09$). Thus, while at the gross level the COSII markers cover the entire Arabidopsis genome, there are specific regions that are either over- or underrepresented.

One possible explanation for this observation may be traced to the now widely accepted inference that the lineage leading up to modern-day Arabidopsis was likely punctuated by whole-genome duplication events followed by gradual loss or divergence of duplicated genes (Blanc *et al.* 2000; Grant *et al.* 2000; Ku *et al.* 2000; Bowers 2003). To qualify as a COSII, a gene must be single copy in Arabidopsis. Thus, regions of the Arabidopsis genome that are still duplicated may be deficient in the COSII gene set. If this is indeed the case, a negative correlation between the number of COSII genes per bin and the average copy number of the Arabidopsis genes should be observed (see materials and methods for defining gene copy number) within each bin. Pearson's correlation coefficients were thus calculated for each of the five Arabidopsis chromosomes and suggested strong negative correlations for all chromosomes ($r = -0.56$–-0.72; $P < 0.001$), supporting the notion that localized heterogeneity in loss of duplicated genes in Arabidopsis after polyploidization may account for the nonrandom distribution of COSII genes within Arabidopsis chromosomes. To further test this hypothesis, only the single-copy genes were subsequently divided into sequential bins of 100 genes each, and the number of COSII genes per 100 single-copy genes for all five chromosomes was also subjected to a chi-square

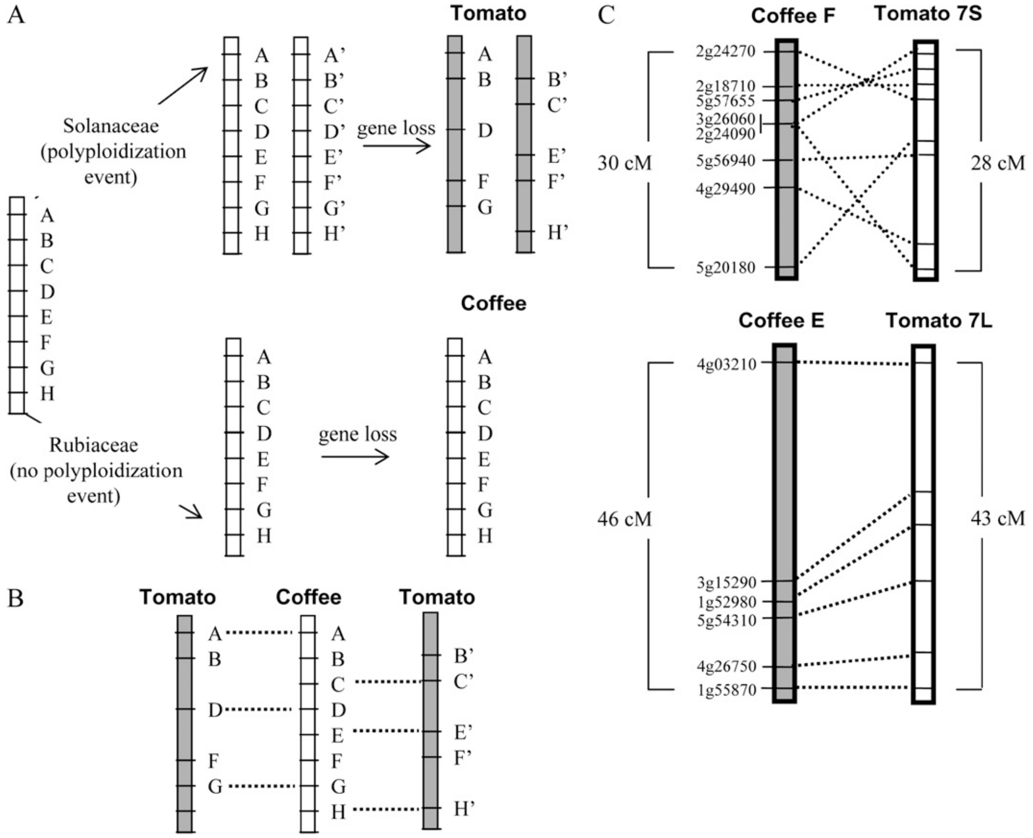

FIGURE 7.—(A) Predicted outcomes when mapping COSII orthologs between a species (Solanaceae, tomato) with a hypothetical whole-genome duplication (polyploidy) at the base of its lineage compared with a related taxon (Rubiaceae, coffee) that did not experience such an event. The lineage with a polyploidization event would have all chromosomes (and genes) duplicated, followed by selective gene loss. (B) When a whole-genome duplication event occurred only in the Solanaceae lineage, mapping of COSII genes between coffee and tomato led to a "network of synteny" as described in KU *et al.* (2000). Note that no such "network of synteny" has been observed in mapping tomato and coffee with a set of >150 COSII genes—casting doubt on the polyploidy event in the Solanaceae lineage proposed by BLANC and WOLFE (2004). (C) Comparative mapping in coffee of COSII markers located on the short arm of tomato chromosome 7 (top) and long arm of tomato chromosome 7 (bottom). Note the one-to-one relationship between the coffee–tomato syntenous regions, which would be predicted if no polyploidy event had occurred in either the Solanaceae or the Rubiaceae lineage either after or just prior to divergence from their last common ancestor.

goodness-of-fit test, which resulted in no significant deviation from binomial distributions ($P = 0.07$–0.90). In conclusion, COSII genes provide good coverage of the single-copy "islands" of the Arabidopsis genome, which are surrounded by extensively duplicated regions.

Distribution of COSII genes in tomato: So far, 525 COSII genes have been mapped onto the high-density tomato genetic map (Figure S3 at http://www.sgn.cornell.edu/markers/cosii_markers.pl). The tomato genome has not yet been sequenced, and thus it was not possible to conduct a heterogeneity test for uniformity of distribution across the tomato genome as was done for Arabidopsis. However, the mapped COSII genes fell on all 12 chromosomes, suggesting that, at least at the gross chromosomal level, the entire tomato genome is likely to be covered. However, we cannot rule out fine-scale distribution heterogeneity similar to what is observed in Arabidopsis.

Application of COSII markers for comparative mapping across wide phylogenetic distances—Solanaceae (tomato) *vs.* Rubiaceae (coffee): Most comparative mapping studies in plants have been restricted to species within the same plant family (PATERSON *et al.* 2000). Coffee (a member of the family Rubiaceae with chromosome number $x = 11$) and tomato (a member of the family Solanaceae with chromosome number $x = 12$) are estimated to have shared a common ancestor ~85 MYA (WIKSTROM *et al.* 2001). To test the efficacy of COSII markers for comparative mapping across such large phylogenetic distances, a subset of COSII markers is being mapped in both tomato and diploid coffee (*C. canephora*). The results indicate that each segment of the coffee genome corresponds to a single segment of the tomato genome (D. CROUZILLAT, unpublished data). For example, the long arm of tomato chromosome 7, encompassing 14 COSII markers and 43 cM, corresponds to a 46-cM segment in coffee linkage group E in which the gene order has been preserved (Figure 7C). Likewise, the short arm of tomato chromosome 7, comprised of 8 COSII makers and 28 cM, corresponds to a 30-cM segment of coffee chromosome F—although the two syntenous segments differ by at least two paracentric inversions (Figure 7C). Thus far we have observed no cases in which single coffee chromosomes (defined by COSII markers) show a networked synteny with two corresponding tomato chromosomal pieces or vice versa. Such would be the case if polyploidization had affected either tomato or coffee lineage (KU *et al.* 2000). These results demonstrate that the COSII gene sets can be used for comparative genome mapping among plant families. Further, they provide insights into genome evolution in euasterids.

Blanc and Wolfe (2004) analyzed existing plant sequence databases and on the basis of those data proposed the possibility of a whole-genome duplication event in the euasterids. Further, this duplication event was predicted to have occurred in the branch leading to the Solanaceae ~20 MYA. Several lines of evidence presented herein (including the coffee–tomato comparative mapping) suggest that the hypothesis of a whole-genome duplication event may not be correct:

1. The families Solanaceae (order Solanales) and Rubiaceae (order Gentianales) are estimated to have diverged ~85 MYA (Wikstrom *et al.* 2001). Thus the whole-genome duplication event predicted to have occurred 20 MYA by Blanc and Wolfe (2004) would have been specific to the Solanaceae lineage, occurring after the divergence between Solanaceae and Rubiaceae from their last common ancestor. If this were the case, many, if not most, coffee genes would correspond to two tomato genes (resulting from the whole-genome duplication in the Solanaceae lineage) (Figure 7, A and B). This prediction is not consistent with the results presented herein as we show that a significant portion of the coffee and tomato gene repertoires show one-to-one orthology matches (*vs.* a two-to-one as would be expected if a whole-genome duplication event affected Solanaceae, but not Rubiaceae) (Tables S1 and S2 at http://www.genetics.org/supplemental/). A similar conclusion was also reached in a study comparing the coffee and tomato EST databases (Lin *et al.* 2005).
2. If a whole-genome duplication event had occurred in the Solanaceae lineage after divergence from Rubiaceae, one would not expect to find stretches of uninterrupted synteny (as reported herein), but rather a network of complex synteny to emerge due to whole-genome duplication followed by selective gene loss (Ku *et al.* 2000; Li *et al.* 2003). Moreover, a similar complex network of synteny would also be observed if a whole-genome duplication event had occurred just prior to divergence of Rubiaceae and Solanaceae followed by selective gene loss in both duplicated lineages (Figure 7, A and B).
3. Phylogenetic analyses in this study indicate that the basal chromosome number of Solanaceae was $x = 12$ and of Rubiaceae was $x = 11$. Thus the last common ancestor of Solanaceae and Rubiaceae probably had a chromosome number of $x =1\ 1$ or 12. This would not be likely if a polyploidization event occurred in the Solanaceae lineage after divergence with the Rubiaceae, in which case the basal chromosome number of Solanaceae would be 22 or 24—a prediction that is not consistent with the empirical data.

On the basis of these lines of evidence we propose that the polyploidization event postulated by Blanc and Wolfe (2004) either did not occur (and their results stemmed from analyzing gene duplications resulting from localized gene duplication, *e.g.*, tandem gene duplication and segmental duplication, rather than a whole-genome duplication event) or the polyploidization event that they postulate occurred much earlier in the evolution of the euasterid I species—well before the divergence of Solanaceae and Rubiaceae. If the polyploidization event did occur, Hughes *et al.* (2003) have proposed that transposable elements in Arabidopsis are responsible, at least in part, for some of the segmental duplication attributed to polyploidization. Further comparative mapping across euasterid I species, using the COSII markers, should provide a means for clarifying this issue.

Use of COSII universal primers for phylogenetic studies: In the past, there has been a paucity of validated nuclear orthologs for phylogenetic studies, and hence most molecular taxonomy studies have relied heavily on a handful of chloroplast and/or ribosomal genes (Small *et al.* 2004). Phylogenies reconstructed with only one or a few independently inherited loci may result in unresolved or incongruent phylogenies due to data sampling (Graybeal 1998), horizontal gene transfer, or differential selection and lineage sorting at individual loci, etc. (Maddison 1997). The magnitude of the risk of basing phylogenies on one or a few loci is well demonstrated by the study of Rokas *et al.* (2003) where it was empirically determined that the sequences from a relatively larger number (>20 in this case) of independently inherited orthologous genes may need to be examined and combined to generate robust phylogenetic trees. The large set of COSII genes and associated universal primers reported herein provides the first broad pool of multiple unlinked, single-copy, orthologous nuclear genes for use in plant phylogenetics, especially in taxa belonging to the euasterid clade.

COSII genes and the associated universal primers described herein were tested for applications in phylogenetic studies involving both more closely (*e.g.*, within genera) and more distantly related taxa (*e.g.*, across genera or families). With regards to more distant taxa, exon regions amplified by 10 eUPAs (one for a COSII gene) were sequenced for eight diverse euasterid I species (tomato, potato, eggplant, pepper, Physalis, tobacco, Petunia, and coffee). For studies among more closely related taxa, intronic regions amplified by five iUPAs and exonic regions by 6 eUPAs as well as one of the above iUPAs were sequenced from eight tomato-related species as well as from the outgroup *S. lycopersicoides* (Figure 8, B and C). COSII genes and primers used for these phylogenetic studies and the GenBank accessions of the sequenced amplicons are listed, respectively, in Tables S4 and S5 at http://www.genetics.org/supplemental/.

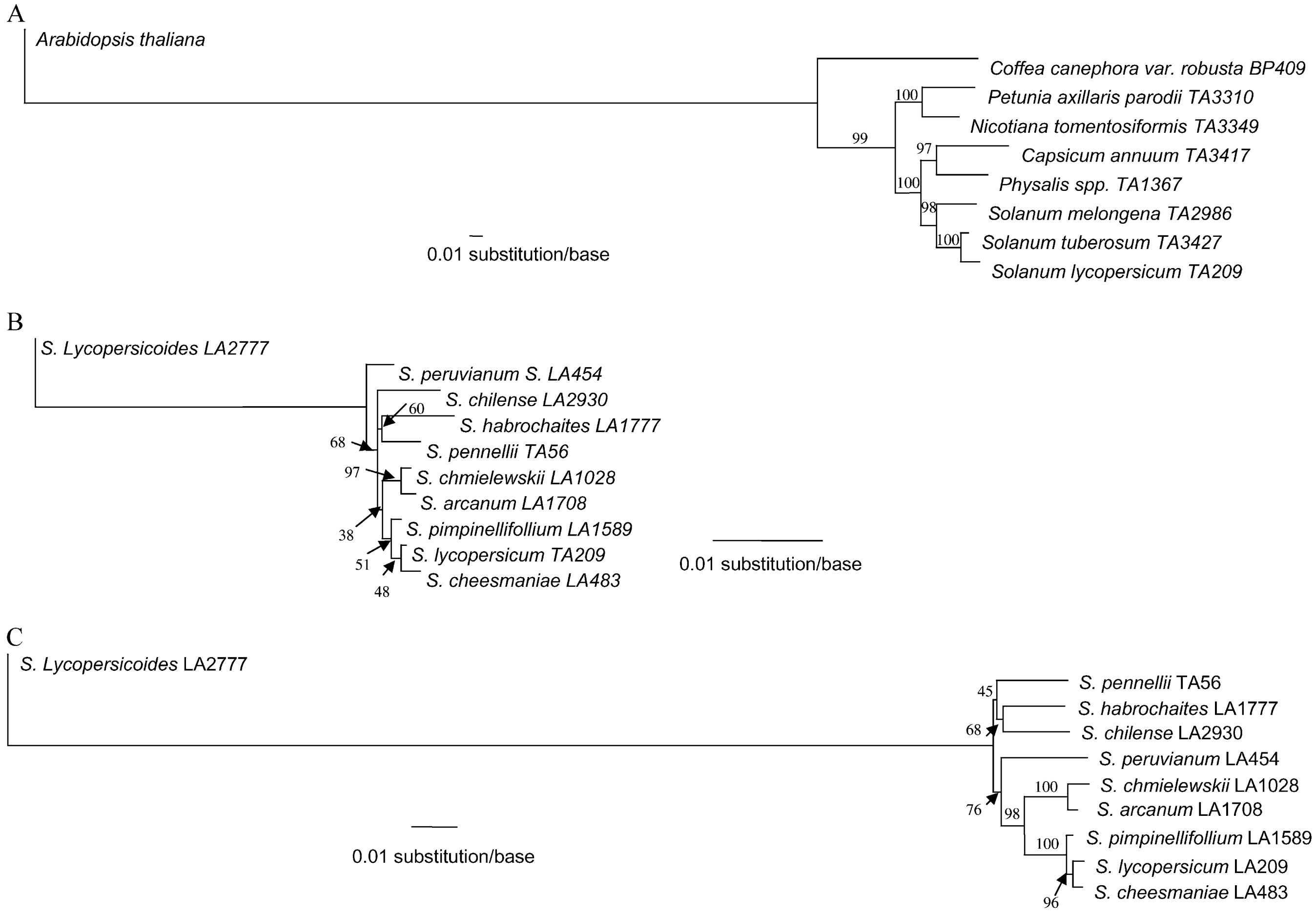

FIGURE 8.—Use of COSII universal primers (UPAs) for phylogenetics. (A) Phylogenetic relationships among euasterid I species from different genera and families using eUPAs that amplify orthologous exons. The tree is based on concatenated exonic sequences corresponding to 10 COSII genes totaling 3750 bp. (B) Phylogenetic relationships among the species most closely related to the cultivated tomato based on concatenated exonic sequences of 2316 bp from 7 COSII genes. (C) Phylogenetic relationships for the same species, but based on concatenated intronic sequences of 2403 bp from 5 COSII genes.

Use of COSII for phylogenetic studies across more distantly related taxa: At the intergeneric to interfamilial level, intronic regions (derived from iUPA amplification) were highly variable in both length and sequence (Figure 5). As a result, intronic sequences were alignable only among closely related species (*e.g.*, the nine Solanum species most closely related to *S. lycopersicum*). Unlike introns, exonic regions were readily aligned among all taxa, even those most distantly related (*e.g.*, coffee *vs.* solanaceous species) (Figure 5B). The 10 eUPAs mentioned earlier were used to amplify and sequence orthologous regions from these taxa, and the concatenated sequences (3750 bp in total) were used to generate a phylogenetic tree (Figure 8A). The result was in general agreement with the chloroplast phylogenies—with, however, greater resolving power on the basis of bootstrap values (Figure 1B, Figure 8A) (CHASE *et al.* 1993; OLMSTEAD 1999). One striking difference is that the COSII phylogeny places tobacco and Petunia as sister taxa *vs.* Petunia being basal to the Solanaceae (bootstrap value = 100%). This result conflicts with studies based on chloroplast DNA, internal transcribed spacer sequences, and a single nuclear gene, SAMT, in which Nicotiana was closer to Solanum, Capsicum, and Physalis, while Petunia was an outgroup to the above taxa, although the bootstrap values for these studies were generally lower than those obtained with the COSII sequences (OLMSTEAD and SWEERE 1994; OLMSTEAD 1999; SANTIAGO-VALENTIN and OLMSTEAD 2003; MARTINS and BARKMAN 2005). The high bootstrap value generated with this study, combined with the fact that 10 unlinked (distributed on different chromosomes) orthologous loci were examined, brings these prior studies into question. Petunia is one of the few solanaceous taxa with a basic chromosome number <12. The prior placement of Petunia as outgroup would be consistent with speculation that the basal chromosome number of the family Solanaceae was $x = 6$ and that all $x = 12$ species were derived from an early polyploidization event (GOODSPEED 1954). The phylogeny generated by this study, in which Petunia is no longer an outgroup, throws significant doubt on that hypothesis and is more consistent with the basal chromosome number for Solanaceae being $x = 12$, which is remarkably similar to the $x = 11$ chromosome number common to most Rubiaceae taxa, including coffee. These results suggest that the last common ancestor of Rubiaceae and Solanaceae may have had a chromosome number of either $x = 11$ or $x = 12$.

Use of COSII for phylogenetic studies across more closely related taxa: A comparison of the sequence divergence in introns *vs.* exons for the closely related tomato species indicates that the base substitution rate in introns is on average 2.7-fold greater ($P < 0.001$) than in exons, presumably due to relaxed selection in the intronic region (Hughes and Yeager 1997; Small and Wendel 2000). Both exonic (2316 bp concatenated from seven COSII genes) and intronic sequences (2403 bp concatenated from five COSII genes) were used to study the relationships among the closely related species in the *S. lycopersicum* (tomato) clade (Figure 8, B and C). The resulting phylogenetic trees derived from both exonic and intronic sequences give similar tree topologies, which in general agree with those derived from other molecular data (Spooner *et al.* 2005). However, the branch lengths were consistently longer and the bootstrap values consistently higher for intronic data *vs.* exonic data (Figure 8, B and C). Evidently, the accelerated rate of intron *vs.* exon divergence provides more resolving power when comparing more closely related species.

The identification of orthologous gene sets across species forms the foundation for much of today's comparative biology and molecular systematics (Eisen 1998; Rokas *et al.* 2003; Blanchette *et al.* 2004; Fay 2006). Methods are now well established for identifying putative orthologs among species whose genomes have been fully sequenced (Tatusov *et al.* 2000; Remm *et al.* 2001; Lee *et al.* 2002; Li *et al.* 2003). However, for most species, either no sequence databases exist or, if they do exist, they are incomplete (*e.g.*, EST databases). We herein report the application of computational and phylogenetic tools to identify a large number of putative orthologs using multiple, incomplete EST sequence databases. In total, we identified 2869 conserved putative ortholog sets (COSII), which are common to most, if not all, euasterid plant species (which encompass one-quarter of all flowering plants) and Arabidopsis. This represents the largest set of single- or low-copy orthologous genes identified thus far for any set of plant species.

Functional annotation revealed that the 2869 putative orthologs encode a higher-than-expected frequency of proteins transported and utilized in organelles and a paucity of proteins associated with cell walls, protein kinases, transcription factors, and signal transduction. The utility of this large ortholog set (and the associated universal primers) was demonstrated in phylogenetic studies across very wide taxonomic intervals (*e.g.*, intergeneric or interfamilial) as well as in closely related species (*e.g.*, congenerics). Combined results of phylogenetic analyses and comparative mapping across plant families such as Solanaceae (tomato) and Rubiaceae (coffee) provide compelling evidence that the ancestral species that gave rise to the euasterid families Solanaceae and Rubiaceae had a basic chromosome number of $x = 11$ or 12 and that taxa with lesser chromosome numbers (*e.g.*, petunia) represent a derived situation. It is apparent that no whole-genome duplication occurred immediately prior to or after the radiation of either the family Solanaceae or the family Rubiaceae as previously proposed by Blanc and Wolfe (2004). Further comparative mapping and sequencing of COSII genes across the wide diversity of the euasterids should allow a much more in-depth understanding of the origins, diversifications, and dispersal/adaptation of this important clade of plants. Finally, the algorithms used to identify and validate these ortholog sets may also prove useful for similar applications in other taxonomic clades where partial sequence data sets are available.

We thank Yimin Xu, Nancy T. Eannetta, Eloisa C. Tedeschi, and Xiaomin Jia for mapping COSII markers; John C. Binns for constructing and maintaining the COSII database and web page; Martin Ganal for testing universal primers; and Anne Frary for critical reading and editing. Seeds and/or DNA for testing primers were received from Walter De Jong (Cornell University; potato tubers), Molly Jahn (Cornell University; pepper), Pal Maliga (Rutgers University), Andrew Leitch (University of London, UK; Nicotiana species), and Cris Kuhlemeier (University of Bern, Switzerland; Petunia species). This project was supported by the National Science Foundation Plant Genome Research Program (no. 0116076) and by a grant from the Nestlé Corporation.

LITERATURE CITED

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796–815.

Berardini, T. Z., S. Mundodi, L. Reiser, E. Huala, M. Garcia-Hernandez *et al.*, 2004 Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. **135:** 745–755.

Blanc, G., and K. H. Wolfe, 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093–1101.

Blanchette, M., E. D. Green, W. Miller and D. Haussler, 2004 Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. **14:** 2412–2423.

Bowers, J. E., B. A. Chapman, J. Rong and A. H. Paterson, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **408:** 796–815.

Bremer, B., K. Bremer, M. W. Chase, J. L. Reveal, D. E. Soltis *et al.*, 2003 An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. **141:** 399–436.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les *et al.*, 1993 Phylogenetics of seed plants: an analysis of nucleotide-sequences from the plastid gene Rbcl. Ann. Mo. Bot. Gard. **80:** 528–580.

Crepet, W. L., K. C. Nixon and M. A. Gandolfo, 2004 Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from cretaceous deposits. Am. J. Bot. **91:** 1666–1682.

de la Torre, J. E., M. G. Egan, M. S. Katari, E. D. Brenner, D. W. Stevenson *et al.*, 2006 Estimating plant phylogeny: lessons from partitioning. BMC Evol. Biol. **6:** 48.

Dehal, P., and J. L. Boore, 2005 Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. **3:** e314.

Doganlar, S., A. Frary, M. C. Daunay, R. N. Lester and S. D. Tanksley, 2002 Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics **161:** 1713–1726.

Eisen, J. A., 1998 Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. **8:** 163–167.

Fay, J., 2006 Human genome: Which proteins contribute to human-chimpanzee differences? Eur. J. Hum. Genet. **14:** 506.

Felsenstein, J., 1978 Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27:** 401–410.

Frary, A., Y. M. Xu, J. P. Liu, S. Mitchell, E. Tedeschi *et al.*, 2005 Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor. Appl. Genet. **111:** 291–312.

Fulton, T. M., R. Van der Hoeven, N. T. Eannetta and S. D. Tanksley, 2002 Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell **14:** 1457–1467.

Gandolfo, M. A., K. C. Nixon and W. L. Crepet, 1998 A new fossil flower from the Turonian of New Jersey: Dressiantha bicarpellata gen. et sp. nov. (Capparales). Am. J. Bot. **85:** 964–974.

Gogarten, J. P., and L. Olendzenski, 1999 Orthologs, paralogs and genome comparisons. Curr. Opin. Genet. Dev. **9:** 630–636.

Goodspeed, T. H., 1954 *The Genus Nicotiana: Originism Relationships and Evolution of Its Species in the Light of Their Distribution, Morphology and Cytogenetics*, pp. 283–310. The Chronica Botanica Company, Waltham, MA.

Grant, D., P. Cregan and R. C. Shoemaker, 2000 Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. USA **97:** 4168–4173.

Graybeal, A., 1998 Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. **47:** 9–17.

Hughes, A. L., and M. Yeager, 1997 Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45:** 125–130.

Hughes, A. L., R. Friedman, V. Ekollu and J. R. Rose, 2003 Nonrandom association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. Mol. Phylogenet. Evol. **29:** 410–416.

Kalendar, R., 2001 A computer program "Oligos" for PCR primers design, in *The Third Major International Bioinformatics Meeting in Scandinavia: "Bioinformatics 2001."* Skovde, Sweden.

Ku, H. M., T. Vision, J. Liu and S. D. Tanksley, 2000 Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA **97:** 9121–9126.

Lee, Y., R. Sultana, G. Pertea, J. Cho, S. Karamycheva *et al.*, 2002 Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res. **12:** 493–502.

Li, L., C. J. Stoeckert and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178–2189.

Lin, C., L. A. Mueller, J. McCarthy, D. Crouzillat, V. Petiard *et al.*, 2005 Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theor. Appl. Genet. **112:** 114–130.

Maddison, W. P., 1997 Gene trees in species trees. Syst. Biol. **46:** 523–536.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu *et al.*, 2005 Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA **102:** 5454–5459.

Martins, T. R., and T. J. Barkman, 2005 Reconstruction of Solanaceae phylogeny using the nuclear gene SAMT. Syst. Bot. **30:** 435–447.

Notredame, C., D. G. Higgins and J. Heringa, 2000 T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302:** 205–217.

O'Brien, K. P., M. Remm and E. L. Sonnhammer, 2005 Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. **33:** D476–D480.

Olmstead, R. G., and J. A. Sweere, 1994 Combining data in phylogenetic systematics: an empirical approach using 3 molecular data sets in the Solanaceae. Syst. Biol. **43:** 467–481.

Olmstead, R. G., J. A. Sweere and R. E. Spangler, 1999 Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA, pp. 111–137 in *Solanaceae IV*, edited by M. Nee, D. E. Symon, R. N. Lester and J. P. Jessop. Royal Botanic Gardens, Kew, United Kingdom.

Paterson, A. H., J. E. Bowers, M. D. Burow, X. Draye, C. G. Elsik *et al.*, 2000 Comparative genomics of plant chromosomes. Plant Cell **12:** 1523–1539.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue and F. Delsuc, 2005 Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. **5:** 50.

Posada, D., and K. A. Crandall, 1998 MODELTEST: testing the model of DNA substitution. Bioinformatics **14:** 817–818.

Remm, M., C. E. V. Storm and E. L. L. Sonnhammer, 2001 Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. **314:** 1041–1052.

Riechmann, J. L., and O. J. Ratcliffe, 2000 A genomic perspective on plant transcription factors. Curr. Opin. Plant Biol. **3:** 423–434.

Rokas, A., B. L. Williams, N. King and S. B. Carroll, 2003 Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425:** 798–804.

Santiago-Valentin, E., and R. G. Olmstead, 2003 Phylogenetics of the Antillean Goetzeoideae (Solanaceae) and their relationships within the Solanaceae based on chloroplast and ITS DNA sequence data. Syst. Bot. **28:** 452–460.

Small, R. L., R. C. Cronn and J. F. Wendel, 2004 Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot. **17:** 145–170.

Small, R. L., and J. F. Wendel, 2000 Copy number lability and evolutionary dynamics of the Adh gene family in diploid and tetraploid cotton (Gossypium). Genetics **155:** 1913–1926.

Sonnhammer, E. L. L., and E. V. Koonin, 2002 Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. **18:** 619–620.

Spooner, D. M., I. E. Peralta and S. Knapp, 2005 Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes. [Solanum L. section Lycopersicon (Mill.) Wettst.] Taxon **54:** 43–61.

Stam, P., 1993 Construction of integrated genetic linkage maps by means of a new computer package: Joinmap. Plant J. **3:** 739–744.

Swofford, D. L., 2003 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, MA.

Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin, 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:** 33–36.

Van Ooijen, J. W., and R. E. Voorrips, 2001 JoinMap version 3.0: software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands.

Wikstrom, N., V. Savolainen and M. W. Chase, 2001 Evolution of the angiosperms: calibrating the family tree. Proc. Biol. Sci. **268:** 2211–2220.

Wolfe, K. H., W.-H. Li and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84:** 9054–9058.

Communicating editor: A. H. D. Brown