Comparison of peach and *Arabidopsis* genomic sequences: fragmentary conservation of gene neighborhoods

Laura L. Georgi, Ying Wang, Gregory L. Reighard, Long Mao, Rod A. Wing, and Albert G. Abbott

Abstract: We examined the degree of conservation of gene order in two plant species, *Prunus persica* (peach) and *Arabidopsis thaliana* (thale cress), whose lineages diverged more than 90 million years ago. In the three peach genomic regions studied, segments with a gene order congruent with *A. thaliana* were short (two to three genes in length); and for any peach region, corresponding segments were found in diverse locations in the *A. thaliana* genome. At the gene level and lower, the *A. thaliana* sequence was enormously useful for identifying likely coding regions in peach sequences and in determining their intron–exon structure. The peach BAC sequence data reported here contained a BLAST-detectable putative coding sequence an average of every 7 kb, and the peach introns identified in this study were, on average, almost twice the length of the corresponding introns in *A. thaliana*.

Key words: conserved microsynteny, genome evolution.

Résumé : Les auteurs ont examiné la conservation de l'ordre des gènes chez deux espèces végétales, le *Prunus persica* (pêcher) et l'*Arabidopsis thaliana* (arabette des dames), dont les ancêtres auraient divergé il y a environ 90 millions d'années. Au sein des trois régions génomiques étudiées chez le pêcher, seuls de courts segments (deux à trois gènes) ont été observés chez lesquels l'ordre génique était identique à celui observé chez l'*A. thaliana*. De plus, pour chaque région du génome du pêcher, des segments correspondants étaient observés à divers endroits dans le génome de l'*A. thaliana*. Au niveau génique ou inférieur, la séquence de l'*A. thaliana* s'avérait très utile pour identifier de possibles régions codantes et la structure des exons–introns au sein des séquences du pêcher. Les séquences des clones BAC du pêcher rapportées ici recelait en moyenne une région codante potentielle (détectable par analyse BLAST) à tous les 7 kb et les introns du pêcher étaient deux fois plus grands en moyenne que ceux de l'*A. thaliana*.

Mots clés : microsynténie conservée, évolution du génome.

[Traduit par la Rédaction]

Introduction

The recent complete genomic sequencing of *Arabidopsis thaliana* (L.) Heynh. (*Arabidopsis* Genome Initiative 2000) provides an unprecedented opportunity to investigate the evolution of plant genomes. The apparent chromosomal-level conservation of genome organization in the grasses maintained over tens of millions of years of evolution (Gale and Devos 1998) suggested that some degree of conservation might extend back two or three times as long to the monocot–dicot divide (Tikhonov et al. 1999). A hybridization-based comparison of a segment of the *A. thaliana* genome with rice (*Oryza sativa* L.) found that some genes clustered in the former are near each other in the latter, but many of the intervening genes in *A. thaliana* are apparently dispersed in rice (van Dodeweerd et al. 1999). Three draft sequences of the rice genome are presently available (Barry 2001; Yu et al. 2002; Goff et al. 2002), and the rice and *Arabidopsis* genomes are now being compared at the sequence level. Mayer et al. (2001) examined a 340-kb segment of rice chromosome 2 and found related genes in several unlinked segments of the *Arabidopsis* sequence. Liu et al. (2001) investigated annotated rice BAC sequences primarily from chromosome 1 and found scant evidence of conserva-

Received 5 September 2002. Accepted 7 January 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 16 March 2003.

Corresponding Editor: O. Rajora.

L.L. Georgi[1] and A.G. Abbott. Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, U.S.A.
Y. Wang[2] and G.L. Reighard. Department of Horticulture, Clemson University, Clemson, SC 29634, U.S.A.
L. Mao[3] and R.A. Wing[4]. Clemson University Genomics Institute, Clemson University, Clemson, SC 29634, U.S.A.

[1]Corresponding author (e-mail: georgil@clemson.edu).
[2]Present address: Department of Plant Breeding, Cornell University, Ithaca, NY 14853, U.S.A.
[3]Present address: Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, U.S.A.
[4]Present address: Arizona Genomics Institute, Tucson, AZ 85721, U.S.A.

 doi: 10.1139/G03-004

tion of gene order with *Arabidopsis*. Although sequences linked in one genome are associated in the other more often than expected on the basis of chance, conserved regions are small and interrupted. These early results were confirmed by Goff et al. (2002), who found limited colinearity and extensive genome rearrangements between the two species in whole-genome sequence comparisons.

Given the apparent stability of genome organization in the grasses, some geneticists (e.g., Tikhonov et al. 1999) hypothesize that the lack of conservation between rice and *Arabidopsis* reflects rapid genome reorganization in dicots, particularly in *Arabidopsis*. Comparisons have been made between *A. thaliana* and its close relatives in the Brassicaceae, all of which are members of the eurosid II clade, as well as with the less closely related legumes (members of the eurosid I clade), and tomato, a distant relative in the other main eudicot clade, the asterids (Soltis et al. 2000). Within the Brassicaceae, comparative physical mapping found conservation of gene order but not repertoire (O'Neill and Bancroft 2000). This is postulated to be a consequence of random gene loss from extensively duplicated sequences in these genomes — 60% of the *A. thaliana* genome is composed of large segmental duplications (*Arabidopsis* Genome Initiative 2000) and *Brassica* species are extensively triplicated (O'Neill and Bancroft 2000). The eurosid I – eurosid II split was investigated by performing homology searches of the *A. thaliana* sequence with molecular genetic markers from soybean (*Glycine max* L. Merr.; Grant et al. 2000; Lee et al. 2001). Shared linkages occur more frequently than would be expected by chance, though genome reorganization was found to be considerable. Again, extensive duplications in both genomes complicated analysis. Several published reports compare sequenced gene-containing regions of tomato and *A. thaliana* (Ku et al. 2000; Rossberg et al. 2001; Mao et al. 2001). As with the *Brassica* comparisons, Ku et al. (2000) and Mao et al. (2001) found the homologs (or homeologs) of their linked tomato sequences scattered around the *A. thaliana* genome on variable-length segments, on which gene order but not repertoire was conserved. This again suggests genome evolution by duplication and selective gene loss. Rossberg et al. (2001) found extensive microcolinearity in the region they studied, although the gene arrangement in tomato (*Lycopersicon esculentum* Mill.) differs by two inversions from that in *A. thaliana*, and one of the tomato genes more strongly resembles genes at other locations in the *A. thaliana* genome. Nonetheless, the *A. thaliana* sequence was successfully exploited to advance the construction of a high-resolution physical map of a region of the tomato genome (Ku et al. 2001).

Thus, comparisons of the degree of conservation of gene organization are now available based on extended (BAC-sized) genomic sequences from monocots, as well as from eudicots representing the asterid and eurosid II clades. In the present study, we present results for peach (*Prunus persica* (L.) Batsch) in the eurosid I (and more specifically the "nitrogen-fixing") clade (Soltis et al. 2000), including one complete BAC-length sequence and a partial sequence from 28 more BACs, predominantly from the region containing the *evg* gene (*evergreen*, also known as *evergrowing*: Wang et al. 2002). In addition to filling in a taxonomic gap in the present data, peach is both the first perennial and the first woody plant examined in this way. It is also at least nominally diploid. With a genome size intermediate between that of *A. thaliana* and rice and having a number of genetic maps, peach is solidly on its way to becoming a model plant species (Abbott et al. 2002).

Materials and methods

The construction, arraying, and hybridization screening of the Nemared peach rootstock BAC library, and the isolation and sequencing of the clone PpN31C7, are described in Georgi et al. (2002). The program Primer3 (Rosen and Skaletsky 1998) was used to develop primers (5′-CCATCACCCAATTAGTCAAT-3′ and 5′-TTCATGCCACTACAGATTCA-3′) flanking a $(GA)_{28}$ simple sequence repeat (SSR) running from nucleotide 39 524 to 39 579 of the PpN31C7 sequence. This marker (pchgms35) was evaluated in 55 F_2 trees from a Nemared × Lovell cross (the K62–68 family; Lu et al. 1998) using the procedure described in Sosinski et al. (2000). In addition, the library was screened with tomato cDNA clone pTC34, corresponding to 240K04.09, a gene encoding a putative auxin (independent) growth promoter (Mao et al. 2001).

Four amplified fragment length polymorphism (AFLP) markers (EAT/MCAC, EAT/MCTA, ETT/CCA2, and ETT/MACC) mapping to the *evg* gene region (Wang et al. 2002) were used to probe the peach BAC library. Positive BAC clones were end sequenced using SP6 and T7 primers (Yu 2000). Additional sequence was obtained by subcloning restriction fragments of the BACs into pUC119 and end sequencing the subclones using M13 forward and reverse primers. Sequencing was performed using Big Dye Terminator Sequencing Kits and an ABI 377 Stretch Sequencer (Applied Biosystems, Foster City, Calif.). The sequence reads were between 300 and 500 bp. Sequences were viewed and compared using Sequencher™ 3.1.1 (Gene Codes Corp., Ann Arbor, Mich.). Dinucleotide frequencies were calculated using GeneWorks 2.3.1 (Intelligenetics, Mountain View, Calif.), and codon frequencies were calculated with the assistance of DNAid+ freeware version 1.8. Only sequences that were not obviously non-functional were used to calculate codon frequencies. Vector sequences were purged using CROSSMATCH (available from http://www.phrap.org).

The peach sequences were compared with the non-redundant (nr) protein database using the BLASTx algorithm (Altschul et al. 1997), either directly on the Web site of the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.gov), or mirrored on the Clemson University Genome Institute server (http://www.genome.clemson.edu). Additional *Arabidopsis* information was obtained from the *Arabidopsis* Information Resource Web site (http://www.arabidopsis.org). tBLASTx comparisons of the PpN31C7 sequence with the *Arabidopsis* AGI BAC genomic sequence database did not result in the detection of significant sequence similarities beyond those identified by BLASTx.

Results and discussion

Complete sequence of a single peach BAC

Peach BAC PpN31C7 was sequenced in its entirety at a one- to seven-fold coverage. Of the 48 443 bases, there was

one base position (11 370) that could not be identified unambiguously. An SSR marker developed from this sequence, pchgms35, was found to segregate in the K62-68 (Nemared × Lovell) mapping population developed by Lu et al. (1998), and mapped between AFLP markers AT/CTC1 and AC/CAA2 on linkage group X. The sequence of PpN31C7 has been deposited in the GenBank database (accession No. AF467900).

Annotation of the PpN31C7 sequence was based on homology searches of GenBank (Table 1). This is an ultraconservative approach that risks missing novel coding regions; however, there is at present insufficient peach sequence data (either genomic or cDNA) on which to train an ab initio gene-finding program. There is little point in relying on programs that have been trained on other species. Even when trained on abundant *Arabidopsis* sequence, these programs can be expected to predict *Arabidopsis* genes correctly only half of the time (Pertea and Salzberg 2002), and in a typical five-exon gene they can be expected to misidentify one of the exons (Brendel and Zhu 2002). Thus any gene structure not supported by cDNA sequence needs to be viewed with a degree of skepticism. On the other hand, the *A. thaliana* sequence is enormously useful for identifying likely coding regions in sequences from other plants and determining their intron–exon structure. Of course, this sort of comparison should also result in improvement of the annotation of the *A. thaliana* sequence. It is also clear that for now, at least, there is no entirely satisfactory substitute for a cDNA sequence. We are presently one third of the way to our goal of sequencing 30 000 peach cDNAs; when this resource is available, a re-evaluation of the annotation of the PpN31C7 sequence will be undertaken. Information on the completed cDNA sequences may be found at http://www.genome.clemson.edu/projects/peach/est/.

Overall G/C content of this sequence was 38.5%. Dinucleotides CG and TA were under represented (Table 2), as is common in eukaryotic sequences (Karlin et al. 1998). This bias was also reflected in codon usage (Table 3). Codon usage patterns generally resembled those previously reported for 46 peach coding sequences in the codon-usage database (http://www.kazusa.jp/codon; 15 August 2002).

PpN31C7.1 resembled *Arabidopsis thaliana* gene *F6F9.9* on chromosome 1, which encodes a hypothetical protein. The genomic sequences contained three introns whose locations were conserved; all three were smaller in peach than in *Arabidopsis* (Table 4). The next identified coding sequence in peach (PpN31C7.2) was on the same strand, and the distance from the TAA stop codon at the end of PpN31C7.1 to the ATG translation start site of PpN31C7.2 was 790 bp. The GeneMark program (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi; relying on *Arabidopsis* settings) interpreted this interval as part of an intron, and combined PpN31C7.1 and PpN31C7.2 in a single transcript.

PpN31C7.2 resembled *A. thaliana* gene *MIJ24.9* on chromosome 5, encoding an unknown protein. Both sequences contain an in-frame stop 15 bp upstream of the presumed ATG translation start site, as well as one intron in the translated region whose position is conserved. In this instance, the peach intron is larger than the corresponding *Arabidopsis* intron (Table 4). There is also a 283-bp intron in the 5′ untranslated region of the *Arabidopsis* sequence. The peach se-

Table 1. Regions of peach BAC PpN31C7 with significant similarity to *A. thaliana* sequences

ID	Strand	Position (nucleotides)	BLASTx *E* value	*A. thaliana* chromosome	*A. thaliana* clone*	Predicted coding sequences	Predicted gene product
PpN31C7.1	+	4 079† – 5 548†	8×10^{-66}	1	F6F9.9	At1g19860–68300.m02157	Hypothetical protein
PpN31C7.2	+	6 350 – 7 133	6×10^{-48}	5	MIJ24.9	At5g39600–68299.m03614	Unknown protein
PpN31C7.3	–	10 574 – 7 787	5×10^{-66}	5	MIJ24.8	At5g39590–68299.m03613	Unknown protein
PpN31C7.4	+	13 083† – 16 713	0	3	MFE16.9	At3g26560–68298.m02926	Putative DEAH helicase
PpN31C7.5	+	17 499 – 18 758†	1×10^{-172}	3	K17E12.5	At3g27230–68298.m03005	Unknown protein
PpN31C7.6	–	24 150† – 19 622†	1×10^{-100}	1	F6F9.10	At1g19850–68300.m02156	IAA24/monopteros/ARF5
PpN31C7.7	+	36 926 – 37 374	6×10^{-43}	1	F6F9.11	At1g19840–68300.m02155	Hypothetical protein
PpN31C7.8	+	41 088† – 42 000†	8×10^{-19}	5	F6O13.20	At2g15650–68297.m01483	Putative retroelement(s)

*GenBank searches performed on 2 October 2001 and 5 December 2001.
†Approximate positions.

Table 2. Dinucleotide frequencies in peach BAC PpN31C7.

Dinucleotide	Relative abundance*
AA/TT	1.12
AC/GT	0.86
AG/CT	1.04
AT	0.94
CA/TG	1.16
CC/GG	1.12
CG	**0.56**
GA/TC	1.05
GC	1.03
TA	**0.75**

*$f_{XY}/f_X f_Y$, where f_{XY} is the frequency of dinucleotide XY, f_Y is the frequency of nucleotide Y, and f_X is the frequency of nucleotide X. Values less than 0.78 or more than 1.23 are considered significant and are shown in bold (Karlin et al. 1998).

quence had a potential splice acceptor sequence at the same position; however, without a cDNA sequence, identification of the corresponding donor site would only be based on speculation. The final exon provided another argument in support of sequencing cDNA clones, because it was missed in the annotation of the *Arabidopsis* genomic sequence and even by tBLASTx comparison of the two BAC sequences using the program's default settings. Thus, the last exon in the peach sequence was identified by direct inspection — this was possible because the deduced amino acid sequences contained six identities and two conservative changes in an eight-residue window.

The next identified coding sequence in peach (PpN31C7.3) was convergently transcribed with the previous one, and the distance between the two stop codons was 651 bp. PpN31C7.3 resembled *A. thaliana* gene *MIJ24.8*, which is convergently transcribed with *MIJ24.9*. It also encodes an unknown protein. The stop-to-stop distance between the two *Arabidopsis* coding sequences is 476 bp. The sequences encoding PpN31C7.3 and *MIJ24.8* each contained six introns (Table 4) and their positions were perfectly conserved. The peach introns in this coding sequence were correctly identified by GeneMark.

The fourth putative coding sequence was once again on the plus strand of the sequence as deposited in GenBank, and the initiation codon, arbitrarily selected from all possibilities to the right of the PpN31C7.3 coding sequence, was over 2 kb away. PpN31C7.4 strongly resembled *MFE16.9* on *Arabidopsis* chromosome 3, which encodes a putative DEAH helicase. The position of the single intron was conserved between peach and *Arabidopsis*. Closely related genes are present in the genomes of *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. The *S. cerevisiae* gene, *PRP22*, is essential, and its product is involved in mRNA maturation (Schneider and Schwer 2001). The *C. elegans* gene *mog-5* is required for post-transcriptional repression of *fem-3*, thus permitting hermaphrodites to cease spermatogenesis and begin oogenesis, and is required maternally for embryogenesis (Puoti and Kimble 2000). *PRP22* and *mog-5* appear to be orthologous (Sanjuán and Marín 2001). Interestingly, PpN31C7.4 appeared to be non-functional. The sequence contained six in-frame stops (one in conserved motif Ib — see Jankowsky and Jankowsky 2000), substitutions for conserved amino acid residues in motifs Ia and Ic and two substitutions in motif VI, a deletion of a conserved glutamate in motif V, and two frame shifts, one of which resulted in the deletion of five amino acid residues in a stretch of identity with *Arabidopsis*. The peach BAC sequence was single pass in parts of this region, but two of the stops were supported by high-quality sequence on both strands, and all suspect sequences appeared convincing on direct inspection. In any case, the sheer number of "fatal" alterations makes it unlikely that this particular gene sequence is functional. The high degree of evolutionary conservation and the lethal consequences of loss of function in budding yeast and *C. elegans* indicate that this is an important gene. It is likely that peach has a functional copy, either allelic to the sequenced gene or elsewhere in the genome. Hybridization of the library filters with a fragment of the 31C7.4 sequence resulted in the identification of 33 positive BACs, including all four BACs originally detected with the pTC11 probe. Because the library coverage is estimated to be six to eight fold (Georgi et al. 2002), there could be as many as three or four related sequences in the peach genome.

The fifth identified coding sequence was on the same strand as the fourth. The distance from the final stop in PpN31C7.4 to the initial ATG in PpN31C7.5 was 782 bp. PpN31C7.5 resembled *A. thaliana* gene *K17E12.5*, which encodes an unknown protein. This is the similarity responsible for the selection of this BAC from the library, because the tomato probe used, pTC11, is related to K17E12.5. The gene appears to consist of a single exon in all three species. The *Arabidopsis* sequence is on chromosome 3, approximately 250 000 bases from the sequence for the DEAH helicase discussed previously. The tomato sequence is contained in tomato BAC clone 240K04, which also contains several other genes that matched peach sequences described below.

PpN31C7.6 resembled *A. thaliana* transcription factor gene *IAA24*/*monopteros*/*ARF5*, also known as *F6F9.10* (Kim et al. 1997; Hardtke and Berleth 1998; Ulmasov et al. 1999). *IAA24* is convergently transcribed with *F6F9.9*; the stop-to-stop distance in the *Arabidopsis* sequence is 471 bp. In peach, PpN31C7.6 was also encoded on the opposite strand from PpN31C7.1, but the distance between their stop codons was 14 kb, a space occupied by sequences corresponding to genes found on two other chromosomes in *Arabidopsis*. The stop codon for PpN31C7.5 (on the same strand as PpN31C7.1) is 850 bp from the stop for PpN31C7.6. The peach sequence contained one additional intron not found in *Arabidopsis IAA24* (Table 4), but the remaining 12 intron positions were conserved. All but one of the peach introns were correctly identified by GeneMark (the 12th splice donor site was misplaced).

PpN31C7.7 resembled *A. thaliana* gene *F6F9.11*, which encodes a hypothetical protein. (A better BLASTx score was obtained with *A. thaliana* gene *F10A5.20*, but *F6F9* yielded the higher BLASTn score). PpN31C7.7 contained a frame shift with respect to the *Arabdopsis* sequence. Only one strand was sequenced in peach in the region in question, but

Table 3. Codon frequency for PpN31c7 (five coding sequences*).

Amino acid	Codon	Frequency	Amino acid	Codon	Frequency	Amino acid	Codon	Frequency	Amino acid	Codon	Frequency
Phe	TTT	62	Leu	CTT	57	Ile	ATT	62	Val	GTT	76
Phe	TTC	46	Leu	CTC	32	Ile	ATC	29	Val	GTC	21
Leu	TTA	15	Leu	CTA	35	Ile	ATA	32	Val	GTA	18
Leu	TTG	58	Leu	CTG	39	Met	ATG	57	Val	GTG	43
Ser	TCT	59	Pro	CCT	66	Thr	ACT	45	Ala	GCT	43
Ser	TCC	33	Pro	CCC	25	Thr	ACC	19	Ala	GCC	36
Ser	TCA	67	Pro	CCA	48	Thr	ACA	43	Ala	GCA	64
Ser	TCG	29	Pro	CCG	18	Thr	ACG	8	Ala	GCG	15
Tyr	TAT	27	His	CAT	38	Asn	AAT	74	Asp	GAT	93
Tyr	TAC	21	His	CAC	19	Asn	AAC	50	Asp	GAC	42
Och	TAA	3	Gln	CAA	54	Lys	AAA	56	Glu	GAA	63
Amb	TAG	0	Gln	CAG	64	Lys	AAG	70	Glu	GAG	68
Cys	TGT	23	Arg	CGT	7	Ser	AGT	47	Gly	GGT	49
Cys	TGC	25	Arg	CGC	12	Ser	AGC	44	Gly	GGC	27
Opa	TGA	2	Arg	CGA	11	Arg	AGA	40	Gly	GGA	60
Trp	TGG	33	Arg	CGG	9	Arg	AGG	30	Gly	GGG	32

*PpN31C7.4, PpN31C7.7, and PpN31C7.8 were excluded from the calculations because they contain frame-shifts, premature termination codons, and (or) other alterations relative to the corresponding *Arabidopsis* sequence and are presumably non-functional (see text).

Table 4. Comparison of intron sizes in putative coding sequences in peach BAC PpN31C7 with the corresponding sequences in *A. thaliana*.

Coding sequence	Peach intron sizes (bp)	*Arabidopsis* intron sizes (bp)
PpN31C7.1	171 129 87	494 298 97
PpN31C7.2	? 397	283 200
PpN31C7.3	95 271 387 114 214 90	87 150 148 83 104 85
PpN31C7.4	774	91
PpN31C7.5	—	—
PpN31C7.6	207 252 110 103 105 174 295 120 262 133 76 91 87	150 108 90 88 79 76 96 82 85 0 7 9 92 96
PpN31C7.7	—	—

both reads were relatively clean. Both PpN31C7.7 and *F6F9.11* were predicted to consist of a single exon encoded on the opposite strand from PpN31C7.6 (*IAA24*). Start-to-start distances were 12 and 14 kb in peach and *Arabidopsis*, respectively. GeneMark detected a possible coding sequence in this interval in peach, but a tBLASTn search of the resulting predicted amino acid sequence failed to find any convincing matches in the database.

Finally, PpN31C7.8 contained two segments with BLASTx similarity with presumed *copia*-like retrotransposons. The intervening 436 bp contained stops in all three frames, and the BLAST-positive fragments on either end contained frame shifts. They appeared not to encode a complete autonomous transposon, either individually or in combination. *Copia*-like retrotransposons have been identified previously in other *Prunus* species using degenerate primer PCR (Hirochika and Hirochika 1993; Asíns et al. 1999), but these sequence entries were not detected when GenBank was queried with PpN31C7. Based on hybridization with a PCR-amplified reverse transcriptase gene fragment from apricot, Asíns et al. (1999) concluded that *copia*-like elements are rare in *Prunus*. Apart from this apparently degenerate transposon, this end of the PpN31C7 sequence contained many repeats and no other sequences with obvious similarity to known genes.

One non-coding region to note was a 28 bp inverted repeat at bases 291–318 and 365–393 in the sequence. It is conceivable that this is a MITE (miniature inverted repeat transposable element, Bureau et al. 1996), although there was no apparent duplication of target sequence. Part of this repeat was a perfect match with a direct repeat on *A. thaliana* chromosome 2. It was not found anywhere else in the entire *Arabidopsis* sequence, so the *Arabidopsis* sequence does not appear to be a MITE.

PpN31C7 contained a BLAST-detectable putative coding sequence on average every 7 kb. By comparison, *A. thaliana* has a gene on average every 4.5 kb (*Arabidopsis* Genome Initiative 2000). This is only a rough preliminary estimate of peach gene density, as it is based on a very limited sample that was biased toward a likely gene-containing region of the genome, and overlooks coding sequences that are unique or too diverged from sequences already identified to allow their detection by BLAST. The peach introns identified in this study were on average almost twice the length of the corresponding introns in *A. thaliana*. The peach genome is

approximately 2.4 times the size of the *A. thaliana* genome (Baird et al. 1994), and our sequence suggests that some of the difference in size may be due to the enlargement of genes and introns within genes.

Partial sequence of multiple peach BACs

One hundred seventy unique peach genomic sequences were obtained, including 114 subclone end sequences from two BACs (43 from PpN106F24 and 71 from PpN70H22) and 56 BAC end sequences (BESs) from 28 BACs. Except for PpN106F24, all of the BAC clones are from an 18.4-cM region that covers the *evg* gene. Because some of the BACs have unknown orientations, the exact physical positions of some sequences were not known. All the peach sequences were compared with the nr protein database using the BLASTx algorithm, which translates the query DNA sequence in all six possible reading frames for comparison (Altschul et al. 1997). Overall, 22 unique peach sequences, about 13%, were matched to similar sequences in the database, and have been deposited in GenBank under accession Nos. BZ412011–BZ413031 and BZ413154. For a majority of the peach sequences, the most similar sequence in the database was from *Arabidopsis*; with the rest being most similar to sequences from other plant species (soybean, tomato, spinach, garden pea), but with sequences from *Arabidopsis* ranking as the second or the third most similar sequence. Therefore, we chose to list only the *Arabidopsis* sequences in Tables 5 and 6.

No one segment in the *Arabidopsis* genome appears to correspond to the *evg* region in peach. Of the 54 BESs from the *evg* region, 12 matched sequences from *Arabidopsis*, but were scattered throughout the genome. Five peach BAC end sequences matched sequences on *Arabidopsis* chromosome 4; but for one pair showing possible conservation of neighbourhood, they did not cluster. Bes_18F12B01 and bes_30D10T7 matched two nearby genes in *Arabidopsis* BAC clone F22K18. However, the bes_18F12B01 sequence was approximately 60 to 70 kb from bes_30D10T7 in the peach genome, according to fingerprinting and contig assembly of the BAC clones in this region (data not shown). Bes_18F12B01 resembled the AGL24 MADS-box protein in *Arabidopsis* (Table 5) and the MADS-box transcription factor JOINTLESS in tomato (tomato BAC 240K04 with E value 2.0E–15). Bes_30D10T7 resembled a PsRT17-1-like gene and an auxin-independent, growth-promoter gene in *Arabidopsis* (on chromosomes 4 and 5, respectively) and a putative auxin growth-promotor gene in tomato BAC 240K04. Therefore, the association of these two genes was conserved among the *Arabidopsis*, tomato, and peach genomes. Beyond these two genes, evidence of conservation was difficult to find. The next closest match on *Arabidopsis* chromosome 4 — that with bes_18F22SP6 — was 2.1 Mb away from the matches with bes_18F12B01 and bes_30D10T7. Although the *Arabidopsis* matches with bes_22H21T7 and bes_89G2SP6 were only about 250 kb from each other, they were 10 Mb from the apparently conserved gene pair. Two complementary overlapping BAC end sequences, bes_35A23T7 (530 bp) and bes_48E24T7 (555 bp), were found homologous to the kinesin-like protein in the MSL3 clone on chromosome 5 in *Arabidopsis*. Interestingly, another tomato gene in 240K04 BAC clone

Table 5. BLASTx matches with *A. thaliana* sequences from the non-redundant (nr) protein database for peach BAC end sequences (BES) in the *evergreen* gene region, listed by chromosome location of the *A. thaliana* sequence.

Peach DNA sequences	*E* value	*Arabidopsis* Chromosome	*Arabidopsis* clone	Predicted coding sequences	*Arabidopsis* proteins
bes_18F22T7	2.00×10^{-12}	1	F21M12.10	At1g09710–68300.m00934	Putative Myb-family transcription factor
bes_96J3T7	5.00×10^{-14}	1	F5O11.11	At1g12390–68300.m01230	Hypothetical protein
bes_83E10sp6	3.00×10^{-20}	2	T9F8.8	At2g06890–68297.m00647	Putative retroelement integrase
bes_96J3SP6	8.00×10^{-13}	2	F4L23.8	At2g45410–68297.m04830	Unknown protein
bes_22H21T7	8.00×10^{-20}	4	F25I24.40	At4g10830–68296.m01075	Putative protein various reverse transcriptases and transposons
bes_89G2sp6	3.00×10^{-39}	4	FCAALL.393	At4g17280–68296.m01784	Full-length cDNA clone sequences (unknown protein)
bes_30D10T7	9.00×10^{-47}	4	F22K18.270	At4g24530–68296.m02594	PsRT17-1 like protein
bes_18F12B01*	3.00×10^{-20}	4	F22K18.260	At4g24540–68296.m02595	MADS-box protein AGL24
bes_18F22SP6	2.00×10^{-29}	4	F9N11.50	At4g30200–68296.m03283	Putative protein
bes_89G2sp6	9.00×10^{-41}	5	MNJ7.12	At5g47530–68299.m04521	Similar to unknown protein
bes_35A23T7 / bes_48E24T7†	3.00×10^{-21} / 6.00×10^{-22}	5	MSL3.50	At5g60930–68299.m06058	Kinesin-like protein
bes_30D10T7‡	3.00×10^{-43}	5	MNA5.21	At5g65470–68299.m06575	Contains similarity to auxin-independent growth promoter gene

Note: Where a peach BES matched two *A. thaliana* sequences with similar *E* values, both are listed.
*this sequence is also homologous to the MADS-box transcription factor JOINTLESS in tomato (240K04.14) (E value 2.0×10^{-15}).
†bes_35A23T7 and bes_48E24T7 are complementary overlapping sequences.
‡This sequence is also homologous to the putative auxin growth promotor protein in tomato (240K04.09) (*E* value 1.0×10^{-33}).

Table 6. BLASTx matches between sequences in the non-redundant protein database and peach sequences from two BAC clones, PpN70H22 (from the *evg* gene region) and PpN106F24, including *Hin*dIII-digested subclone end sequences and BAC end sequences (BES).

Peach DNA sequences	*E* value	*Arabidopsis* chromosome	*Arabidopsis* clone	Predicted coding sequence	*Arabidopsis* protein
Sequences from PpN106F24					
106F24H_P2B6R	5.00×10^{-34}	1	T14P4.9	At1g02630–68300.m00190	Hypothetical protein
106F24H_P2D3R	1.00×10^{-20}	1	F18B13.21	At1g80130–68300.m08149	Unknown protein
106F24H_P2D3F	7.00×10^{-14}	2	F8D23.4	At2g18180–68297.m01770	Putative phosphatidylinositol–phophatidylcholine transfer protein
bes_106F24_T7	1.00×10^{-53}	2	F13B15.15	At2g25490–68297.m02595	F-box protein family
106F24H_P3G12R	5.0×10^{-10}	4	T22A6.230	At4g24400–68296.m02580	Serine (threonine) kinase like protein
Sequences from PpN70H22					
70H22H_P1H9F	6.0×10^{-19}	1	F21M12.35	At1g09960–68300.m00961	Similar to *Vicia* sucrose transport protein (gb|Z93774)
70H22H_P1D9R	3.0×10^{-15}	1	T17H3.15	At1g27595–68300.m02937	Hypothetical protein
70H22H_P3A9F	5.0×10^{-14}	1	T22C5.3	At1g27595–68300.m02937	Hypothetical protein
70H22H_P1B10F	2.0×10^{-15}	1	F19C14.5	At1g58340–68300.m05765	Unknown protein contains Pfam profile
bes_70H22_SP6	9.0×10^{-32}	2	F13B15.15	At2g25490–68297.m02595	F-box protein family
70H22H_P1H3R	1.0×10^{-32}	5	MCM23.1	At5g17930–68299.m01866	Unknown protein

(*240K04.15*) is homologous to the kinesin heavy-chain protein gene (*E* value 5.00×10^{-80}). In contrast, peach BAC PpN31C7 contained apparently the sole peach homolog of a fourth tomato gene in 240K04 (240K04.11, encoding an unknown protein). In tomato, 240K04.11 is between the hypothetical auxin growth promoter gene and the MADS box gene; whereas in peach, it was in a separate BAC clone that not only did not overlap the *evg* region, but was also apparently at least 40 cM from it on the genetic map (Wang et al. 2002), based on the map locations of pchgms35, an SSR marker developed from nearby PpN31C7 sequence. Similarly, a tomato cDNA corresponding to the hypothetical auxin growth-promoter gene (pTC34) hybridized on the peach BAC library array to 52 clones, including five from the *evg* region (PpN12E12, PpN18G7, PpN30D10, PpN48E24, PpN94F16), but not to PpN31C7 or BACs known to overlap it (PpN18B20, PpN35O1, PpN92L14). Further investigation of collinearity between this peach region and tomato BAC 240K4 will be pursued.

Most of the peach sequences match a single region in *Arabidopsis* at very low *E* value, but two peach sequences (bes_89G2Sp6 and bes_30D10T7) each matched two *Arabidopsis* sequences from different chromosomal positions. This is not surprising, because 37.4% of *Arabidopsis* proteins are present in families of more than 5 members (Bevan et al. 2001). Large gene families complicate the identification of true orthologs from different species, especially when only subsets of sequences are available (Yamamoto and Knap 2001).

Table 6 shows matches with the non-redundant protein database for sequences from the two peach BAC clones PpN106F24 and PpN70H22, with an inclusion value of $E < 10^{-10}$. Five regions in *Arabidopsis* matched sequences in PpN106F24, including four random subclone end sequences and one BAC end sequence. However, these five genes were located on three different *Arabidopsis* chromosomes, and the two on the same chromosome (1; *Arabidopsis* BAC clones T14P4 and F18B13) were separated by about 28.7 Mb. Because the two corresponding peach sequences were from random subclones from a single BAC, we know the maximum distance between them was about the size of the BAC clone, 50 kb. Therefore, the level of conservation of gene neighborhoods between peach and the *Arabidopsis* genome was very low in the region represented by this BAC. For the peach BAC clone PpN70H22 (estimated to be 120 kb), six matches were found with *Arabidopsis*: four from chromosome 1, one from chromosome 2, and one from chromosome 5. On chromosome 1, two of homologues were very close to each other (within approximately 60 kb), whereas the other two were about 6 and 10 Mb away, respectively. Therefore, conservation of gene neighborhoods between these two species was limited and fragmentary at best. Of note is one database hit (F13B15.15) that is shared between BACs PpN106F24 and PpN70H22. Because microsatellites developed from these two BACs have been shown not to be genetically linked (Wang et al. 2002), this gene may be duplicated in the peach genome.

Conclusions

We have examined the degree of conservation of gene order in two plant species, *Prunus persica* and *Arabidopsis*

thaliana, whose lineages diverged more than 90 million years ago (Magallón et al. 1999). Toward this end, we obtained sequence data from three genomic regions. In all three peach genomic regions we studied, gene neighborhoods conserved with *A. thaliana* were small and corresponded to regions dispersed throughout the *A. thaliana* genome. This outcome is not surprising, given earlier reports that gene order, but not repertoire, is conserved even within the Brassicaceae (O'Neill and Bancroft 2000). Likewise, molecular markers from soybean revealed considerable genome rearrangement when mapped in silico to *A. thaliana* (Grant et al. 2000; Lee et al. 2001). Thus, genetic mapping projects in peach and other dicots may do better to use markers derived from the *A. thaliana* sequence rather than entirely random markers, but the *A. thaliana* based markers need to be treated with a healthy skepticism. In this regard, dicot researchers need not envy their colleagues working with monocots, because even in monocots, colinearity contains holes when examined closely. For example, Tikhonov et al. (1999) found rearrangements between maize and sorghum even in the *adh1* region, which is substantially conserved, and the *Adh1* gene is in a completely unrelated region in rice (Tarchini et al. 2000). Nonetheless, it does appear that dicot genomes have been unusually prone to rearrangements on a frustratingly fine scale, and it is tempting to see a connection between the huge evolutionary success of the eudicots and their genomic diversification.

Acknowledgments

This work was supported by the South Carolina Agriculture and Forestry Research System (SCAFRS) under project SC-1700120 and by the United States Department of Agriculture Cooperative State Research, Education, and Extension Service (CSREES) special research grant "Peach Tree Short Life in South Carolina". This report is technical contribution No. 4793 of the SCAFRS.

References

Abbott, A., Georgi, L., Yvergniaux, D., Iñigo, M., Sosinski, B., Wang, Y., Blenda, A., and Reighard, G. 2002. Peach: the model genome for Rosaceae. Acta Hort. **575**: 145–155.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

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

Asíns, M.J., Monforte, A.J., Mestre, P.F., and Carbonell, E.A. 1999. *Citrus* and *Prunus copia*-like retrotransposons. Theor. Appl. Genet. **99**: 503–510.

Baird, W.V., Estager, A.S., and Wells, J.K. 1994. Estimating nuclear DNA content in peach and related diploid species using laser flow cytometry and DNA hybridization. J. Am. Soc. Hort. Sci. **119**: 1312–1316.

Barry, G.F. 2001. The use of the Monsanto draft rice genome sequence in research. Plant Physiol. **125**: 1164–1165.

Bevan, M., Mayer, K., White, O., Eisen, J.A., Preuss, D., Bureau, T., Salzberg, S.L., and Mewes, H.W. 2001. Sequence and analysis of the *Arabidopsis* genome. Curr. Opin. Plant Biol. **4**: 105–110.

Brendel, V., and Zhu, W. 2002. Computational modelling of gene structure in *Arabidopsis thaliana*. Plant Mol. Biol. **48**: 49–58.

Bureau, T.E., Ronald, P.C., and Wessler, S.R. 1996. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Natl. Acad. Sci. U.S.A. **93**: 8524–8529.

Gale, M.D., and Devos, K.M. 1998. Comparative genetics in the grasses. Proc. Natl. Acad. Sci. U.S.A. **95**: 1971–1974.

Georgi, L.L., Wang, Y., Yvergniaux, D., Ormsbee, T., Iñigo, M., Reighard, G.L., and Abbott, A.G. 2002. Construction of a BAC library and its application to the identification of simple sequence repeats in peach (*Prunus persica* [L.] Batsch). Theor. Appl. Genet. **105**: 1151–1158.

Goff, S., Ricke, D., Lan, T., Presting, G., Wang, R., et al. 2002. A draft sequence of the rice genome (*Oryza sative* L. ssp. *japonica*). Science (Washington, D.C.) **296**: 92–100.

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

Hardtke, C.S., and Berleth, T. 1998. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor medating embryo axis formation and vascular development. EMBO J. **17**: 1405–1411.

Hirochika, H., and Hirochika, R. 1993. *Ty1-copia* group retrotransposons as ubiquitous components of plant genomes. Jpn. J. Genet. **68**: 35–46.

Jankowsky, E., and Jankowsky, A. 2000. The DExH/D protein family database. Nucleic Acids Res. **28**: 333–334.

Karlin, S., Campbell, A.M., and Mrázek, J. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**: 185–225.

Kim, J., Harter, K., and Theologis, A. 1997. Protein-protein interactions among the Aux/IAA proteins. Proc. Natl. Acad. Sci. U.S.A. **94**: 11786–11791.

Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. 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. U.S.A. **97**: 9121–9126.

Ku, H.-M., Liu, J., Doganlar, S., and Tanksley, S.D. 2001. Exploitation of *Arabidopsis*–tomato synteny to construct a high-resolution map of the *ovate*-containing region in tomato chromosome 2. Genome, **44**: 470–475.

Lee, J.M., Grant, D., Vallejos, C.E., and Shoemaker, R.C. 2001. Genome organization in dicots. II. *Arabidopsis* as a "bridging species" to resolve genome evolution events among legumes. Theor. Appl. Genet. **103**: 765–773.

Liu, H., Sachidanandam, R., and Stein, L. 2001. Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. Genome Res. **11**: 2020–2026.

Lu, Z.X, Sosinski, B., Reighard, G.L., Baird, W.V., and Abbott, A.G. 1998. Construction of a linkage map in peach rootstocks [*Prunus persica* (L.) Batsch.], and localization of genes conferring resistance to root-knot nematodes (*Meloidogyne incognita* and *M. javanica*). Genome, **41**: 199–207.

Magallón, S., Crane, P.R., and Herendeen, P.S. 1999. Phylogenetic pattern, diversity, and diversification of eudicots. Ann. Mo. Bot. Gard. **86**: 297–372.

Mayer, K, Murphy, G., Tarchini, R., Wambutt, R., Volckaert, G., et al. 2001. Conservation of microstructure between a sequenced region of the genome of rice and mutiple segments of the genome of *Arabidopsis thaliana.* Genome Res. **11**: 1167–1174.

Mao, L., Begum, D., Goff, S.A., and Wing, R.A. 2001. Sequence and analysis of the tomato *JOINTLESS* locus. Plant Physiol. **126**: 1331–1340.

O'Neill, C.M., and Bancroft, I. 2000. Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. 2000. Plant J. **23**:233–243.

Pertea, M. and. Salzberg, S.L. 2002. Computational gene finding in plants. Plant Mol. Biol. **48**: 39–48.

Puoti, A., and Kimble, J. 2000. The hermaphrodite sperm–oocyte switch requires the *Caenorhabditis elegans* homologs of PRP2 and PRP22. Proc. Natl. Acad. Sci. U.S.A. **97**: 3276–3281.

Rosen, S., and Skaletsky, H.J. 1998. Primer3. Code available at http://www.genome.wi.mit.edu/genome_software/other/primer3.html.

Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., and Schmidt, R. 2001. Comparative sequence analysis reveals extensive microcolinearity in the *lateral suppressor* regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell, **13**: 979–988.

Sanjuán, R., and Marín, I. 2001. Tracing the origin of the compensasome: evolutionary history of DEAH helicase and MYST acetyltransferase gene families. Mol. Biol. Evol. **18**:330–343.

Schneider, S., and Schwer, B. 2001. Functional domains of the yeast splicing factor Prp22p. J. Biol. Chem. **276**: 21184–21191.

Soltis, D. E., Soltis, P.S., Chase, M.W., Mort, M.E., Albach, D.C., et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbc*L and *atp*B sequences. Bot. J. Linn. Soc. **133**: 381–461.

Sosinski, B., Gannavarapu, M., Hager, L.D., Beck, L.E., King, G.J., Ryder, C.D., Rajapakse, S., Baird, W.V., Ballard, R.E., and Abbott, A.G. 2000. Characterization of microsatellite markers in peach [*Prunus persica* (L.) Batsch.]. Theor. Appl. Genet. **101**: 421–428.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. 2000. The complete sequence of 340 kb of DNA around the rice *Adh1–Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell, **12**: 381–391.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. U.S.A. **96**:7409–7414.

Ulmasov, T., Hagen, G., and Guilfoyle, T.J. 1999. Activation and repression of transcription by auxin-response factors. Proc. Natl. Acad. Sci. U.S.A. **96**:5844–5849.

van Dodeweerd, A.-M., Hall, C.R., Bent, E.G., Johnson, S.J., Bevan, M.W., and Bancroft, I. 1999. Identification and analysis of homoeologous segments of the genomes of rice and *Arabidopsis thaliana*. Genome, **42**: 887–892.

Wang, Y., Georgi, L., Reighard, G.L., Scorza, R., and Abbott, A.G. 2002. Genetic mapping the evergreen gene in peach [*Prunus persica* (L) Batsch]. J. Hered. **93**: 352–358.

Wang, Y., Georgi, L.L., Zhebentyayeva, T.N., Reighard, G.L., Scorza, R., and Abbott, A.G. 2002. High throughput targeted SSR marker development in peach (*Prunus persica*). Genome, **45**: 319–328.

Yamamoto, E., and Knap H.T. 2001. Soybean receptor-like protein kinase genes: paralogous divergence of a gene family. Mol. Biol. Evol. **18**: 1522–1531.

Yu, J., Hu, S., Wang, J., Wong, G.K., Lin, S., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science (Washington, D.C.) **296**: 79–92.

Yu, Y. 2000. Development and application of genomics tools for analysis of grass genome. Ph.D. dissertation, Clemson University, Clemson, S.C. pp. 22–45.