Characterizing the citrus cultivar Carrizo genome through 454 shotgun sequencing

William R. Belknap, Yi Wang, Naxin Huo, Jiajie Wu, David R. Rockhold, Yong Q. Gu, and Ed Stover

Abstract: The citrus cultivar Carrizo is the single most important rootstock to the US citrus industry and has resistance or tolerance to a number of major citrus diseases, including citrus tristeza virus, foot rot, and Huanglongbing (HLB, citrus greening). A Carrizo genomic sequence database providing approximately 3.5× genome coverage (haploid genome size approximately 367 Mb) was populated through 454 GS FLX shotgun sequencing. Analysis of the repetitive DNA fraction indicated a total interspersed repeat fraction of 36.5%. Assembly and characterization of abundant citrus Ty3/*gypsy* elements revealed a novel type of element containing open reading frames encoding a viral RNA-silencing suppressor protein (*RNA binding protein*, *rbp*) and a plant cytokinin riboside 5′-monophosphate phosphoribohydrolase-related protein (*LONELY GUY*, *log*). Similar *gypsy* elements were identified in the *Populus trichocarpa* genome. Gene-coding region analysis indicated that 24.4% of the nonrepetitive reads contained genic regions. The depth of genome coverage was sufficient to allow accurate assembly of constituent genes, including a putative phloem-expressed gene. The development of the Carrizo database (http://citrus.pw.usda.gov/) will contribute to characterization of agronomically significant loci and provide a publicly available genomic resource to the citrus research community.

Key words: Citrus sinensis, *Poncirus trifoliata*, sequencing, retrotransposon.

Résumé : Le citrange Carrizo est le plus important porte-greffes pour l'industrie des agrumes aux Etats-Unis et il confère la résistance ou la tolérance à bon nombre des plus importantes maladies des agrumes dont le virus de la tristeza, la pourriture brune et le Huanglongbing (HLB, maladie du Dragon jaune). Une base de données de séquences génomiques du Carrizo fournissant une couverture d'environ 3,5× du génome (génome haploïde d'environ 367 Mb) a été produite en procédant à du séquençage génomique aléatoire sur un appareil 454 GS FLX. L'analyse de la faction d'ADN répété a indiqué que les répétitions dispersées constituent environ 36,5 % du génome. L'assemblage et une caractérisation des éléments abondants Ty3/*gypsy* a révélé un nouveau type d'élément qui contient des cadres de lectures codant pour un inhibiteur viral de l'inactivation génique (*RNA binding protein*, *rbp*) et une protéine apparentée à une cytokinine riboside 5′-monophosphate phosphoribohydrolase végétale (*LONELY GUY*, *log*). Des éléments *gypsy* similaires ont été identifiés au sein du génome du *Populus trichocarpa*. Une analyse des régions codantes a indiqué que 24,4 % des séquences dépourvues de répétitions correspondaient à des régions géniques. Le degré de couverture génomique était suffisant pour permettre un assemblage précis des gènes de cette espèce, incluant un gène qui serait exprimé dans le phloème. La mise sur pied d'une base de données Carrizo (http://citrus.pw.usda.gov/) contribuera à la caractérisation de locus d'importance agronomique et fournira une ressource génomique publique pour la communauté de chercheurs travaillant sur les agrumes.

Mots-clés : Citrus sinensis, *Poncirus trifoliata*, séquençage, rétrotransposon.

[Traduit par la Rédaction]

Introduction

Citrus is an economically important fruit crop in many countries, with the highest total global production levels of any tree crop (FAO 2010). The production of citrus worldwide is currently threatened by the presence and spread of the disease Huanglongbing (HLB), which was initially reported in China but is now in most of the major citrus growing countries in the world (Bove 2006). The devastating consequences of this disease and high cost of control have made it the focus of an extraordinarily broad research effort.

The lack of HLB resistance in major citrus scion types has compelled interest in identifying transgenic solutions for developing HLB-resistant cultivars. Ideally, commercial

Received 6 July 2011. Accepted 19 September 2011. Published at www.nrcresearchpress.com/gen on 1 December 2011.

Paper handled by Associate Editor G. Scoles.

W.R. Belknap,* D.R. Rockhold, and Y.Q. Gu. USDA-ARS, Western Regional Research Center, Albany, CA 94710, USA.
Y. Wang,* N. Huo, and J. Wu. USDA-ARS, Western Regional Research Center, Albany, CA 94710, USA; Department of Plant Sciences, University of California, Davis, CA 95616, USA.
E. Stover. USDA-ARS, U.S. Horticultural Research Laboratory, Fort Pierce, FL 34945, USA.

Corresponding authors: William R. Belknap (e-mail: william.belknap@ars.usda.gov) and Yong Gu (e-mail: Yong.Gu@ars.usda.gov).

*These authors contributed equally to this publication.

HLB-resistant cultivars would be constructed in a manner that limits introduced DNA to that found within the sexual compatibility group (i.e., intragenic (Rommens et al. 2007) or cisgenic constructs to enhance consumer acceptance (Jacobsen and Schouten 2007). The application of these strategies to HLB-resistance in citrus will require a detailed understanding of genomic sequence from diverse citrus genotypes, to provide the sequence data for construct development.

The diversity of citrus phenotypes, reproductive compatibility between related genera and the extended history of citrus cultivation have combined to complicate analysis of citrus taxonomy and phylogeny. Although past classification systems recognized numerous citrus species (Swingle 1943; Tanaka 1977), more recent evidence (Barkley et al. 2006) suggests an extremely limited number of natural citrus species. The citrus cultivar Carrizo (Savage and Gardner 1965) is the single most important rootstock to the US citrus industry and is widely employed worldwide. This diploid cultivar was generated by a 'Washington' navel orange × *Poncirus trifoliata* cross. The genus *Poncirus* is closely related to *Citrus*, so F_1 hybrids and later generations are relatively easily produced; some recent treatments (Zhang and Mabberley 2008) include *Poncirus* within *Citrus*. Carrizo shows resistance or tolerance to diseases such as citrus tristeza virus and foot rot. In addition, this cultivar has recently been demonstrated to have considerable resistance to HLB (Folimonova et al. 2009), viewed as the most serious current threat to citrus production globally.

Characterization of the citrus genome is essential to the efficient application of current crop improvement methodologies to this commodity. Towards this effort, the citrus genome genetic (Cai et al. 1994; Carlos de Oliveira et al. 2007; Chen et al. 2008; Bernet et al. 2010) and physical (Bernet and Asins 2003; Luo and Dvorak 2011; Moraes et al. 2008) maps are being developed. Recently, the draft genome assemblies of the haploid Clementine mandarin (*Citrus clementina*) and sweet orange (*Citrus sinensis*) were released for the user community (http://www.citrusgenomedb.org/gb/gbrowse/citrus_clementina_v0.9/, http://www.citrusgenomedb.org/species/sinensis/genome1.0), with certain restrictions on genomewide data analysis and publication. The potential applications of genomic data in breeding and crop improvement efforts make acquisition of genome sequences imperative (Talon and Gmitter 2008). In this work, we report the acquisition and analysis of a 3.5× citrus Carrizo genome sequence. These data will provide a foundation for studying genetics/genomics of rootstock characteristics to facilitate efficient screening and selection of improved rootstock cultivars using marker-assisted selection. A Carrizo genomic sequence database (USDA Public Citrus Genome Database, http://citrus.pw.usda.gov/) has been established to allow freely open access to this data.

Materials and methods

Carrizo genotype

Three plants of Carrizo maintained in the U.S. Horticultural Research Laboratory, Fort Pierce, Florida, greenhouse supplied the leaf material for the genomic DNA preparation. Carrizo is highly apomictic (Roose and Traugh 1988) and seedlings that are not rogued because of obvious abnormality are almost always genetically and phenotypically indistinguishable from the seed source tree. The identity of the source trees as Carrizo was confirmed by simple sequence repeat (SSR) analysis (Barkley et al. 2006).

Roche 454 sequencing

Preparation and sequencing of the 454 sequencing library was performed according to the manufacturer's instructions (GS FLX Titanium General Library preparation kit/emPCR kit/sequencing kit, Roche Diagnostics, http://www.roche.com). In brief, 10 μg of citrus genomic DNA were sheared by nebulization and fractionated on agarose gel to isolate 400–750 base fragments. These were used to construct a single-stranded shotgun library that was used as a template for single-molecule PCR. The amplified template beads were recovered after emulsion breaking and selective enrichment. The Genome Sequencer FLX Titanium flows 200 cycles of four solutions containing either dTTP, αSdATP, dCTP, and dGTP reagents, in that order, over the cell.

Repeat DNA analysis

The repetitive sequences in the Carrizo 454 reads were identified and masked using the RepeatMasker program (http://repeatmasker.org/). Repeat database version 20090604 was used with the parameter defined as Eukaryote species. For identification of unique citrus unique repeats, 454 reads representing 10% of the Carrizo genome were extracted from the known-repeat masked sequences. These 454 reads were further scanned using RepeatScout (Price et al. 2005) to identify de novo repeats. A fasta file that contained all the repetitive elements that RepeatScout could find was generated. The novel repeats were used as a custom library for RepeatMasker to further mask repetitive sequences in the 454 reads.

Putative citrus MITE sequences were identified in the de novo repeat data set by identification of terminal inverted repeated domains using Pustell matrix analysis (Pustell and Kafatos 1982, 1986) to compare individual DNA sequences (MacVector11.1). Full-length MITEs were identified by BLAST searches of the NCBI EST database, HarvEST (http://harvest-blast.org/) and USDA Public Citrus Genome Database. The actual boundaries of the repeated domains described here were determined by direct comparison of related repeats from multiple loci (EST and genomic sources) and identifying points of divergence. Putative recombinational target sequences were identified employing MITE-flanking sequences to BLAST probe available citrus databases.

Ty3/*gypsy* elements were identified using a TBLASTN search of the Carrizo database employing the *Arabidopsis ATHILA* polyprotein as a probe. Reads immediately flanking the original Carrizo polyprotein sequences were identified by BLASTN searches of the database employing original Carrizo-read 150-bp terminal domains (e value cut off of 1×10^{-65}). Sequential BLASTN searches were used to complete the Ty3 elements. Long terminal repeat (LTR) 5′ and 3′ ends were identified by using sequential 400-bp segments of the LTR domain as BLASTN probes and identifying those that returned BLAST results in which approximately half the hits aligned to a portion of the probe. These BLASTN hits were then characterized to identify junctions of LTR and nonrepetitive DNA sequences. The structure of each of the

Table 1. Occurrence and distribution of known repetitive DNA sequences in the Carrizo 454 sequence read database.

	No. of elements	Length occupied (bp)	Percentage of the genome (%)
Class I retrotransposon	302 568	75 105 042	5.94
LTR retrotransposon	292 085	73 154 971	5.79
Ty1/*copia*	141 952	35 659 581	2.82
Ty3/*gypsy*	146 015	37 225 461	2.95
Non-LTR retroelement	10 483	1 950 071	0.15
SINEs	7	287	0.00
LINEs	10 476	1 949 784	0.15
Class II DNA transposons	27 244	4 990 094	0.39
hobo-Activator	7497	1 306 924	0.10
Tc1-IS630-Pogo	1969	298 704	0.02
En-spm	5702	1 332 862	0.11
MuDR-IS905	9163	1 619 130	0.13
PiggyBac	0	0	0.00
Tourist/Harbinger	1894	332 392	0.03
Other (Mirage, P-element)	0	0	0.00
Rolling-circles (RC/Helitron)	0	0	0.00
Unclassified	267	18 736	0.00
Total interspersed repeats		80 113 872	6.34
Small RNA	49 115	16 353 398	1.29
Satellites	19	2283	0.00
Simple repeats	8569	819 185	0.06
Low complexity	7	808	0.00
Total repetitive DNA		97 289 546	7.70

three Carrizo Ty3 elements constructed in this manner was verified by constructing the same elements using an identical procedure employing the DOE Joint Genome Institute (JGI) sweet orange database (HarvEST, http://harvest-blast.org/).

CitPhlo2 gene assembly and amplification

The CitPhlo2 protein was identified by characterization of citrus proteins encoded by ESTs over-represented in phloem/bark-derived libraries in the NCBI database as previously described (McCue et al. 2007). Approximately full-length ESTs were employed to generate the CitPhlo2 protein sequence. CitPhlo2-encoding Carrizo sequences were identified using a TBLASTN search of the Carrizo 454 reads. Sequential BLASTN searches employing Carrizo-read terminal domains were then employed to assemble a putative Carrizo *CitPhlo2* gene sequence as described for the Ty3 elements above.

Genomic DNA from the *Citrus sinensis* cultivar Olinda was used in a PCR amplification of the *CitPhlo2* gene. The amplification reaction employed the proofreading polymerase Phusion (New England Biolabs, Ipswich, Massachusetts, USA) and forward (5′-GCCATTTCATGGCTAGAGTGA-3′) and reverse (5′-AACATTTGTGTTGTAACGCTGTA-3′) primers.

Results

454 sequencing

Genomic DNA was prepared from Carrizo leaves as described by Peterson et al. (1997). The identity of the source material was verified by SSR analysis (Barkley et al. 2006). The DNA was randomly sheared and sequenced using the 454 GS FLX shotgun sequencing method. A total of 3 455 302 shotgun reads were generated and used for analysis. The mean read length was 365 bp and the genomic raw sequence produced was 1264.4 Mb, which corresponds to almost 3.5× of the citrus genome based on the estimated citrus genome size of 367 Mb per 1C genome (Arumuganathan and Earle 1991).

Repetitive DNA composition and content in the Carrizo genome

The total transposable element (TE) content was estimated in two steps. First, the 454 shotgun reads were compared with the Eukaryote part of the RepeatMasker database. Based on similarity searches of the known repeat database, 7.7% of the nucleotides in the citrus shotgun reads were identified as belonging to known repeats (Table 1). The most common repeat families identified using known repeats were LTR retrotransposons (5.94%), which primarily includes Ty1/*copia* (2.82%) and Ty3/*gypsy* (2.95%) elements. Although this analysis appeared consistent with previously reported repetitive DNA content (Terol et al. 2008), it represented a relatively low content as compared with many plant genomes even given the relative small size of the citrus genome (Arumuganathan and Earle 1991).

The observed low percentage of known repetitive DNA contents in the 454 reads appeared likely to be due to poor representation of Carrizo repetitive DNA sequences in the known repeat database. To more completely identify repetitive sequences within the database, a de novo citrus unique repeat database library was constructed using RepeatScout software (Price et al. 2005). The resulting library contained a total of 2592 unique citrus repetitive sequences (identified

Table 2. High abundance de novo repetitive DNA in Carrizo.

ID		No. of reads	Length occupied (bp)	Percentage of the genome (%)
$R = 506$	Satellite[a]	54 224	4 835 360	0.38
$R = 0$	Ty1/*copia LTR*[b]	17 298	5 645 910	0.44
$R = 1$	Ty1/*copia* LTR[b]	15 294	4 686 150	0.37
$R = 31$	Ty3/*gypsy gag*[c]	5883	2 041 494	0.16
$R = 10$	Ty3/*gypsy gag*[c]	5640	2 038 678	0.16
$R = 16$	Ty3/*gypsy* RP[c]	5233	1 802 643	0.14
$R = 23$	Ty3/*gypsy gag*[c]	5144	1 534 833	0.12
$R = 2$	Ty1/*copia* Integrase[b]	4799	1 118 458	0.08
$R = 27$	Ty3/*gypsy LOG*[c]	4582	1 554 802	0.12

[a](Fann et al. 2001).
[b](Yang et al. 2003).
[c]Ty3-3.

Fig. 1. Alignment of Carrizo MITE-containing sequences with recombinational target sequences. Putative citrus MITE sequences were identified in the de novo repeat dataset by identification of terminal inverted repeated domains using DNA Pustell analysis as described in the Materials and methods. Complete MITE domains for two separate elements (A, Cit-MITE1; B, Cit-MITE2) were identified in the NCBI EST database (A, NCBI accession CX636022) and the HarvEST database of citrus JGI genomic sequences (B, http://harvest-blast.org/, accession BUZY200111). Recombinational target sequences in the two 454 reads, GI30BAM01A3JZE and GI1478W01BMX8G, were identified by BLASTN search of the Carrizo 454 database (http://citrus.pw.usda.gov/). Positions in individual NCBI, HarvEST , and USDA Public Citrus Genome Database accession numbers are indicated and target site duplications are underlined.

A

```
CX636022  (298) TTTGGGATGACCGTACGAACATAATAAATAAGCAAGTAACCGAACGAGTTCAAGGACCTGTCTATGTTACAGATTCTGTAACAAACATCTCCCGTCATCCCACCATCCGATTT
GI30BAM01A3JZE  TTTGGGATGACCGTACGAACATAATAAATAAGCAAGTAACCGAACGAGTTCAA (53)

CX636022        TCTTTTGTTTAGTTTCTCACTCCCCATCCAACCACATCCGACGGCTGATGCTGCAGTCTGTACCTTACAGAATCTGTAACGTAGTTTTGTCCCGAGTTCAAATGCCAAAAATG
GI30BAM01A3JZE                                                                                                (45) CGAGTTCAAATGCCAAAAATG

CX636022        GAAACCCATAAAAACAATAAATAAAAATAAATTATTTAACAAAAGAGATCGATCGAACAGTAGTTTTATCTTTTTGAATA (603)
GI30BAM01A3JZE  GAACCCCATAAAAACAATAAATAAAAATAAATTATTTAACAAAAGAGATCGATCGAACAGTAGTTTTATCTTTTTGAATA (145)
```

B

```
BUZY200111 (26) GTTGAGGAAGCGTGCAGCATTAATTAAATGGGTAATTTTTAAAAACTTACCCTGAGGTTTGGACTTGTTGCAAGTAGATGACGAGAATTTATTTATTTGTAAAAAATCCCCTA
GI1478W01BMX8G  GTTGAGGAAGCGTGCAGCATTAATTAAAT (195)

BUZY200111      CCGTCAGTTAACTTTAACATTGACCGTTAGTTGACTGTGCAAAGATAATATTACCCTTAACAATAATTTGTAGACGGAAATAACTAAAAAAAAATAAAATTATTGGCTATTTT
GI1478W01BMX8G

BUZY200111      ATCCTCTTTAAATTATTGATATTTTCTTAAAGTTCCTACAAGAAAAATAAAATTAAAAATTTTAAAAATCCTCTTTAAATTATTGATATTTTCTTAAAGTTTCTACAAAAAAG
GI1478W01BMX8G

BUZY200111      ATAAAATTAAAAACTTGAAAATGAAATAGTTATAATCTTTACCTATGATATTGTAAATCTTTGGAATTGACAAGGATAAAATTGTCAAAAAATAGGGTTTGTGCTTTTCGCGC
GI1478W01BMX8G

BUZY200111      CAATAATTTTAAATTTTTTTTTAGTTATTTCGGTCTACAAATTGTTGTTAAGGGTAATATTGTCTTTGCACAGTCAACTAACGGTTAATGTTAAAGTTAACTGATGGTAGAGG
GI1478W01BMX8G

BUZY200111      GTGTTTTACAAATAAATAAATTCTCACCATTTACTTGTAACAAGCACGAACCTCAGGGGAGGTTTTTAAAAATTACCCTTAATTAAATACGAAACACGGCGCTAGAGTTGAAA
GI1478W01BMX8G                                                                                                (186) TTAATTAAATACGAAACATGGCGCTGGAGTTGAAA


BUZY200111      ATTTTTCTAGTTGTGGA
GI1478W01BMX8G  ATTTTTCTAGTTGTGGA
```

using BLAST e value 1×10^{-15} to mask the repetitive sequences). Comparison of this library with the Carrizo database indicated that 955 814 reads, not identified by similarity to known repeats, produced significant matches with the de novo citrus unique repeat sequences. Thus, the Carrizo unique repeat DNA represents 28.8% of the genomic sequences in the 454 reads. Combining with the 7.7% of known repeats, the total repetitive DNA in the Carrizo genome was approximately 36.5%. The complete de novo citrus unique repeat database is available (http://citrus.pw.usda.gov/). The analysis of the highly representative de novo repeats in the reads is shown in Table 2.

The de novo repeat data set was also employed to identify common miniature inverted repeat TEs (MITEs, (Wessler 1998)), as indicated in Fig. 1. The Cit-MITE1 element (Fig. 1A) is 152 bp in length and defined by 28 bp terminal inverted repeats. This element is flanked by 9 bp target site duplications. BLAST searching of the Carrizo database indicated Cit-MITE1 within approximately 4000 reads, with MITE sequences representing approximately 0.6 Mb (0.05%) of the Carrizo genome. The Cit-MITE2 element (Fig. 1B) is 614 bp in length and defined by 191 bp terminal inverted repeats. This element is flanked by 10 bp target site duplications. BLAST searching of the Carrizo database indicated Cit-MITE2 within approximately 6000 reads, with MITE sequences representing approximately 1.5 Mb (0.1%) of the Carrizo genome. A search of the NCBI EST database reveals that the Cit-MITE1 element is found within a large number of ESTs derived from a variety of genes from both sweet orange and *Poncirus*. In contrast, the Cit-MITE2 element is

Fig. 2. Representation of full length Ty3/*gypsy*-like retrotransposons from Carrizo and *Populus trichocarpa*. LTR regions are indicated by grey boxes, coding sequences are indicated by black arrows, Carrizo Ty3-1 and Ty3-2 retrotransposons (A and B, respectively) reconstructed from 454 reads and polyprotein-coding regions defined as described in the Materials and methods. (B) Carrizo Ty3-2 contains two additional ORFs, the *rbp* ORF encodes a viral RNA-binding-like protein, the *log* ORF encodes a Lonely Guy-like protein. (C) The poplar retrotransposon was identified in locus AC216453 (poplar chromosome 4, reverse complement of 127 259 to 139 470) by a BLAST search of the NCBI database. Retrotransposon-coding domains and coding sequences identified as described above.

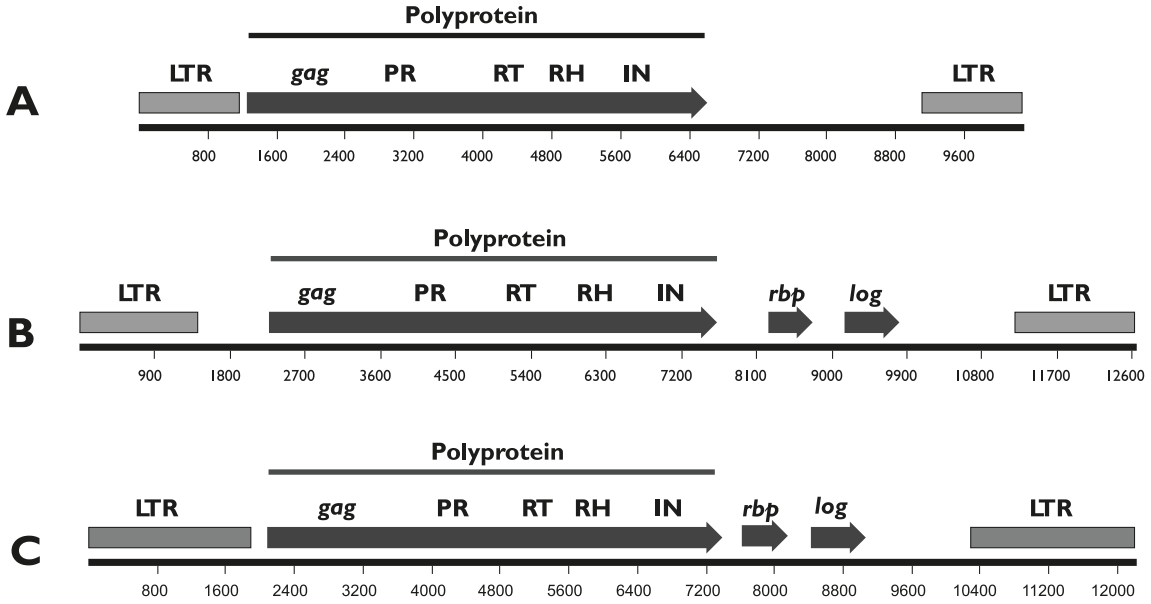


Table 3. Identification of Carrizo retrotransposon Ty3-2-coding regions.

Coding region	Locus	Protein	Expect
gag	ABD63142	Asparagus *gag*	9×10^{-55}
RP	ABD63142	Asparagus retropepsin	6×10^{-48}
RT	ABD63142	Asparagus reverse transcriptase	3×10^{-69}
RH	ABD63142	Asparagus RNaseH	4×10^{-51}
IN	ABD63142	Asparagus integrase	2×10^{-120}
rbp	AAL76174	Grapevine Virus A RNA binding protein	0.9
log	NP_001043439	Rice Lonely Guy	3×10^{-47}
	NP_565668	*Arabidopsis* LOG1	9×10^{-49}

observed only in three citrus ESTs from sweet orange and *Poncirus* (data not shown). It is not clear if frequent association of Cit-MITE1 with gene sequences have any biological relevance regarding gene evolution and expression.

Characterization of Carrizo Ty3/*gypsy* retrotransposons

Our repeat DNA analysis indicated that the TEs demonstrating the highest genome percentages are a Ty3 element (Ty3-2, described below) and a previously identified Ty1 element (Yang et al. 2003), which account for 4.0% and 1.4% of the total Carrizo reads, respectively. The Ty3/*gypsy* class of retrotransposons are common features of plant genomes (Kumar and Bennetzen 1999), including citrus (Bernet and Asins 2003). Two Ty3/*gypsy* elements were identified in the Carrizo database employing the *Arabidopsis ATHILA* polyprotein as a probe, and full length Carrizo retrotransposons were assembled from the Carrizo database by sequential BLASTN searches employing Carrizo-read terminal domains. The first Carrizo element (Carrizo Ty3-1), shown schematically in Fig. 2A, is 10.2 kb in length and is flanked by 1.2 kb LTRs. This element contains a single polyprotein-encoding ORF similar to other Ty3 elements. The second Carrizo element (Carrizo Ty3-2, Fig. 2B) is 12.6 kb in length and is flanked by 1.4 kb LTRs. These two elements share very little DNA sequence similarity (Supplementary data,[1] Fig. S1). Assignment of coding region domains of Carrizo Ty3-2 was established by BLASTP search of the NCBI protein database (Table 3) using partial (polyprotein) or complete (*rbp* and *log*) peptide sequences as probes.

As indicated in Fig. 2, Carrizo Ty3-2 contains two ORFs not found in elements with standard Ty3/*gypsy* architecture (Carrizo Ty3-1). The first ORF (*rbp*) encodes a 184 residue protein weakly similar to ORF5 of Grapevine Virus A (Minafra et al. 1997) Fig. S2). This putative RNA-binding protein has been identified as an RNA-silencing suppressor (Zhou et al. 2006). The second additional Carrizo Ty3-2 ORF (224 residues, *log*) shares significant identity with Lonely Guy (LOG)-like proteins from a variety of plant species, including

[1]Supplementary data are available with the article through the journal Web site (http://nrcresearchpress.com/doi/suppl/10.1139/g11-070).

Fig. 3. Alignment of the endogenous and retrotransposon-encoded LOG proteins. Protein sequences from Carrizo Ty3-2 and poplar Ty3 log proteins were aligned with rice, *Arabidopsis*, citrus, and poplar LOG proteins using the ClustalW program. Sequences representing conserved amino acids among these compared LOG proteins are shaded with grey.

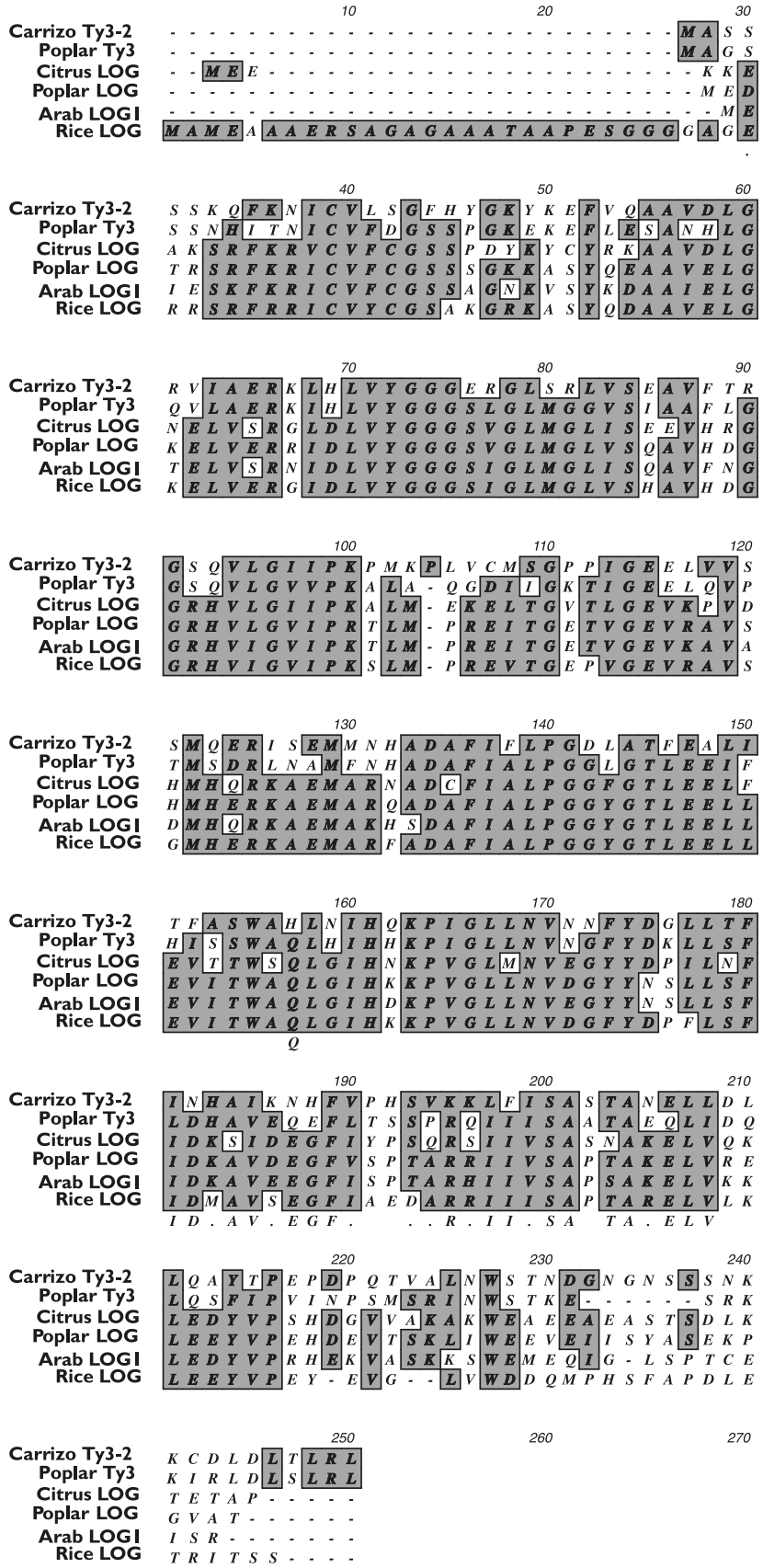
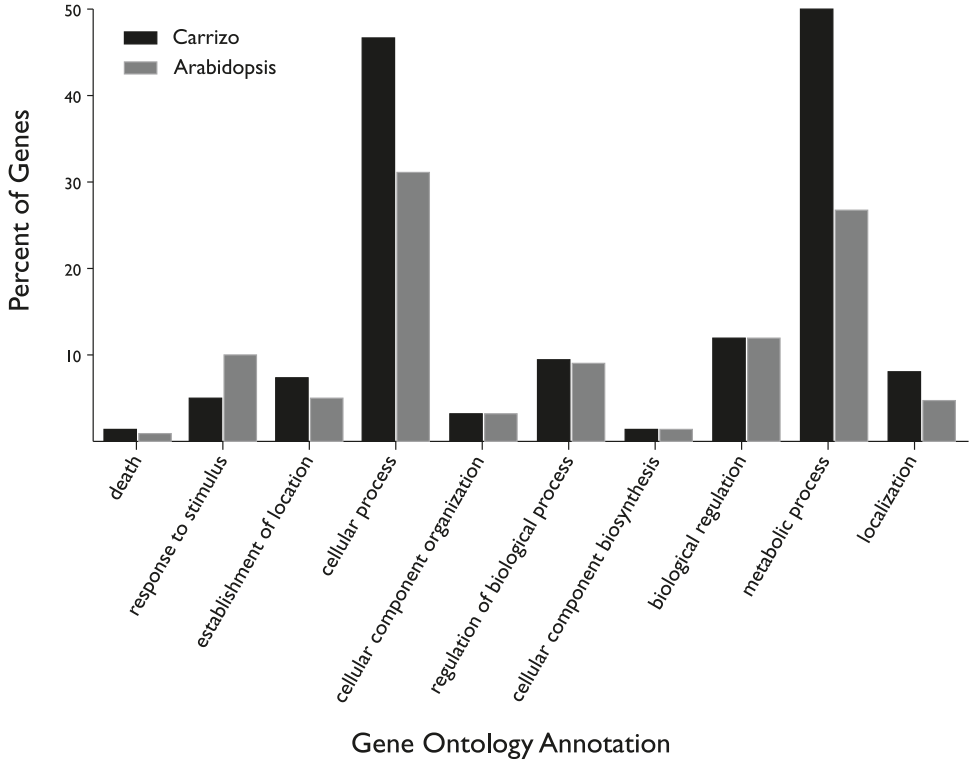


Fig. 4. Blast2Go gene ontology analysis of Carrizo and *Arabidopsis*. Functional annotation of 18 698 proteins of a total of 31 186 identified employing assembled Carrizo contigs and singletons as described in text.



rice LOG and *Arabidopsis* LOG1 (Kurakawa et al. 2007; Kuroha et al. 2009) (Table 3; Fig. 3). The LOG proteins have cytokinin riboside 5′-monophosphate phosphoribohydrolase activity involved in meristem maintenance. However, alignment of the Carrizo Ty3-2 *log*-encoded protein to the Conserved Domain Database (NCBI) (Marchler-Bauer et al. 2009) indicates inclusion in the DNA Processing A Superfamily. This superfamily includes the single stranded DNA-binding protein dprA and the SMF bacterial protein involved in RecA-dependent homologous recombination (Mortier-Barrière et al. 2007). In contrast to the endogenous flowering plant *LOG*-like genes that have six conserved introns, the *log* domain in Carrizo Ty3-2 has no introns and appears to be a retrotransposed cDNA from a processed LOG-like mRNA.

Given the unusual structure of the Carrizo Ty3-2 element, available plant genomes (*Arabidopsis*, grape, poplar, rice, *Brachypodium*) were evaluated for the presence of similar retrotransposons employing the *log*-encoded protein as a probe. Although *LOG*-like genes were easily detected in all genomes, only the poplar genome (Tuskan et al. 2006) contained a Ty3/*gypsy* element with similar structure (Fig. 2C). Although DNA sequence similarity between the two elements is limited to the protein-coding domains, the structural and protein sequence similarity are apparent (Fig. 3; Fig. S3). It seems that the two retrotransposon-encoded LOG proteins have slightly higher sequence conservation as compared with the endogenous LOG proteins (Fig. 3).

Coding region analysis

The complete pool of shotgun reads was filtered to remove repetitive DNA sequences as described above, and the remaining 2 275 916 nonrepetitive reads were examined for citrus EST homology. Similarity searches were performed with BLASTN (Altschul et al. 1990) against the citrus EST sequences (as downloaded from the plant genome database). Comparison of the 454 reads with the citrus EST database revealed that 24.4% of the reads contained "genic" regions using BLASTN (e value cut off of 1×10^{-15}). The 2 275 916 nonrepetitive reads were also employed for sequence assembly using the Roche GS De novo Assembler software. This resulted in 191 153 contigs with a mean contig length of 914 bp and the longest contig of 5778 bp. When the contigs and singletons were searched against the nonredundant protein database, a total of 31 186 significant protein hits were identified at 1×10^{-10}. Blast2Go gene ontology analysis was performed for functional annotation of these matched proteins. Among these 31 861 proteins, 18 698 were assigned to GO categories. In general, the resulting GO category assignment of these proteins was similar to that found in *Arabidopsis* (Fig. 4), with the exception of the cellular and metabolic process categories. The complete Carrizo gene annotation database is available (http://citrus.pw.usda.gov/).

One of the primary incentives for acquisition of the Carrizo genomic sequence was to act as a source for transcriptional control element sequences for application to specific citrus improvement projects (Rommens et al. 2007). For example, a phloem-specific promoter would be desirable for expression of introduced HLB-resistance genes (Graham and Timmer 2000; Folimonova and Achor 2010). The CitPhlo2 protein was identified by characterization of citrus proteins encoded by ESTs over-represented in phloem/bark-derived libraries in the NCBI database (McCue et al. 2007). CitPhlo2-

Fig. 5. Representation of putative phloem-specific gene *CitPhlo2* assembled from Carrizo 454 reads and PCR amplification from citrus genomic DNA. (A) Positions of 454 individual reads are indicated by grey boxes. The protein-coding domain and transcribed region (as derived by similarity to citrus EST NCBI locus EY769260) are indicated by the black box and grey arrow, respectively. Positions of the source sequences for PCR, the amplification primer pair, are indicated by the black arrows. (B) PCR amplification products from *Citrus sinensis* genomic DNA using primers from positions indicated in (A).

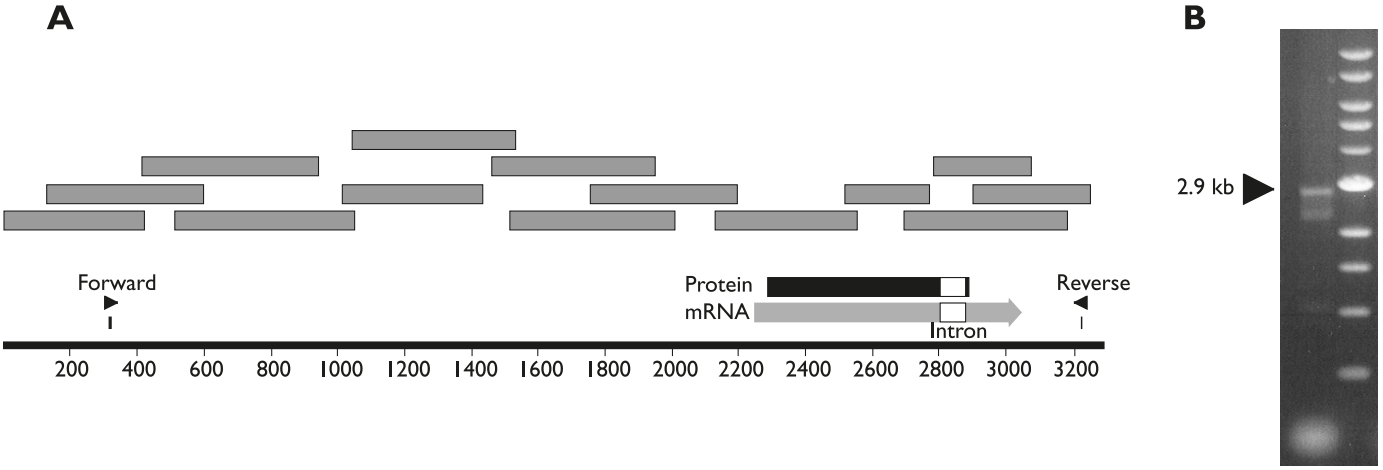
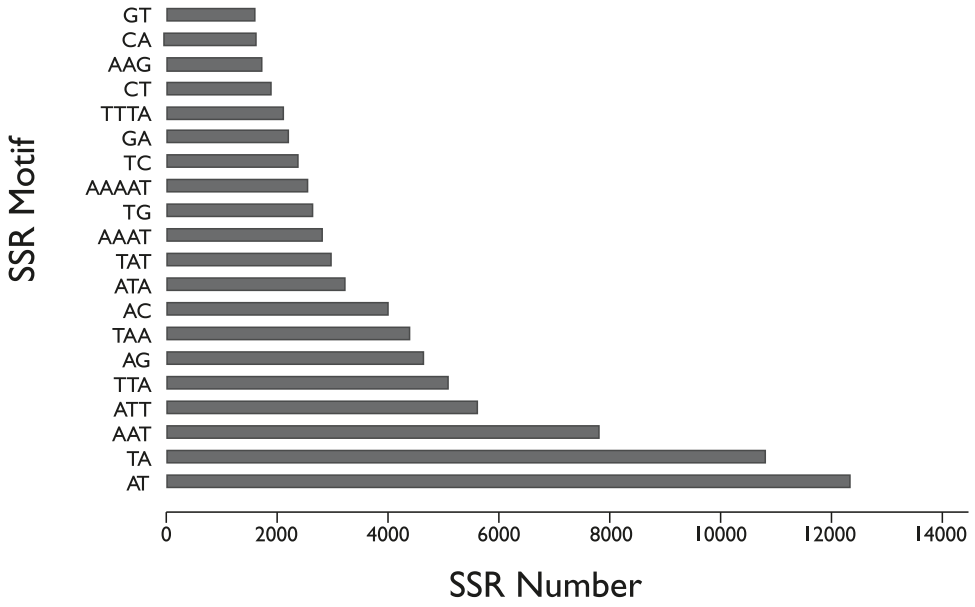


Fig. 6. Number of repeats of the top 20 abundant SSRs in shotgun reads from the Carrizo 3.5× sequence database. SSRs were identified using SSR Perl Script (ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl).



encoding Carrizo sequences were then identified by BLAST search (TBLASTN) of the Carrizo 454 reads (Fig. 5A). Sequential BLASTN searches employing Carrizo-read terminal domains were then employed to assemble a putative Carrizo *CitPhlo2* gene sequence (Fig. 5A).

The accuracy of the assembly represented in Fig. 5A was evaluated by PCR amplification from *C. sinensis* genomic DNA as described in the Materials and methods. The expected 2.9 kb amplification product was obtained (Fig. 5B). Sequence analysis of the amplification product revealed >99.5% identity to the assembly sequence (data not shown). This divergence from the assembled sequence may reflect the expected *Poncirus/Citrus* heterozygosity in the database. These data indicate that the 3.5× Carrizo genome coverage described here is sufficient for accurate assembly of selected citrus genes for isolation of putative transcriptional control sequences.

SSR analysis

SSR Perl Script was used to identify a total of 168 273 SSRs longer than 15 bp in the assembled sequence reads. The occurrence of SSRs in the Carrizo genome had a frequency of 0.14 SSR per kilo base pair (kb), a value almost identical to that reported in a study based on citrus BAC sequences (0.19 SSR per kb) (Terol et al. 2008; Table S1).

In the Carrizo SSR set, there were 29 095 class I (more than 10 repeats) and 13 9178 class II (less than 10 repeats) SSRs. In general, those motifs containing A/T nucleotides were far more abundant than G/C rich repeats, especially ATT/TAA and AT/TA tri- and dinucleotides (Fig. 6). The

SSRs identified here represent useful molecular markers for citrus research. Both the SSR marker sequences and the flanking primer sequences for SSR detection are available at http://citrus.pw.usda.gov/.

Discussion

We report here acquisition and analysis of a 3.5× draft sequence of the citrus Carrizo genome. Approximately 36.5% of the reads in the database constitute repetitive DNA sequences, representing both class I and class II TEs. Of the nonrepetitive reads, approximately 24% of the reads contained genic regions, identified by comparison with the available protein and citrus EST databases.

The unusual Ty3/*gypsy* elements from Carrizo and poplar (Figs. 2B and 2C) represent the first reported that contain *rbp*- and *log*-like genes. Although a number of plant retrotransposons encoding envelope (env) proteins similar to those required for retrovirus infectivity have been identified (Vicient et al. 2001), retrotransposons with viral *rbp* genes with the potential to affect element transcriptional silencing (Lippman and Martienssen 2004) have not. Perhaps more surprising is the presence of a retrotransposon-associated *LOG*-like gene, conserved in both citrus and poplar. This feature suggests functions of members of the plant LOG-like protein family in recombinational pathways, in addition to phytohormone biosynthesis (Kurakawa et al. 2007; Kuroha et al. 2009). Despite high sequence conservation between the endogenous and the retrotransposon-encoded LOG proteins, it is not clear if the latter proteins are expressed and possess any biological function. Annotated sequences of the two Carrizo Ty3/*gypsy* elements presented here are available (http://citrus.pw.usda.gov/).

Plant breeding is a time consuming and labor intensive process. Characterization of new hybrids, to select for desirable traits, is perhaps the most expensive and laborious component of the plant improvement cycle. Molecular markers have wide application in crop improvement, including genotyping, genetic mapping and gene cloning, and marker-assisted selection. Use of marker-assisted selection can reduce a process that might require many years to one that can be conducted in a lab in just a few hours, for a few select traits, reducing the numbers of progenies subjected to more laborious and lengthy analysis.

SSRs, in which the same pattern of two or more bases is repeated multiple times in consecutive nucleotide sequence, are an important class of DNA markers. They are widespread within eukaryotic genomes, permit acquisition of mutations because they are often in noncoding regions or do not influence gene product function, and develop variants at a high rate because their repetitive nature results in greater replication error. SSRs' relatively high degree of variability at any individual locus and ease of assessing alleles in a co-dominant manner makes them especially suited to use as markers for selection of closely linked traits and pedigree analysis (Song 1999). A total of 168 273 genome-wide SSR markers have been generated through mining the assembled citrus sequence reads. The sequence data also represent a useful genomics resource for developing other molecular markers such as single nucleotide polymorphism (SNP) markers for citrus research (You et al. 2011).

With the development of high-throughput sequencing technologies, sequence assembly and characterization of complex plant genomes through genome shotgun sequencing is becoming feasible. Previously, only limited citrus genomics resources were available, mainly derived from Sanger-based sequencing technologies (Roose et al. 2007; Terol et al. 2008; Yang et al. 2003). Recently, the raw sequence data of the haploid Clementine mandarin (*C. clementina*) and sweet orange (*C. sinensis*) genomes have been generated primarily using next generation and Sanger sequencing technologies and are accessible at the phytozome Web site (www.phytozome.org). The potential applications of these data in breeding and crop improvement efforts make acquisition of this sequence imperative (Talon and Gmitter 2008). The citrus genome genetic (Cai et al. 1994; Carlos de Oliveira et al. 2007; Chen et al. 2008; Bernet et al. 2010) and physical (Bernet and Asins 2003; Luo and Dvorak 2011; Moraes et al. 2008) maps will certainly help accuracy assembly of the citrus genomes by anchoring and ordering sequence contigs or scaffolds onto citrus chromosomes. In addition, genome sequencing of closely related species is important to understand the evolutionary processes of speciation and domestication. In this work, we generated a 3.5× draft genome sequence of the citrus cultivar Carrizo, the most important rootstock in US citrus industry. Both the data and analytical tools described here are available on the USDA Public Citrus Genome Database (http://citrus.pw.usda.gov/) Web site. Included in this database are the complete set of 454 reads with BLAST server, the library of unique citrus repetitive sequences, and a catalog of assembled TEs and citrus SSRs. This database was established specifically to allow open access to the citrus research community, with no limitations on use or dissemination of the information.

Acknowledgements

The authors thank M. Roose and C. Federici for SSR analysis of sequencing library source DNA, K. Bowman for preparing Carrizo leaf tissue, and G.R. Lazo for assistance in bioinformatics

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**(3): 403–410. PMID:2231712.

Arumuganathan, K., and Earle, D.E. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Reporter, **9**(3): 208–218. doi:10.1007/BF02672069.

Barkley, N.A., Roose, M.L., Krueger, R.R., and Federici, C.T. 2006. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theor. Appl. Genet. **112**(8): 1519–1531. doi:10.1007/s00122-006-0255-9. PMID:16699791.

Bernet, G.P., and Asins, M.J. 2003. Identification and genomic distribution of *gypsy* like retrotransposons in *Citrus* and *Poncirus*. Theor. Appl. Genet. **108**(1): 121–130. doi:10.1007/s00122-003-1382-1. PMID:12937896.

Bernet, G.P., Fernandez-Ribacoba, J., Carbonell, E.A., and Asins, M.J. 2010. Comparative genome-wide segregation analysis and map construction using a reciprocal cross design to facilitate citrus germplasm utilization. Mol. Breed. **25**(4): 659–673. doi:10.1007/s11032-009-9363-y.

Bove, J.M. 2006. Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. J. Plant Pathol. **88**(1): 7–37.

Cai, Q., Guy, C., and Moore, G.A. 1994. Extension of the linkage map in *Citrus* using random amplified polymorphic DNA (RAPD) markers and RFLP mapping of cold-acclimation-responsive loci. Theor. Appl. Genet. **89**(5): 606–614. doi:10.1007/BF00222455.

Carlos de Oliveira, A., Bastianel, M., Cristofani-Yaly, M., Morais do Amaral, A., and Machado, M.A. 2007. Development of genetic maps of the citrus varieties 'Murcott' tangor and 'Pêra' sweet orange by using fluorescent AFLP markers. J. Appl. Genet. **48**(3): 219–231. doi:10.1007/BF03195216. PMID:17666774.

Chen, C., Bowman, K.D., Choi, Y.A., Dang, P.M., Rao, M.N., Huang, S., et al. 2008. EST-SSR genetic maps for *Citrus sinensis* and *Poncirus trifoliata*. Tree Genet. Genomes, **4**(1): 1–10. doi:10.1007/s11295-007-0083-3.

Fann, J.-Y., Kovarik, A., Hemleben, V., Tsirekidze, N.I., and Beridze, T.G. 2001. Molecular and structural evolution of *Citrus* satellite DNA. Theor. Appl. Genet. **103**(6–7): 1068–1073. doi:10.1007/s001220100719.

FAO. 2010. Food and Agricultural Organization of the United Nations: FAOSTAT [online]. Available from http://faostat.fao.org/default.aspx [accessed 13 September 2011].

Folimonova, S.Y., and Achor, D.S. 2010. Early events of citrus greening (Huanglongbing) disease development at the ultrastructural level. Phytopathology, **100**(9): 949–958. doi:10.1094/PHYTO-100-9-0949. PMID:20701493.

Folimonova, S.Y., Robertson, C.J., Garnsey, S.M., Gowda, S., and Dawson, W.O. 2009. Examination of the responses of different genotypes of citrus to Huanglongbing (citrus greening) under different conditions. Phytopathology, **99**(12): 1346–1354. doi:10.1094/PHYTO-99-12-1346. PMID:19900000.

Graham, J.H., and Timmer, L.W. (*Editors*). 2000. Compendium of citrus diseases. 2nd ed. APS Press, St. Paul, Minn.

Jacobsen, E., and Schouten, H.J. 2007. Cisgenesis strongly improves introgression breeding and induced translocation breeding of plants. Trends Biotechnol. **25**(5): 219–223. doi:10.1016/j.tibtech.2007.03.008. PMID:17383037.

Kumar, A., and Bennetzen, J.L. 1999. Plant retrotransposons. Annu. Rev. Genet. **33**(1): 479–532. doi:10.1146/annurev.genet.33.1.479. PMID:10690416.

Kurakawa, T., Ueda, N., Maekawa, M., Kobayashi, K., Kojima, M., Nagato, Y., et al. 2007. Direct control of shoot meristem activity by a cytokinin-activating enzyme. Nature, **445**(7128): 652–655. doi:10.1038/nature05504. PMID:17287810.

Kuroha, T., Tokunaga, H., Kojima, M., Ueda, N., Ishida, T., Nagawa, S., et al. 2009. Functional analyses of *LONELY GUY* cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. Plant Cell, **21**(10): 3152–3169. doi:10.1105/tpc.109.068676. PMID:19837870.

Lippman, Z., and Martienssen, R. 2004. The role of RNA interference in heterochromatic silencing. Nature, **431**(7006): 364–370. doi:10.1038/nature02875. PMID:15372044.

Luo, M.C., and Dvorak, J. 2011. Citrus physical mapping database [online]. Available from http://phymap.ucdavis.edu/citrus/ [accessed 13 September 2011].

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. **37**(suppl 1): D205–D210. doi:10.1093/nar/gkn845. PMID:18984618.

McCue, K.F., Allen, P.V., Shepherd, L.V., Blake, A., Maccree, M.M., Rockhold, D.R., et al. 2007. Potato glycosterol rhamnosyltransferase, the terminal step in triose side-chain biosynthesis. Phytochemistry, **68**(3): 327–334. doi:10.1016/j.phytochem.2006.10.025. PMID:17157337.

Minafra, A., Saldarelli, P., and Martelli, G.P. 1997. Grapevine virus A: nucleotide sequence, genome organization, and relationship in the *Trichovirus* genus. Arch. Virol. **142**(2): 417–423. doi:10.1007/s007050050088. PMID:9125055.

Moraes, A.P., Mirkov, T.E., and Guerra, M. 2008. Mapping the chromosomes of *Poncirus trifoliata* Raf. by BAC-FISH. Cytogenet. Genome Res. **121**(3–4): 277–281. doi:10.1159/000138897. PMID:18758171.

Mortier-Barrière, I., Velten, M., Dupaigne, P., Mirouze, N., Piétrement, O., McGovern, S., et al. 2007. A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. Cell, **130**(5): 824–836. doi:10.1016/j.cell.2007.07.038. PMID:17803906.

Peterson, D.G., Boehm, K.S., and Stack, S.M. 1997. Isolation of milligram quantities of nuclear DNA from tomato (*Lycopersicon esculentum*), a plant containing high levels of polyphenolic compounds. Plant Mol. Biol. Rep. **15**(2): 148–153. doi:10.1007/BF02812265.

PlantGDB. 2011. PlantGDB: resources for comparative plant genomics [online]. Available from http://www.plantgdb.org/ [accessed 13 September 2011].

Price, A.L., Jones, N.C., and Pevzner, P.A. 2005. De novo identification of repeat families in large genomes [online]. Available from http://bix.ucsd.edu/repeatscout/ [accessed 13 September 2011].

Pustell, J., and Kafatos, F.C. 1982. A high speed, high capacity homology matrix: zooming through SV40 and polyoma. Nucleic Acids Res. **10**(15): 4765–4782. doi:10.1093/nar/10.15.4765. PMID:6290988.

Pustell, J., and Kafatos, F.C. 1986. A convenient and adaptable microcomputer environment for DNA and protein sequence manipulation and analysis. Nucleic Acids Res. **14**(1): 479–488. doi:10.1093/nar/14.1.479. PMID:3753784.

Rommens, C.M., Haring, M.A., Swords, K., Davies, H.V., and Belknap, W.R. 2007. The intragenic approach as a new extension to traditional plant breeding. Trends Plant Sci. **12**(9): 397–403. doi:10.1016/j.tplants.2007.08.001. PMID:17692557.

Roose, M.L., and Traugh, S.N. 1988. Identifcation and performance of citrus trees on nucellar and zygotic rootstocks. J. Am. Soc. Hortic. Sci. **113**(1): 100–105.

Roose, M., Niedz, R., Gmitter, F., Close, T., Dandekar, A., and Rokhsar, D.S. 2007. Analysis of a 1.2× whole genome sequence of *Citrus sinensis*. *In* Plant & Animal Genome XV Conference, San Diego, Calif., 13–17 January 2007. pp. 81.

Savage, E.M., and Gardner, F.E. 1965. The Troyer and Carrizo citranges. Calif. Citrog. **40**: 255, 275–278.

Song, Q.J. 1999. A review of development and application of simple sequence repeat (SSR) in soybean. Soybean Science, **18**: 248–254.

SSRFinder. 2011. SSRFinder [online]. Available from http://www.maizemap.org/bioinformatics/SSRFINDER/SSR_Finder_Download.html [accessed 13 September 2011].

Swingle, W.T. 1943. The botany of citrus and its wild relatives of the orange subfamily (family Rutaceae, subfamily Aurantioideae). University of California Press, Berkeley, Calif.

Talon, M., and Gmitter, F.G., Jr.. 2008. Citrus genomics. Int. J. Plant Genomics, 2008: 528361. PMID:18509486.

Tanaka, T. 1977. Fundamental discussion of Citrus classification. Stud. Citrologica, **14**: 1–6.

Terol, J., Naranjo, M.A., Ollitrault, P., and Talon, M. 2008. Development of genomic resources for *Citrus clementina*: characterization of three deep-coverage BAC libraries and analysis

of 46,000 BAC end sequences. BMC Genomics, **9**: 423. doi:10.1186/1471-2164-9-423. PMID:18801166.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, **313**(5793): 1596–1604. doi:10.1126/science.1128691. PMID:16973872.

Vicient, C.M., Kalendar, R., and Schulman, A.H. 2001. *Envelope*-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res. **11**(12): 2041–2049. doi:10.1101/gr.193301. PMID:11731494.

Wessler, S.R. 1998. Transposable elements associated with normal plant genes. Physiol. Plant. **103**(4): 581–586. doi:10.1034/j.1399-3054.1998.1030418.x.

Yang, Z.-N., Ye, X.-R., Molina, J., Roose, M.L., and Mirkov, T.E. 2003. Sequence analysis of a 282-kilobase region surrounding the citrus Tristeza virus resistance gene (*Ctv*) locus in *Poncirus trifoliata* L. Raf. Plant Physiol. **131**(2): 482–492. doi:10.1104/pp.011262. PMID:12586873.

You, F.M., Huo, N., Deal, K.R., Gu, Y.Q., Luo, M.C., McGuire, P.E., et al. 2011. Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics, **12**: 59. doi:10.1186/1471-2164-12-59. PMID:21266061.

Zhang, D.-X., and Mabberley, D.J. 2008. Citrus. *In* Flora of China. *Edited by* Z.Y. Wu, P.H. Raven, and D.Y. Hong. China and Missouri Botanical Garden Press, St. Louis, Mo.

Zhou, Z.S, Dell'Orco, M., Saldarelli, P., Turturo, C., Minafra, A., and Martelli, G.P. 2006. Identification of an RNA-silencing suppressor in the genome of *Grapevine virus A*. J. Gen. Virol. **87**(8): 2387–2395. doi:10.1099/vir.0.81893-0. PMID:16847135.