Plant Biotechnology Journal

aab Association of Applied Biologists SEB Society for Experimental Biology

Plant Biotechnology Journal (2016) **14**, pp. 938–950 doi: 10.1111/pbi.12447

An RNA-Seq-based reference transcriptome for Citrus

Javier Terol*, Francisco Tadeo, Daniel Ventimilla and Manuel Talon

Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias (IVIA), Moncada, Valencia, Spain

Received 1 April 2015;
revised 4 June 2015;
accepted 1 July 2015.
**Correspondence* (Tel +34 3424000;
fax +34 3424001; email terol_javalc@gva.es)

Keywords: citrus, RNA-seq, transcriptome.

Summary

Previous RNA-Seq studies in citrus have been focused on physiological processes relevant to fruit quality and productivity of the major species, especially sweet orange. Less attention has been paid to vegetative or reproductive tissues, while most *Citrus* species have never been analysed. In this work, we characterized the transcriptome of vegetative and reproductive tissues from 12 *Citrus* species from all main phylogenetic groups. Our aims were to acquire a complete view of the citrus transcriptome landscape, to improve previous functional annotations and to obtain genetic markers associated with genes of agronomic interest. 28 samples were used for RNA-Seq analysis, obtained from 12 *Citrus* species: *C. medica, C. aurantifolia, C. limon, C. bergamia, C. clementina, C. deliciosa, C. reshni, C. maxima, C. paradisi, C. aurantium, C. sinensis* and *Poncirus trifoliata*. Four different organs were analysed: root, phloem, leaf and flower. A total of 3421 million Illumina reads were produced and mapped against the reference *C. clementina* genome sequence. Transcript discovery pipeline revealed 3326 new genes, the number of genes with alternative splicing was increased to 19 739, and a total of 73 797 transcripts were identified. Differential expression studies between the four tissues showed that gene expression is overall related to the physiological function of the specific organs above any other variable. Variants discovery analysis revealed the presence of indels and SNPs in genes associated with fruit quality and productivity. Pivotal pathways in citrus such as those of flavonoids, flavonols, ethylene and auxin were also analysed in detail.

Introduction

Citrus, including species such as sweet orange, mandarin, lemon or grapefruit, is one of the most important fruit crops in the world, both in terms of fruit production and economical value. Many efforts have been carried out to characterize the genome sequence of the main *Citrus* species: the draft genome of sweet orange, *Citrus sinensis*, was released in 2012 (Xu *et al.*, 2013); more recently a high-quality reference genome sequence of a haploid clementine, *C. clementina*, as well as the genome sequences of mandarin (*C. reticulata*), pummelo (*C. maxima*), sweet orange (*C. sinensis*) and sour orange (*C. aurantium*) was obtained and compared (Wu *et al.*, 2014). Citrus fruits have been traditionally classified into different groups based on the use of molecular markers, although the phylogeny of the species is not yet clear due to the presence of numerous hybrids. Lineages that gave rise to the most modern cultivars are still under discussion (Nicolosi *et al.*, 2000).

The analysis of the transcriptome is a crucial step to characterize any species genome, and during the past years, these studies have been boosted by the development of RNA-Seq technique (Egan *et al.*, 2012; Wang *et al.*, 2009). This approach has been greatly used to improve functional annotation of model plants like Arabidopsis (Filichkin *et al.*, 2010; Ossowski *et al.*, 2008), rice (Lu *et al.*, 2010; Mizuno *et al.*, 2010) and poplar (Ko *et al.*, 2012), with outstanding results. Deep sequencing of the transcriptome has also been applied for the identification of candidate genes in processes of agronomical interest (Canales *et al.*, 2014; Chen *et al.*, 2013; Venu *et al.*, 2011), or to obtain markers for large scale genotyping (Haseneyer *et al.*, 2011; Scaglione *et al.*, 2012).

Transcriptome studies in citrus have been mostly focused on the characterization of physiological processes of high relevance to fruit quality and productivity, especially of sweet orange, as it is the most important citrus fruit for the juice industry. Thus, several works analysed transcriptome changes during fruit ripening of *C. Sinensis* (Shalom *et al.*, 2014; Yu *et al.*, 2012; Yun *et al.*, 2012), and *C. Paradisi* (Patel *et al.*, 2014). RNA-Seq was also used to study the level of heterozygosity of sweet orange and its effect on gene expression (Jiao *et al.*, 2013). The transcriptome profiling of responses to huanglongbing infection of *C. Sinensis* (Martinelli *et al.*, 2012) and *Xylella fastidiosa* infection of *C. reticulata* (Rodrigues *et al.*, 2013) has been also addressed.

However, only a few works have been performed on nonfruit organs (Xu *et al.*, 2013), and most of the *Citrus* species have never been analysed. Therefore, in this work, we carried out RNA-Seq studies of 4 nonfruit organs (flower, leaf, root and phloem) from 12 citrus species, including key members from all main phylogenetic groups, providing a comprehensive view of the citrus transcriptome.

Results and discussion

Overview of RNA-seq analysis

Twenty-eight samples obtained from 4 different organs of 12 *Citrus* species (Table 1) were used for RNA-Seq analysis. The selected species constitute a wide representation of the *Citrus* genus, with species from the 5 main *Citrus* clusters (Nicolosi *et al.*, 2000): citron cluster including *C. medica, C. limon* and *C. bergamia*; mandarin cluster including *C. clementina*, *C. deliciosa* and *C. reshni*; pummelo cluster including *C. maxima, C. paradisi, C. aurantium* and *C. sinensis*; micrantha cluster with *C. aurantifolia;* and *Poncirus trifoliata*, from the Poncirus cluster (Figure 1).

The organs analysed were root, phloem (bark), leaf and flower. Young leaves and flowers were collected from *C. aurantifolia*,

Table 1 Description of the samples analysed with RNA-seq

Sample	Species	Cultivar	Cluster	Organ/Organ
ERS485732	*C. aurantifolia*	Mexican lime	Micrantha	Young leaf
ERS485733	*C. aurantifolia*	Mexican lime	Micrantha	Open flower
ERS485734	*C. aurantium*	Sevillano	Pummelo	Phloem
ERS485735	*C. aurantium*	Sevillano	Pummelo	Root
ERS485736	*C. aurantium*	Sevillano	Pummelo	Young leaf
ERS485737	*C. aurantium*	Sevillano	Pummelo	Open flower
ERS485738	*C. bergamia*	Bergamoto	Citron	Young leaf
ERS485739	*C. bergamia*	Bergamoto	Citron	Open flower
ERS485740	*C. clementina*	Clemenules	Mandarin	Flowers (green button)
ERS485741	*C. clementina*	Clemenules	Mandarin	Flowers (white button)
ERS485742	*C. clementina*	Clemenules	Mandarin	Flowers (petals elongation)
ERS485743	*C. clementina*	Clemenules	Mandarin	Open flower
ERS485744	*C. deliciosa*	Willowleaf	Mandarin	Young leaf
ERS485745	*C. deliciosa*	Willowleaf	Mandarin	Open flower
ERS485746	*C. limon*	Fino lemon	Citron	Young leaf
ERS485747	*C. limon*	Fino lemon	Citron	Open flower
ERS485748	*C. maxima*	Deep Red	Pummelo	Young leaf
ERS485749	*C. maxima*	Deep Red	Pummelo	Open flower
ERS485750	*C. medica*	Citron Diamond	Citron	Young leaf
ERS485751	*C. medica*	Citron Diamond	Citron	Open flower
ERS485752	*C. paradisi*	Star Ruby	Pummelo	Young leaf
ERS485753	*C. paradisi*	Star Ruby	Pummelo	Open flower
ERS485754	*C. reshni*	Cleopatra	Mandarin	Young leaf
ERS485755	*C. reshni*	Cleopatra	Mandarin	Open flower
ERS485756	*C. sinensis*	Navelina	Pummelo	Young leaf
ERS485757	*C. sinensis*	Navelina	Pummelo	Open flower
ERS485758	*Poncirus trifoliata*	Rubidoux	Poncirus	Root
ERS485759	*P. trifoliata*	Rubidoux	Poncirus	Phloem

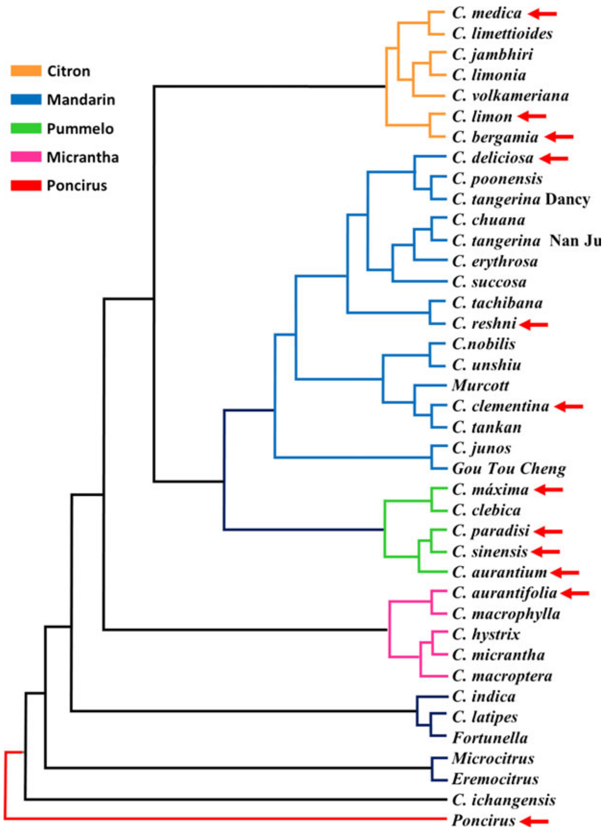

Figure 1 *Citrus* phylogenetic tree according to Nicolosi *et al.* (2000) showing the relationships among the species analysed in this work (arrows). Colour of the branches indicates main citrus groups represented in the RNA-Seq analysis: citron (orange), mandarin (blue), pummelo (green), micrantha (pink) and Poncirus (red).

C. aurantium, *C. bergamia*, *C. reshni*, *C. deliciosa*, *C. limon*, *C. maxima*, *C. medica*, *C. paradisi*, *C. clementina* and *C. sinensis*. Phloem and roots were obtained from *Poncirus* and *C. aurantium*, two species that are used as root stock.

RNA-Seq was carried out as described in Experimental procedures section, and the results are summarized in Table 2. Eight samples were sequenced with single fragment libraries and 50-bp reads, with an average number of 92.4 million reads per sample after quality trimming. The remaining samples were sequenced with paired-end libraries and 75-bp reads, and after quality trimming, the average number of reads per sample was 133.7 million.

Overall, a total of 28 libraries were constructed and sequenced and 3.42 billion reads were produced. After quality trimming to remove low-quality bases and reads, 3.4 billion reads remained, with a total of 235.6 Gb of useful sequence.

Transcript assembly

To obtain a set of reference transcripts and genes, high-quality reads from all samples were mapped to the *C. clementina* reference genome sequence (Wu *et al.*, 2014) as described in Experimental procedures section. Mapped reads were the input for the Transcript Discovery tool, using existing annotations from the Citrus Genome Database (http://www.citrusgenomedb.org/), but adding new transcripts or genes when suggested by mapped reads.

About 2592 million reads were mapped, with 585.2 million reads in pairs (19%), 1079.8 million broken in paired reads (35%) and 256.3 million of gapped reads (10%) (Table S1). The CLC transcriptome assembly tool was the only one that, in a comparative study with ABySS and Velvet, consistently returned large numbers of quality transcripts regardless of the reference used (Misner *et al.*, 2013).

About 341 million reads were mapped to exons resulting in 28 203 genes found and annotated. The average transcript size was 3048.8 bp, and the total transcriptome size was estimated in 77.3 Mb. As the *C. clementina* genome project annotation provided 24 533 genes (Wu *et al.*, 2014), 3326 new genes were discovered in this work (Figure 2). Most of the genes annotated by the international consortium were confirmed by the RNA-seq, except 3891 genes that had <10 reads and did not overcome the above background. From this group, 1086 genes produced corresponding citrus ESTs when a BLASTN search (Camacho *et al.*, 2009) was carried out against the EST section of the GenBank. On the contrary, 2805 predicted genes had no reads mapped and produced no ESTs, but they cannot be discarded as active genes because of the limited treatments and organs used in this work.

Table 2 Sequencing results and quality filtering of reads

Sample	Number of reads	Avg. length	Number of reads after trim	Percentage trimmed	Avg. length after trim
ERS485740	92 861 948	49	92 004 770	100.00	48.7
ERS485741	99 134 150	49	98 432 513	100.00	48.6
ERS485742	87 514 337	49	86 741 095	100.00	48.7
ERS485743	92 066 308	49	91 445 162	100.00	48.6
ERS485758	104 556 381	49	103 197 149	100.00	48.8
ERS485735	94 430 909	49	93 203 308	100.00	48.8
ERS485759	92 814 285	49	91 607 700	100.00	48.8
ERS485734	83 414 573	49	82 413 599	100.00	48.8
ERS485732	152 884 984	76	152 757 332	99.92	74.9
ERS485756	130 703 840	76	130 496 682	99.84	74.9
ERS485754	149 338 896	76	148 972 476	99.75	74.9
ERS485748	137 444 426	76	137 321 017	99.91	74.9
ERS485744	150 501 592	76	150 384 560	99.92	74.9
ERS485750	139 341 630	76	139 220 101	99.91	74.9
ERS485738	125 912 176	76	125 797 182	99.91	75
ERS485736	147 367 278	76	147 259 164	99.93	75
ERS485749	151 137 704	76	150 891 456	99.84	74.6
ERS485733	128 476 212	76	128 335 998	99.89	74.6
ERS485751	126 657 076	76	126 370 891	99.77	74
ERS485747	106 779 768	76	106 515 273	99.75	74
ERS485745	127 634 138	76	127 531 652	99.92	74.6
ERS485757	116 252 938	76	116 088 605	99.86	74.4
ERS485739	100 528 756	76	100 445 936	99.92	75.2
ERS485737	100 161 652	76	99 762 013	99.6	74.9
ERS485753	128 176 568	76	128 014 853	99.87	74.7
ERS485755	152 959 696	76	152 604 182	99.77	74.7
ERS485746	160 659 302	76	160 530 460	99.92	74.7
ERS485752	141 377 378	76	141 256 870	99.91	74.7
TOTAL	3 421 088 901		3 409 601 999	99.90	

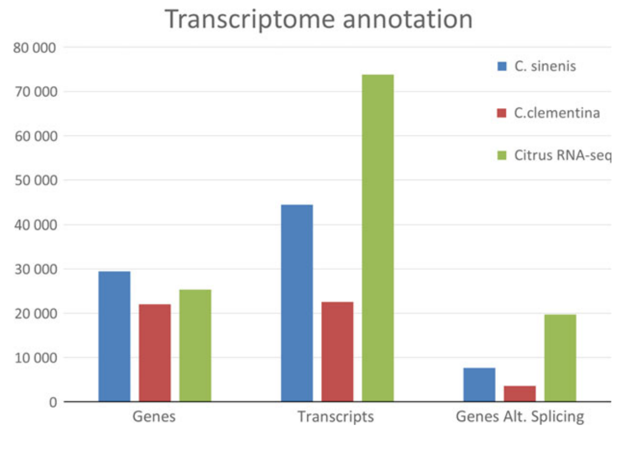

Figure 2 Summary of the transcriptome annotation compared with the ones from the genome projects of *Citrus clementina* (Wu *et al.*, 2014) and *C. Sinensis* (Xu *et al.*, 2013). The total number of genes, transcripts and genes with alternative splicing are shown for the 3 annotations.

A similar comparison with the *C. sinensis* genome project (Xu *et al.*, 2013) is more difficult to interpret, as the quality of the assembly is much lower and the number of scaffolds is very high, which has probably caused an overestimation in the number of genes (Figure 2). A de novo transcriptome analyses carried out in *C. paradisi* flavedo with six different assemblers followed by meta-assembly obtained 29 882 transcripts, with 17 129 ones provided by the CLC assembler (Patel *et al.*, 2014), which is in agreement with the results obtained in this work.

Our analysis reported 5619 genes with 1 transcript, while the number of genes with alternative splicing was 19 739 that had an average number of 2.9 transcripts per gene. 48 875 new alternative acceptor/donor sites and 39 879 new exons were found, with a total of 73 797 transcripts, with an average of 14038.7 reads and 849-fold coverage per transcript that strongly support these results. In a previous work based on the analysis of 1.6 million ESTs from different sources (Wu *et al.*, 2014), 3567 genes with alternative splicing producing 22 536 transcripts were described in *C. clementina*. Furthermore, the *C. sinensis* project identified 7640 genes alternatively spliced and 29 445 different transcripts using RNA-Seq (Xu *et al.*, 2013). Our analysis allowed the identification of 51 261 (3.0-fold increase) and 29 410 (1.7-fold increase) additional transcripts for clementine and sweet orange.

Therefore, this work provides an unprecedented view of the complexity of the transcriptome in *Citrus* species. Our results are in agreement with those works that evidence the substantial increase of sensitiveness of RNA-seq as related to cDNA sequence tag sequencing. Thus, deep transcriptome sequencing in Arabidopsis identified thousands of novel alternatively spliced mRNA isoforms (Filichkin *et al.*, 2010), uncovered additional exons and previously unannotated 5′ and 3′ untranslated regions for pollen-expressed genes (Loraine *et al.*, 2013). Functional annotation of the rice transcriptome by RNA-seq identified 15 708 novel transcriptional active regions and found that ~48% of rice genes showed alternative splicing pattern (Lu *et al.*, 2010). Furthermore, 5877 unannotated transcripts were identified in stress-induced shoot and root (Mizuno *et al.*, 2010). Similarly, deep sequencing of *Populus trichocarpa* xylem transcriptome identified 27902 alternative splicing events, suggesting that at least 36% of the xylem-expressed genes in poplar are alternatively spliced (Bao *et al.*, 2013).

A Circos plot (Krzywinski *et al.*, 2009) showing the distribution of genes, transcripts and reads along the chromosomes of *C. clementina*, the reference genome, is presented in Figure 3. It is worth noting that gene-rich regions accumulate more transcripts and display higher levels of expression, while those with lowest gene density, like centromeric regions, show low expression levels. However, a total of 46 regions, 23 Mb of the genome, had an expression rate 1.5 times higher than the expected considering the number of genes or transcripts. On the contrary, 122 regions, comprising 61 Mb, showed a level of expression half of what could be expected. In Arabidopsis, for instance, it has been reported that physical location along the chromosome affects gene activity. Thus, genes in close proximity are much more likely to be co-expressed than would be expected by chance, while centromeric regions and other stretches had greatly reduced transcriptional activity (Schmid *et al.*, 2005).

In humans, it has also been described the presence of domains with a significant clustering of highly expressed or low-expressed genes, suggesting they are an integral part of a higher order structure in the genome related to transcriptional regulation (Versteeg *et al.*, 2003). Our results suggest a similar organization of the transcriptomic activity in *Citrus* species.

Functional annotation of transcripts

The longest transcript from each gene was selected for functional annotation performed with Blast2GO (Conesa *et al.*, 2005), InterProScan (Jones *et al.*, 2014), EC enzyme codes and KEGG (Kanehisa *et al.*, 2014) pathways, which resulted in 24 502 transcripts annotated with GO terms and/or functional domains.

The annotation showed that 307 transcripts were classified as transposable elements (TEs) and therefore should not be considered as genes. Consequently, the number of *real* citrus genes should be closer to 27 530, than to 27 837 initially found. These TEs corresponded mainly to mutator-like (61), copia (51) or gypsy (8) elements, with 38 unclassified TEs. The fact that these transposable elements were found in the RNA-Seq analysis as well as the high number of reads mapped to them (1 183 654) indicates that they are very active in the citrus genome, an observation that may be in part related to the relevant number of spontaneous mutations that are found in citrus (Butelli *et al.*, 2012; De Felice *et al.*, 2009; Terol *et al.*, 2015).

The functional annotation obtained for the genes previously described was almost identical to the one reported by the International Citrus Consortium, available at phytozome (Goodstein *et al.*, 2012). Therefore, a summary of the annotation of the 3326 new genes is provided. Functional annotation was found for 1262 of these genes, while the rest remained as unknown. Thirty-seven new GO terms from 58 genes were added to the annotation, corresponding to 18 molecular functions, 15 biolog-

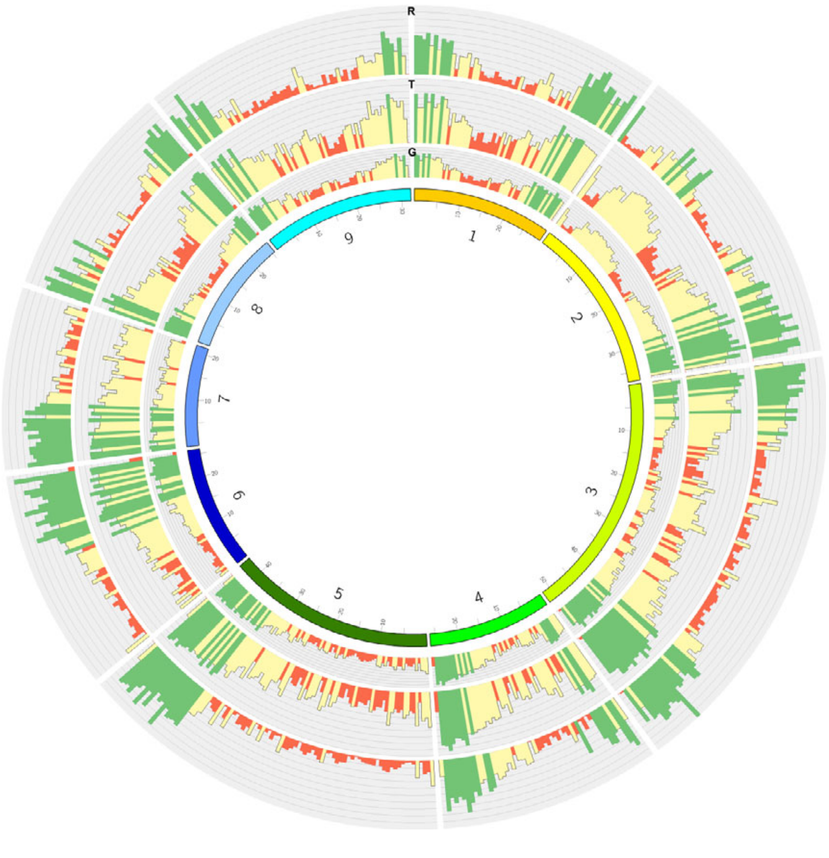

Figure 3 Circos plot showing the transcriptional activity of the citrus genome. Inner circle represents the 9 chromosomes of the reference genome, *Citrus clementina,* and the different concentric layers show the number of genes (G), transcripts (T) and reads (R) per 500 Kb. Scales are relative for each layer; red and green bars indicate bins that are 1.5 times below and above average, respectively.

ical processes and 4 cellular components that were described for the first time in citrus. In other cases, the number of genes associated with a GO term increased remarkably, as is shown in Table S2. Homologs of 290 genes were described for the first time, including 29 transcription factors belonging to MADs box (29), ethylene responsive (2) or WRKY (2) families. A total of 139 enzymatic activities (ECs) from 65 different pathways were found and 10 of them were novel ones in citrus. The number of genes related to several enzymatic activities increased significantly (Table 3).

In summary, our analysis provides a significant improvement in the description of the citrus transcriptome, both in terms of new genes and new transcripts from known genes that will allow a better understanding of the genetic regulation controlling important biological processes that are responsible of desirable traits for citrus improvement.

Differential expression analysis in organs

To perform differential expression analyses, RNA-seq reads were grouped by organ and mapped against the reference transcriptome (Table S1). Gene expression was assessed in each organ compared to the rest and genes were counted as expressed in an organ if a minimum of reads per kb per million reads (RPKM) of 1 was observed. The expression of 13 614 genes was detected in all four organs, while 1620, 294, 356 and 329 genes were exclusively expressed in flower, root, phloem and leaf organs, respectively (Figure 4).

The study of transcription factors (TFs) found a total of 409 genes belonging to the MYB (136), bHLH (72), WD40 (17), MADs (80), WRKY (55) and ERF (49) families. A total of 152 genes were expressed in all four organs, while 7, 4, 12 and 34 were leaf, phloem, root and flower specific, respectively (Table S3). This distribution might reflect the different regulatory roles of these factors in the analysed organs, in a similar manner that was found in peach (Wang *et al.*, 2013).

Table 3 The 20 most increased enzymatic activities

EC	New	ICGSC	Total	Enzymatic activity
4.2.3.22	4	0	4	Germacradienol synthase
2.7.9.1	1	0	1	Pyruvate, phosphate dikinase
2.7.8.11	1	0	1	CDP-diacylglycerol-inositol 3-phosphatidyltransferase
4.2.3.20	2	0	2	(R)-limonene synthase
6.3.3.1	2	0	2	Phosphoribosylformylglycinamidine cyclo-ligase
1.1.1.14	1	0	1	L-iditol 2-dehydrogenase
1.14.13.76	1	0	1	Taxane 10-beta-hydroxylase
5.5.1.12	1	0	1	Copalyl diphosphate synthase
1.14.21.3	1	0	1	Berbamunine synthase
4.2.1.17	3	3	6	Enoyl-CoA hydratase
3.6.1.15	18	278	296	Nucleoside-triphosphate phosphatase
1.1.1.35	3	3	6	3-hydroxyacyl-CoA dehydrogenase
1.11.1.7	12	83	95	Peroxidase
3.2.1.67	11	9	20	Galacturan 1,4-alpha-galacturonidase
3.2.1.15	11	34	45	Polygalacturonase
2.7.7.6	9	53	62	DNA-directed RNA polymerase
3.2.1.22	4	5	9	Alpha-galactosidase
3.1.1.11	6	104	110	Pectinesterase
2.6.1.1	1	16	17	Aspartate transaminase
1.14.14.1	1	27	28	Unspecific monooxygenase

In a parallel approach, differentially expressed genes (DEGs) were examined using the EdgeR package (Robinson *et al.*, 2010), and the results were filtered with a FDR *P*-value correction value <0.05 and a fold change value >1.5 or <−1.5. As a result, 4466, 631, 272 and 5825 genes were up-regulated, while 6273, 270, 41 and 6294 genes were down-regulated in leaf, phloem, root and flower, respectively. It was noticeable that the number of DEGs in phloem and root was one order of magnitude lower than in flower or leaf, however, when both tissues were grouped and compared against the other organs, a total of 3287 genes were found to be down-regulated and 3346 overexpressed, suggesting that root and phloem share many DEGs that are quenched when one organ is compared against the other. Our data support a correlation of gene expression in shoot and root that has been previously reported (Dash *et al.*, 2014; Kelly *et al.*, 2014; Sarkar *et al.*, 2007).

To evaluate the functional properties of the organ-specific genes, annotation enrichment analyses were carried out with the Fisher's exact test, considering those genes that where exclusive or overexpressed in a given organ. This way, 11, 14, 437 and 468 GO terms were significantly enriched in root, phloem, leaf and flower, respectively, that were in agreement with the main functions performed by the analysed organs (Figure 5).

Functional enrichment in flower was related to the morphogenesis of floral organs, the role of auxin in floral differentiation, pollen differentiation, tube development, etc. This enrichment was rather similar to that obtained in the analysis of the transcriptome during flower development in chickpea (Singh *et al.*, 2013).

About 367 genes with significant homology to TFs were overexpressed in flowers, including the most important regulatory families: bHLH (34), zinc finger (92), MADs box (30), MYB (40), homeo-box (36) or AP2/ERF (31). Some of them perform crucial roles during flower differentiation and development: CONSTANS (CO)that induces flower differentiation (Valverde, 2011);. the floral homeotic gene APETALA2 (AP2) (Jofuku *et al.*, 1994); AGAMOUS and Sepallata 1, 2 and 3; MADs box homeotic genes (Gómez-Mena *et al.*, 2005; Pelaz *et al.*, 2000); or MYB transcription factor r2r3-myb that activates the biosynthesis of anthocyanins (Petroni and Tonelli, 2011).

Genes overexpressed in leaf displayed functions related to its development and the organization of photosynthetic machinery, photosynthesis itself, or response to stresses with the involvement

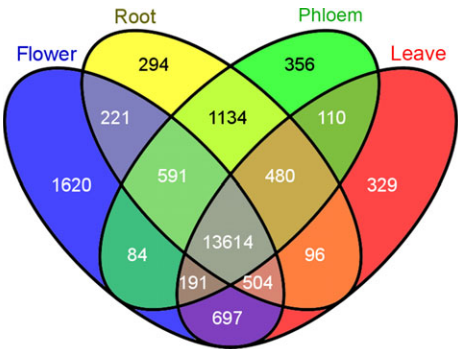

Figure 4 Venn diagram showing gene expression in flower, root, phloem and leaf. Samples were grouped by tissues, and expression was normalized to RPKM. Genes were counted as expressed in an organ if a minimum of RPKM = 1 was observed in the organ.

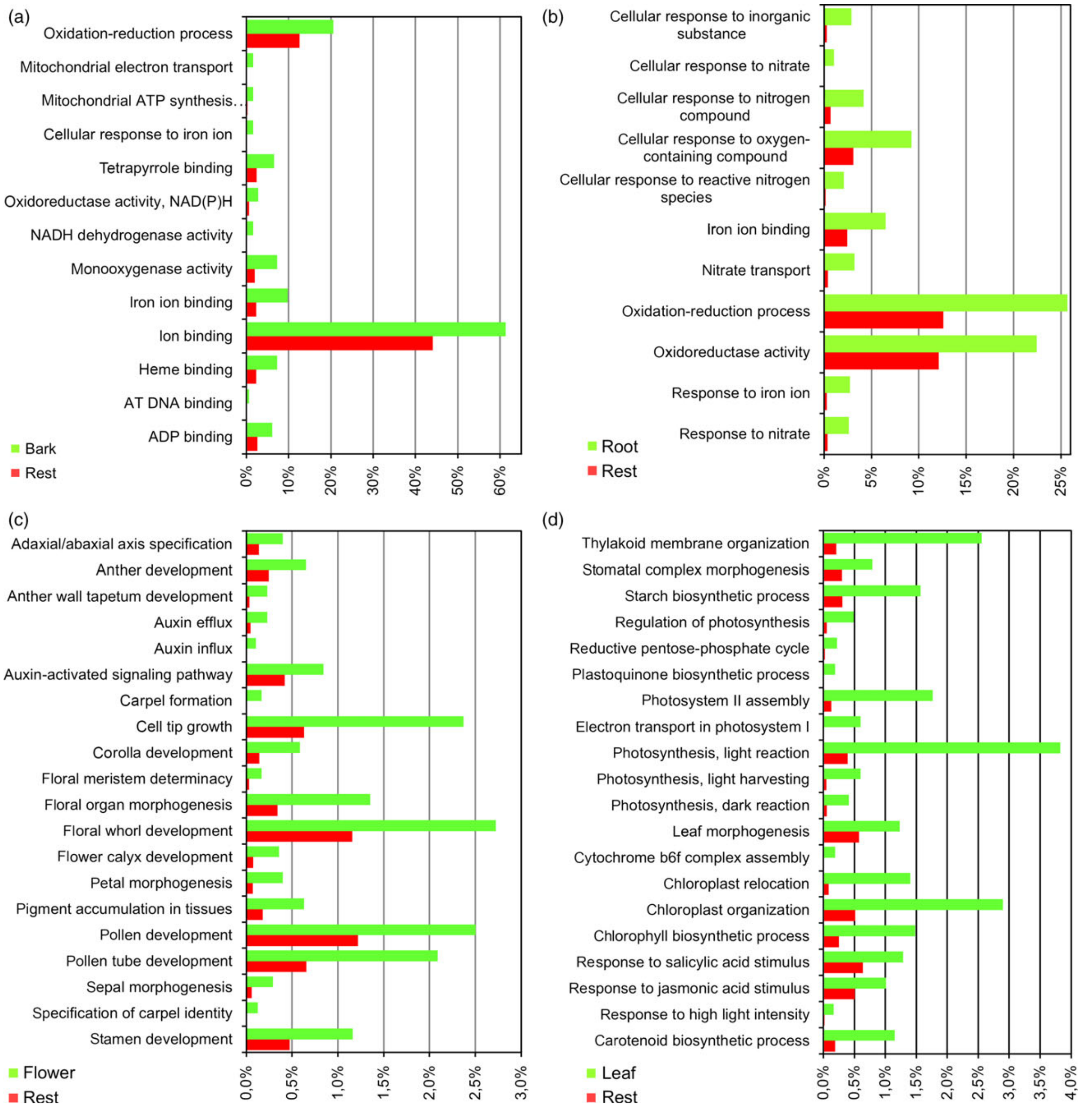

Figure 5 Significantly enriched GO terms in the differentially expressed genes in phloem (a), root (b), flower (c) and leaf (d). Horizontal axis shows the percentage of DEGs displaying a GO annotation (green) and in the control group (red).

of the jasmonic and salicylic acid pathways. The number of TFs overexpressed in flower (367) was much larger than in leaf (248), probably reflecting the complexity of the regulatory pathways controlling the development of reproductive organs. On the contrary, the number of chloroplastic genes overexpressed in leaf (286) was more than 3 times larger than in flower (86).

A total of 35 cytochrome P450 genes were found to be overexpressed in leaf, an identical number that was found in a co-expression analysis of the cytochrome P450 superfamily in Arabidopsis, that showed an unexpectedly large subset of 35 P450 genes being mapped to pathways identified as ‘plastidial isoprenoids’, ‘photosystems’, ‘photosynthesis’ and ‘biogenesis of the chloroplast’ with very high expression in all green organs. These data might indicate that a number of plant P450 enzymes have functions related to primary photosynthetic metabolism for the synthesis of antioxidants, plastidial structural components, signalling molecules related to energetic metabolism or light perception (Ehlting *et al.*, 2008).

Fifty-four members of the family of genes coding for LRR receptor-like serine/threonine proteins were also highly expressed in leaf. These receptors play an important role in signalling during pathogen recognition and the subsequent activation of plant defence mechanisms (Afzal *et al.*, 2008). These data are in agreement with the enrichment in stress-response genes activated by jasmonic and salicylic acid. Actually, it has been found that genes encoding for receptor-like protein kinases are targets of pathogen, and salicylic acid-induced WRKY DNA-binding proteins in Arabidopsis (Du and Chen, 2000).

Differential expression in root and phloem is displayed by genes related to binding and transport of inorganic substances: nitrogen compounds, iron, copper, aluminium or calcium. Among the most highly expressed genes in root and phloem was the superoxide dismutase (SOD) gene that is expressed in response to the oxidative stress caused by drought and salinity, the most serious abiotic stresses affecting citrus culture in the Mediterranean Basin (Gueta-Dahan *et al.*, 1997). Homologs of the Arabidopsis copper (Cu) chaperones, antioxidant protein1 (ATX1) and ATX1-like copper chaperone (CCH) were also highly expressed transcripts in root and shoot, which are required to maintain Cu homoeostasis to facilitate its use and avoid its toxicity (Shin *et al.*, 2012). A homolog to a two-pore calcium channel 1 (AtTPC1), gene that in Arabidopsis is part of a signalling system based on Ca2+ waves that contribute to whole-plant stress tolerance (Choi *et al.*, 2014), was also found.

The different expression profile of the four analysed tissues becomes evident in the principal component analysis (PCA) carried out with the 28 samples (Figure 6), which shows the correlation between the origin of the sample and the expression of the genes. On the contrary, no correlation is observed when the samples were grouped by species (Figure S1), except for the Poncirus ones, probably due to the larger genetic distance of this species with respect the other citrus. In fact, differential expression analysis between species using the same organ (data not shown) yielded very low number of DEGs, indicating that the organs and stages used in this work were not suitable for comparison studies between species.

The differential expression observed between the four organs can also be observed at the genomic level (Figure 7), as there were regions where gene expression in a specific organ was much higher than in the rest. We identified 41 regions that showed expression levels significantly higher than the average of the four tissues in that bin: 20 regions were overexpressed in flower, nine in leaf, seven in phloem and nine in root. Two regions showed simultaneous overexpression in leaf/flower and six in root/phloem, indicating possible co-expression patterns.

In general, these results agree with the concept that gene expression is overall related to the physiological function of the specific organs, above any other consideration. Thus, our results are similar to those obtained in a comprehensive microarray study of the tomato transcriptome that identified 465 co-expression/functional modules, and found differential expression in leaf, fruit and root (Fukushima *et al.*, 2012). In Arabidopsis, a genomewide expression analysis of 18 organ or tissue types showed that they had a defining genome expression pattern and that the degree to which organs share expression profiles was highly correlated with the biological relationship of organ types (Ma *et al.*, 2005). In a microarray study in Arabidopsis, the largest differences in gene expression were observed when comparing samples from different organs: on average, 10-fold more genes were differentially expressed between organs as compared to any other experimental variables (Aceituno *et al.*, 2008).

To validate the differential expression, 10 DEGs were selected for qRT-PCR analysis. Total RNA extracted used in the RNA-Seq was also utilized in these experiments that were carried out as described in Experimental procedures. The genes and the primers used for PCR are shown in Table S4. The results obtained confirmed in all cases the differential expression observed with RNA-Seq, and all the analysed genes showed higher levels of expression in the tissues where they had been identified as DEGs (Figure S2).

Metabolic pathways

Three relevant pathways were analysed in detail: flavonoids, auxin and ethylene biosynthesis. Flavonoids are plant secondary metabolites implicated in the control of auxin transport, defence, flower colouring, seed dispersal and many other processes (Brunetti *et al.*, 2013; Buer *et al.*, 2007; Taylor and Grotewold, 2005; Treutter, 2005). Auxin is a plant hormone involved in an extraordinarily broad variety of biological processes, such as cell polarity, cell cycle control and organ patterning (Luschnig, 2001; Sauer *et al.*, 2013). Finally, ethylene plays numerous roles in the development and environmental responses of the plant like germination, senescence, abscission or ripening, as well as stress

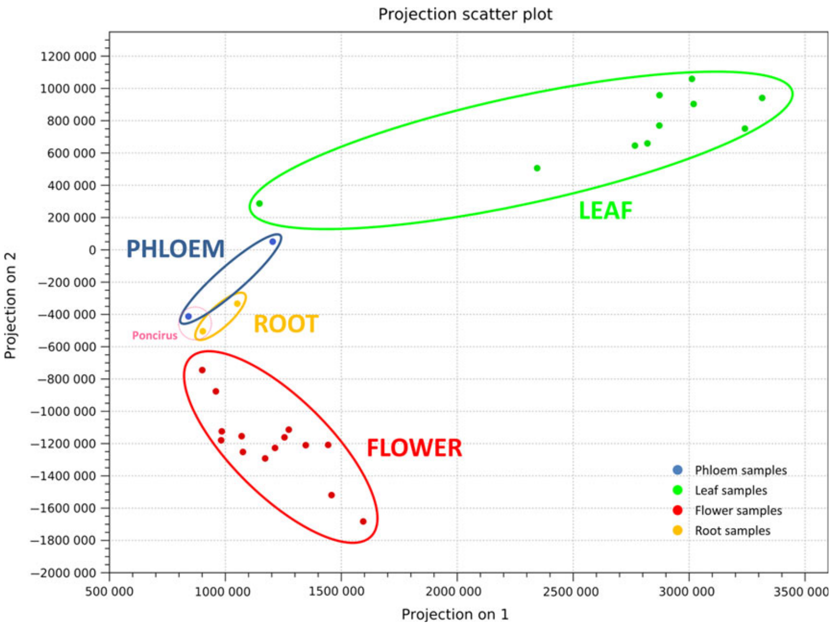

Figure 6 Principal component analysis in 2 dimensions showing the correlation between gene expression and the tissue origin of the samples that cluster together, separating clearly the 4 tissues. Poncirus is the only exception, as samples from the species are closer between them than with the ones from the same tissue.

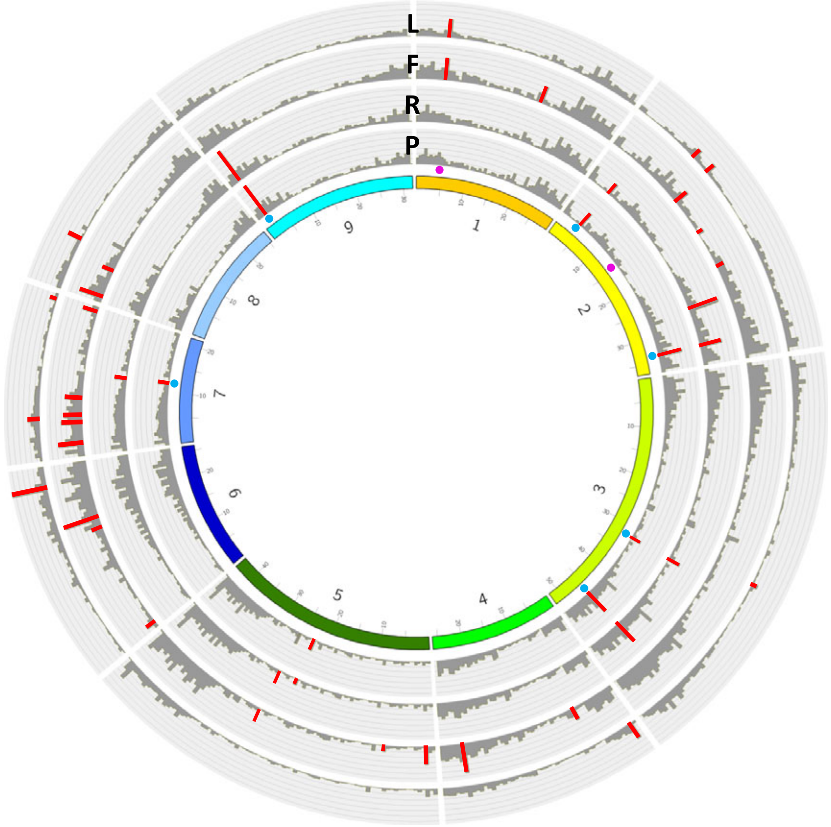

Figure 7 Circos plot showing the transcriptional activity of the four organs analysed on the citrus genome. Inner circle represents the 9 chromosomes of the reference genome, *Citrus clementina,* and the different concentric layers show the number of reads per million reads per 500 Kb in phloem (P), root (R), flower (F) and leaf (L). Those bins showing expression significantly above average in that region are highlighted in red. Several regions of the genome display higher levels of expression in a tissue-specific manner. Regions displaying simultaneous over expression of phloem/root and leaf/flower are indicated with blue and magenta circles, respectively.

and pathogen responses (Iqbal *et al.*, 2013; Merchante *et al.*, 2013). In addition, auxin and ethylene act synergistically to control root elongation and root hair formation, as well as antagonistically in lateral root formation and hypocotyl elongation (Muday *et al.*, 2012). In some of these processes, they regulate and/or interact with flavonoids (Buer *et al.*, 2006; Smith *et al.*, 2003).

Initially, the EC number was used to identify the enzymes involved in these 3 pathways in citrus. Then, the closest protein showing empirical evidences of its activity was used to perform a phylogenetic analysis with the citrus proteins (see an example in Figure S3). Only proteins clustering with the control sequences were considered for the analysed pathway, so the large initial number of proteins was reduced to a more accurate one. The transcriptional activity of the selected genes was estimated using the RPKM data obtained from the RNA-Seq.

In the flavonoid and flavonol biosynthetic pathway (Kanehisa *et al.*, 2014), 18 enzymatic activities were represented by 51 citrus genes (Figure S4). Chalcone-flavanone isomerase, which catalyses the formation of naringenin, and dihydrokaempferol 4-reductase, which forms final products for flavonoid biosynthesis, were highly expressed in flowers but at a very low level in the other organs. Actually, all essential enzymes catalysing the formation of naringenin, apiforol, luteforol and leucocyanidin were found in flower samples, except flavonoid 3′,5′-dihydroxyflavanone, that was also produced in leaf and phloem. Our data would be in agreement with the relevant role of flavonoids in flowers and later in fruits (Hichri *et al.*, 2011).

For the ethylene biosynthesis route, 3 SAM-synthase, 5 ACC synthase and 3 ACC oxidase genes were found in citrus (Figure S5). The SAM-synthase displayed the highest expression levels compared with the other enzymes, with maximum values in flower; ACC synthase genes had the lowest expression, but were present in the 4 organs; ACC oxidase genes were also expressed in the four organs with low values, except for Ciclev10015962 that showed an expression peak in leaf. The rate-limiting step of ethylene synthesis is the conversion of SAM to ACC by ACC synthase: expression of the ACC synthase genes is highly regulated by a variety of signals and active ACC synthase is labile and present at low levels (Wang *et al.*, 2002). The low levels of expression of ACC synthase in citrus, as well as the variations found in the different genes and organs, would be in agreement with the role of regulation of ACC synthase in the control of ethylene levels. Differences in the regulation of ACC synthase isoforms have been already described in tomato (Barry *et al.*, 2000; Nakatsuka *et al.*, 1998).

The synthesis of auxins in plants proceeds through both the indole-3-pyruvic acid (IPA) pathways and the indole-3-acetamide (IAM) (Figure S6) (Mano and Nemoto, 2012). In the IPA pathway, two homologs for TAR, a gene coding for tryptophan aminotransferase, and six for YUC that encodes for flavin monooxygenase were identified. The higher level of expression of the genes from the IPA pathway in citrus would be in agreement with the fact that most of IAA in plants is produced through this pathway (Brumos *et al.*, 2014).

In the IAM pathway, homologs of AM1, which produces IAA from IAM, have been found in Arabidopsis and tobacco (Mano *et al.*, 2010), and we identified 9 putative homologs in citrus. Indole-3-acetamide (IAM) has also been found in Arabidopsis and other species, including *Citrus unshiu* (Takahashi *et al.*, 1975), and indole-3-acetamide hydrolase activity has been detected in crude extract from young fruits of *P. trifoliata* (Kawaguchi *et al.*, 1991), indicating that IAM is a native compound and that the IAM pathway is operative in citrus. However, no clear homolog of the bacterial gene aux1/iaaM has been found in plants, including our analyses in citrus, despite of several evidences if its activity (Klee *et al.*, 1987).

Variant analysis

RNA-seq has proven to be a suitable resource for the development of molecular markers in many species, including plants (Piskol *et al.*, 2013; Severin *et al.*, 2010). The improvement of RNA-seq de novo assembly has been especially useful for species with no genome sequence available like onion, *Allium cepa* (Kim *et al.*, 2014); rye, *Secale cereale* (Haseneyer *et al.*, 2011); red clover, *Trifolium pratense* (Yates *et al.*, 2014); or artichoke, *Cynara cardunculus* (Scaglione *et al.*, 2012).

A major benefit of RNA-seq in SNP calling relative to whole genome sequencing is the reduction in the effective size of the genome. Utilization of RNA-seq requires much less sequence depth to identify the majority of the variants in medium to highly expressed transcripts relative to whole genome sequencing. This advantage has been used, for example, to explore genetic variability in a large number of populations and varieties from *Zea mays* (Hansey *et al.*, 2012) or *Brassica rapa* (Devisetty *et al.*, 2014; Paritosh *et al.*, 2013). Validation of SNPs by different methods showed rates higher than 90% indicating the accuracy of this method (Devisetty *et al.*, 2014; Hansey *et al.*, 2012; Scaglione *et al.*, 2012).

For our variant calling analysis, read samples were grouped by species and proceeded as described in the Experimental procedure section. Results are summarized in Table 4. A total of 6.5 million high-quality variants were found, including 5.74 million SNPs and 534 thousand indels. The number of variants ranged from 220 557 in *C. clementina* to 905 707 in *C. aurantifolia* a difference proportional to the phylogenetic distance of each species with respect to the haploid *C. clementina* used as the reference genome. Approximately 890 000 variants, 14% of the total, were privative of the 12 species, and again for each one, the number of privative variants reflected the phylogenetic distance with respect to clementine.

Analysis of the SNP set shows that transitions (60%) were more abundant than transversions (40%), and the most abundant changes were the transitions C/T and A/G, with about more than 990 000 each, independently of the species analysed.

As the reference used was the haploid clementine genome, the only possible variants are heterozygous, and therefore, the expected number of homozygous variants in the clementine samples should be 0. However, a total of 38 809 homozygotic variants were found in our samples, suggesting the presence of false positives. Further analysis showed that in all these positions, the alternative allele was in a proportion of reads higher than 85%, with reads in both strands, revealing mistakes in the Sanger sequencing (Figure S7).

An investigation of the heterozygosity in *C. sinensis* identified 226 000 SNPs and 47 700 indels in 17 765 protein-coding loci annotated in the sweet orange genome and predicted that 32 460 SNPs potentially affect gene function (Jiao *et al.*, 2013). In our work, we identified 436 196 SNPs and 41 139 indels, affecting 13 664 genes, probably due to the fact that the haploid *C. clementina* genome sequence was used as reference, which would increase significantly the number of variants.

We analysed the effect of the variants on the coding regions and found that a total of 20 328 genes displayed changes in

Table 4 Variant discovery summary

Type/Species	*Citrus clementina*	*Citrus deliciosa*	*Citrus reshni*	*Citrus sinensis*	*Citrus paradisi*	*Citrus maxima*	*Citrus aurantium*	*Citrus bergamia*	*Citrus medica*	*Poncirus trifoliata*	*Citrus limon*	*Citrus aurantifolia*
Total	220 557	231 269	261 766	482 212	531 640	556 422	627 874	628 462	646 468	654 534	770 313	905 707
SNV	198 070	202 063	227 876	425 980	469 411	486 073	562 922	553 168	562 794	587 573	673 054	793 334
Deletion	8568	10 691	12 896	20 652	22 708	25 715	22 474	26 845	29 541	23 009	34 792	38 842
Insertion	7122	10 794	12 035	19 371	20 869	23 886	18 655	25 681	28 988	19 294	33 097	38 108
MNV	6796	7721	8959	16 209	18 652	20 748	23 823	22 768	25 145	24 658	29 370	35 423
Heterozygous	220 556	173 489	131 965	394 843	358 059	171 368	493 681	416 749	101 201	511 391	437 055	565 198
Homozygous	0	57 780	129 801	87 369	173 581	385 054	134 193	211 713	545 267	143 143	333 258	340 509
Privative variants	31 555	33 962	52 842	38 897	40 969	91 548	124 867	22 177	31 415	110 086	34 557	275 882
SNV/Kb	2.6	2.6	2.9	5.5	6.1	6.3	7.3	7.2	7.3	7.6	8.7	10.3
Transitions	119 679	121 694	137 076	256 887	283 858	291 869	599 881	332 131	336 960	354 735	401 247	474 641
Transversions	78 391	80 369	90 800	169 093	185 553	194 204	393 652	221 037	225 834	232 838	271 807	318 693
Genes with coding region changes	9369	9038	9369	13 664	14 791	72 933	15 105	15 029	83 769	14 500	16 269	16 685
Coding region changes	74 422	71 756	74 422	146 750	169 630	174 317	179 116	196 214	198 769	202 597	224 776	267 627
Average changes/gene	7.9	7.9	7.9	10.7	11.5	2.4	11.9	13.1	2.4	14.0	13.8	16.0
Aa changes	44 230	41 460	44 230	85 152	98 483	101 223	103 071	113 579	116 295	118 460	130 325	156 023
Aa changes/gene	4.7	4.6	4.7	6.2	6.7	1.4	6.8	7.6	1.4	8.2	8.0	9.0

the coding regions. *C. deliciosa* with 9038 genes affected had the lowest number, while *C. aurantium* with 16 736 genes had the largest one. A large number of changes on the coding regions were privative of each species that could be interesting targets for the development of markers associated with differential traits.

Overall, the variant analysis performed in this work provides a valuable resource of genetic markers close to or within coding sequences that make them especially useful for citrus breeding programs. Furthermore, for 6 of the species, *C. medica*, *C. limon*, *C. bergamia*, *C. paradisi*, *C. aurantifolia* and *P. trifoliata*, this is the first genomewide set of markers described so far, which increases the value of the results obtained in our work.

Conclusions

In this study, we have used deep RNA-Seq of 12 species using 4 different organs, which provided enough coverage for the discovery of unknown transcripts, the exponential increase in the quantity of transcript variants, the detection of a large number of polymorphisms and an update of existing annotation.

Experimental procedures

Plant material

Vegetal material was obtained from *C. aurantifolia* (Mexican lime), *C. aurantium* (Sevillano sour orange), *C. bergamia* (Bergamota de Reggio Calabria), *C. clementina* (Clemenule*s mandarin*), *C. deliciosa* (Willowleaf mandarin), *C. limon* (Fino lemon), *C. maxima* (Deep Red pummelo), *C. medica* (Diamond citron), *C. paradisi* (Star Ruby grapefruit), *C. reshni* (Cleopatra mandarin), *C. sinensis* (Navelina sweet orange) and *P. trifoliata* (Rubidoux). The four organs analysed were leaf (young and fully developed), phloem (bark from last sprout and old branches), root (young secondary roots) and flowers (complete flowers in anthesis). Flowers at different developmental stages— green button, white button, petals elongation and anthesis—were collected from *C. clementina* (Table 1). Samples were stored at −80°C until RNA extraction.

RNA extraction

Total RNA was isolated from frozen organs using acid phenol extraction and lithium chloride precipitation method as described in Ecker and Davis (Ecker and Davis, 1987). PolyA RNA was isolated with RNEASY™ kit from Qiagen, following provider's protocol. Purified polyA RNA was diluted in 100 μL of free RNAase water and quantified using Nanodrop.

Illumina TruSeq™ RNA sequencing library preparation

Libraries from total RNA were prepared using the TruSeq™ RNA sample preparation kit (Illumina Inc., San Diego, California) according to manufacturer's protocol. Briefly, 0.5 μg of total RNA was used for polyA-based mRNA enrichment selection using oligo-dT magnetic beads followed by fragmentation by divalent cations at elevated temperature resulting into fragments of 80–250 nt, with the major peak at 130 nt. First-strand cDNA synthesis by random hexamers and reverse transcriptase was followed by the second-strand cDNA synthesis performed using RNAseH and DNA Pol I. Double-stranded cDNA was end-repaired, 3′ adenylated and the 3′-'T' nucleotide at the Illumina adaptor was used for the adaptor ligation. The ligation product was amplified with 15 cycles of PCR.

Sequencing, base calling and quality trimming

Paired end libraries: Each library was sequenced using TruSeq SBS Kit v3-HS, in paired-end mode with the read length 2 × 76 bp. A minimum of 50 million paired-end reads for each sample were generated on HiSeq2000 (Illumina, Inc) following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files by CASSAVA. Low-quality bases with a Phred score lower than 13 (base-calling error probability limit = 0.05) were removed with CLC Genomics Workbench 7.0.3. All the reads are available at ENA, study accession number PRJEB6342.

Transcript discovery

A set of reference transcripts and genes was inferred after reads from all samples were mapped to *the C. clementina* reference genome using the large gap mapper tool and Transcript Discovery Plug-in from CLC-Bio Genomics Workbench 7.0.3 with default parameters. Previous annotation of the clementine genome was used; therefore, the existing gene and mRNA annotations were kept and new ones added only when the mapped reads suggested a new transcript or gene.

Functional annotation

Blast2Go (Conesa *et al.*, 2005) was used for functional annotation of the longest transcript from each gene. Sequences were also searched for conserved protein domains with IPRscan 5.0 (Jones *et al.*, 2014) using the Blast2Go suite.

RNA-Seq and differential expression analyses

RNA-Seq analysis was carried out by mapping next-generation sequencing reads and counting and distributing the reads across genes and transcripts with CLC-Bio Genomics Workbench 7.0.3 tool (Mortazavi *et al.*, 2008), with default parameters. The genome sequence of *C. clementina* annotated with the results of the transcripts detection analyses was the reference. Differential expression studies were carried out with EdgeR package (Robinson *et al.*, 2010) with a total read count filter of 5 and *P*-values with FDR correction. Samples were grouped based on the organ they were collected from.

qRT-PCR analysis

RNA extractions were performed as described above and RNA concentration was determined by fluorometric assays in triplicate using RiboGreen dye (Molecular Probes Eugene, Oregon) following the manufacturer's instructions. qRT-PCR was performed with a LightCycler 2.0 Instrument (Roche, Basel, Switzerland) equipped with LightCycler Software version 4.0 as described previously (Agustí *et al.*, 2007). Transformation of fluorescence intensity data into relative mRNA levels was carried out using a standard curve constructed with a 10-fold dilution series of a single RNA sample. The relative expression level of each gene was calculated using the 2-ΔΔCT (cycle threshold) method, and CitUBQ was used as an internal control (Livak and Schmittgen, 2001). Specificity of the amplification reactions was assessed by postamplification dissociation curves and product sequencing. Results were expressed as contrasts between compared tissues. The sequences of the forward and reverse primers and the size of the resulting fragments are listed in Table S4.

Variants discovery

Fixed Ploidy Variant Detection tool from CLC-Bio Genome Workbench was used for variant discovery, mainly SNPs and indels. The parameters were set for an expected ploidy of 2, a required variant probability of 90%, a minimum coverage of 10 and a minimum frequency for the alternative allele of 20%. Base quality of the variant position and of the 5 neighbour up- and downstream positions was set to 20 and 15, respectively, to filter the results. The amino acid changes caused by the variants were also determined when they were located at coding regions, focusing on nonsynonymous changes, using the tool from CLC-Bio Genome Workbench.

Acknowledgements

We want to acknowledge Isabel Sanchis and Antonio Lopez for their participation in this project as field and laboratory technicians. This project was funded by the Spanish Grant from the Ministerio De Economia y Competitividad AGL 2011-30240. JT was funded by the IVIA- Consellería de Agricultura de la Generalitat Valenciana Grant Proy-IVIA-2013/07.

References

Aceituno, F.F., Moseyko, N., Rhee, S.Y. and Gutierrez, R.A. (2008) The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genom.* **9**, 438.

Afzal, A.J., Wood, A.J. and Lightfoot, D.A. (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol. Plant Microbe Interact.* **21**, 507–517.

Agustí, J., Zapater, M., Iglesias, D.J., Cercós, M., Tadeo, F.R. and Talón, M. (2007) Differential expression of putative 9-cis-epoxycarotenoid dioxygenases and abscisic acid accumulation in water stressed vegetative and reproductive tissues of citrus. *Plant Sci.* **172**, 85–94.

Bao, H., Li, E., Mansfield, S., Cronk, Q., El-Kassaby, Y. and Douglas, C. (2013) The developing xylem transcriptome and genome-wide analysis of alternative splicing in *Populus trichocarpa* (black cottonwood) populations. *BMC Genom.* **14**, 359.

Barry, C.S., Llop-Tous, M.I. and Grierson, D. (2000) The regulation of 1-aminocyclopropane-1-carboxylic acid synthase gene expression during the transition from system-1 to system-2 ethylene synthesis in tomato. *Plant Physiol.* **123**, 979–986.

Brumos, J., Alonso, J.M. and Stepanova, A.N. (2014) Genetic aspects of auxin biosynthesis and its regulation. *Physiol. Plant.* **151**, 3–12.

Brunetti, C., Di Ferdinando, M., Fini, A., Pollastri, S. and Tattini, M. (2013) Flavonoids as antioxidants and developmental regulators: relative significance in plants and humans. *Int. J. Mol. Sci.* **14**, 3540–3555.

Buer, C.S., Sukumar, P. and Muday, G.K. (2006) Ethylene modulates flavonoid accumulation and gravitropic responses in roots of Arabidopsis. *Plant Physiol.* **140**, 1384–1396.

Buer, C.S., Muday, G.K. and Djordjevic, M.A. (2007) Flavonoids are differentially taken up and transported long distances in Arabidopsis. *Plant Physiol.* **145**, 478–490.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G. and Martin, C. (2012) Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell,* **24**, 1242–1255.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Canales, J., Bautista, R., Label, P., Gomez-Maldonado, J., Lesur, I., Fernandez-Pozo, N., Rueda-Lopez, M., Guerrero-Fernandez, D., Castro-Rodriguez, V., Benzekri, H., Canas, R.A., Guevara, M.A., Rodrigues, A., Seoane, P., Teyssier, C., Morel, A., Ehrenmann, F., Le Provost, G., Lalanne, C., Noirot, C., Klopp, C., Reymond, I., Garcia-Gutierrez, A., Trontin, J.F., Lelu-Walter, M.A., Miguel, C., Cervera, M.T., Canton, F.R., Plomion, C., Harvengt, L., Avila, C., Gonzalo Claros, M. and Canovas, F.M. (2014) De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnol. J.* **12**, 286–299.

Chen, X., Zhu, W., Azam, S., Li, H., Zhu, F., Li, H., Hong, Y., Liu, H., Zhang, E., Wu, H., Yu, S., Zhou, G., Li, S., Zhong, N., Wen, S., Li, X., Knapp, S.J., Ozias-Akins, P., Varshney, R.K. and Liang, X. (2013) Deep sequencing analysis of the transcriptomes of peanut aerial and subterranean young pods identifies candidate genes related to early embryo abortion. *Plant Biotechnol. J.* **11**, 115–127.

Choi, W., Toyota, M., Kim, S., Hilleary, R. and Gilroy, S. (2014) Salt stress-induced Ca2+ waves are associated with rapid, long-distance root-to-shoot signaling in plants. *Proc. Natl Acad. Sci. USA,* **111**, 6497–6502.

Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Dash, P.K., Cao, Y., Jailani, A.K., Gupta, P., Venglat, P., Xiang, D., Rai, R., Sharma, R., Thirunavukkarasu, N., Abdin, M.Z., Yadava, D.K., Singh, N.K., Singh, J., Selvaraj, G., Deyholos, M., Kumar, P.A. and Datla, R. (2014) Genome-wide analysis of drought induced gene expression changes in flax (*Linum usitatissimum*). *GM Crops Food,* **5**, 106–119.

De Felice, B., Wilson, R., Argenziano, C., Kafantaris, I. and Conicella, C. (2009) A transcriptionally active copia-like retroelement in *Citrus limon*. *Cell. Mol. Biol. Lett.* **14**, 289–304.

Devisetty, U.K., Covington, M.F., Tat, A.V., Lekkala, S. and Maloof, J.N. (2014) Polymorphism identification and improved genome annotation of brassica rapa through deep RNA sequencing. *G3: Genes - Genomes - Genetics*, **4**, 2065–2078.

Du, L. and Chen, Z. (2000) Identification of genes encoding receptor-like protein kinases as possible targets of pathogen- and salicylic acid-induced WRKY DNA-binding proteins in Arabidopsis. *Plant J.* **24**, 837–847.

Ecker, J.R. and Davis, R.W. (1987) Plant defense genes are regulated by ethylene. *Proc. Natl Acad. Sci. USA,* **84**, 5202–5206.

Egan, A.N., Schlueter, J. and Spooner, D.M. (2012) Applications of next-generation sequencing in plant biology. *Am. J. Bot.* **99**, 175–185.

Ehlting, J., Sauveplane, V., Olry, A., Ginglinger, J., Provart, N. and Werck-Reichhart, D. (2008) An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in *Arabidopsis thaliana*. *BMC Plant Biol.* **8**, 47.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**, 45–58.

Fukushima, A., Nishizawa, T., Hayakumo, M., Hikosaka, S., Saito, K., Goto, E. and Kusano, M. (2012) Exploring tomato gene functions based on coexpression modules using graph clustering and differential coexpression approaches. *Plant Physiol.* **158**, 1487–1502.

Gómez-Mena, C., de Folter, S., Costa, M.M.R., Angenent, G.C. and Sablowski, R. (2005) Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development*, **132**, 429–438.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186.

Gueta-Dahan, Y., Yaniv, Z., Zilinskas, B.A. and Ben-Hayyim, G. (1997) Salt and oxidative stress: similar and specific responses and their relation to salt tolerance in Citrus. *Planta*, **203**, 460–469.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M. and Buell, C.R. (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS ONE*, **7**, e33071.

Haseneyer, G., Schmutzer, T., Seidel, M., Zhou, R., Mascher, M., Schon, C.C., Taudien, S., Scholz, U., Stein, N., Mayer, K.F. and Bauer, E. (2011) From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.* **11**, 131.

Hichri, I., Barrieu, F., Bogs, J., Kappel, C., Delrot, S. and Lauvergeat, V. (2011) Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *J. Exp. Bot.* **62**, 2465–2483.

Iqbal, N., Trivellini, A., Masood, A., Ferrante, A. and Khan, N.A. (2013) Current understanding on ethylene signaling in plants: the influence of nutrient availability. *Plant Physiol. Biochem.* **73**, 128–138.

Jiao, W.B., Huang, D., Xing, F., Hu, Y., Deng, X.X., Xu, Q. and Chen, L.L. (2013) Genome-wide characterization and expression analysis of genetic variants in sweet orange. *Plant J.* **75**, 954–964.

Jofuku, K.D., den Boer, B.G., Van Montagu, M. and Okamuro, J.K. (1994) Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. *Plant Cell*, **6**, 1211–1225.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R. and Hunter, S. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205.

Kawaguchi, M., Kobayashi, M., Sakurai, A. and Syōno, K. (1991) The presence of an enzyme that converts indole-3-acetamide into IAA in wild and cultivated rice. *Plant Cell Physiol.* **32**, 143–149.

Kelly, G., Sade, N., Attia, Z., Secchi, F., Zwieniecki, M., Holbrook, N.M., Levi, A., Alchanatis, V., Moshelion, M. and Granot, D. (2014) Relationship between hexokinase and the aquaporin PIP1 in the regulation of photosynthesis and plant growth. *PLoS ONE*, **9**, e87888.

Kim, S., Kim, M., Kim, Y., Yeom, S., Cheong, K., Kim, K., Jeon, J., Kim, S., Kim, D., Sohn, S., Lee, Y. and Choi, D. (2014) Integrative structural annotation of de novo RNA-Seq provides an accurate reference gene set of the enormous genome of the onion (*Allium cepa* L.). *DNA Res.* **22**(1): 19–27.

Klee, H.J., Horsch, R.B., Hinchee, M.A., Hein, M.B. and Hoffman, N.L. (1987) The effects of overproduction of two *Agrobacterium tumefaciens* T-DNA auxin biosynthetic gene products in transgenic petunia plants. *Genes Dev.* **1**, 86–96.

Ko, J.H., Kim, H.T., Hwang, I. and Han, K.H. (2012) Tissue-type-specific transcriptome analysis identifies developing xylem-specific promoters in poplar. *Plant Biotechnol. J.* **10**, 587–596.

Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**: 1639–1645.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **25**, 402–408.

Loraine, A.E., McCormick, S., Estrada, A., Patel, K. and Qin, P. (2013) RNA-Seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant Physiol.* **162**, 1092–1109.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Li, W., Huang, X. and Han, B. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20**, 1238–1249.

Luschnig, C. (2001) Auxin transport: why plants like to think BIG. *Curr. Biol.* **11**, R831–R833.

Ma, L., Sun, N., Liu, X., Jiao, Y., Zhao, H. and Deng, X.W. (2005) Organ-specific expression of Arabidopsis genome during development. *Plant Physiol.* **138**, 80–91.

Mano, Y. and Nemoto, K. (2012) The pathway of auxin biosynthesis in plants. *J. Exp. Bot.* **63**, 2853–2872.

Mano, Y., Nemoto, K., Suzuki, M., Seki, H., Fujii, I. and Muranaka, T. (2010) The AMI1 gene family: indole-3-acetamide hydrolase functions in auxin biosynthesis in plants. *J. Exp. Bot.* **61**, 25–32.

Martinelli, F., Uratsu, S.L., Albrecht, U., Reagan, R.L., Phu, M.L., Britton, M., Buffalo, V., Fass, J., Leicht, E., Zhao, W., Lin, D., D'Souza, R., Davis, C.E., Bowman, K.D. and Dandekar, A.M. (2012) Transcriptome profiling of citrus fruit response to huanglongbing disease. *PLoS ONE*, **7**, e38039.

Merchante, C., Alonso, J.M. and Stepanova, A.N. (2013) Ethylene signaling: simple ligand, complex regulation. *Curr. Opin. Plant Biol.* **16**, 554–560.

Misner, I., Bicep, C., Lopez, P., Halary, S., Bapteste, E. and Lane, C.E. (2013) Sequence comparative analysis using networks: software for evaluating de novo transcript assembly from next-generation sequencing. *Mol. Biol. Evol.* **30**, 1975–1986.

Mizuno, H., Kawahara, Y., Sakai, H., Kanamori, H., Wakimoto, H., Yamagata, H., Oono, Y., Wu, J., Ikawa, H., Itoh, T. and Matsumoto, T. (2010) Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (*Oryza sativa* L.). *BMC Genom.* **11**, 683.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods,* **5**, 621–628.

Muday, G.K., Rahman, A. and Binder, B.M. (2012) Auxin and ethylene: collaborators or competitors? *Trends Plant Sci.* **17**, 181–195.

Nakatsuka, A., Murachi, S., Okunishi, H., Shiomi, S., Nakano, R., Kubo, Y. and Inaba, A. (1998) Differential Expression and internal feedback regulation of 1-aminocyclopropane-1-carboxylate synthase, 1-aminocyclopropane-1-carboxylate oxidase, and ethylene receptor genes in tomato fruit during development and ripening. *Plant Physiol.* **118**, 1295–1305.

Nicolosi, E., Deng, Z.N., Gentile, A., La Malfa, S., Continella, G. and Tribulato, E. (2000) *Citrus* phylogeny and genetic origin of important species as investigated by molecular markers. *Theor. Appl. Genet.* **100**, 1155–1166.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033.

Paritosh, K., Yadava, S., Gupta, V., Panjabi-Massand, P., Sodhi, Y., Pradhan, A. and Pental, D. (2013) RNA-seq based SNPs in some agronomically important oleiferous lines of *Brassica rapa* and their use for genome-wide linkage mapping and specific-region fine mapping. *BMC Genom.* **14**, 463.

Patel, M., Manvar, T., Apurwa, S., Ghosh, A., Tiwari, T. and Chikara, S.K. (2014) Comparative de novo transcriptome analysis and metabolic pathway studies of *Citrus paradisi* flavedo from naive stage to ripened stage. *Mol. Biol. Rep.* **41**, 3071–3080.

Pelaz, S., Ditta, G.S., Baumann, E., Wisman, E. and Yanofsky, M.F. (2000) B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature*, **405**, 200–203.

Petroni, K. and Tonelli, C. (2011) Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* **181**, 219–229.

Piskol, R., Ramaswami, G. and Li, J.B. (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodrigues, C., de Souza, A., Takita, M., Kishi, L. and Machado, M. (2013) RNA-Seq analysis of *Citrus reticulata* in the early stages of *Xylella fastidiosa* infection reveals auxin-related genes as a defense response. *BMC Genom.* **14**, 676.

Sarkar, A.K., Luijten, M., Miyashima, S., Lenhard, M., Hashimoto, T., Nakajima, K., Scheres, B., Heidstra, R. and Laux, T. (2007) Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature*, **446**, 811–814.

Sauer, M., Robert, S. and Kleine-Vehn, J. (2013) Auxin: simply complicated. *J. Exp. Bot.* **64**, 2565–2577.

Scaglione, D., Lanteri, S., Acquadro, A., Lai, Z., Knapp, S.J., Rieseberg, L. and Portis, E. (2012) Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotechnol. J.* **10**, 956–969.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Severin, A.J., Peiffer, G.A., Xu, W.W., Hyten, D.L., Bucciarelli, B., O'Rourke, J.A., Bolon, Y., Grant, D., Farmer, A.D., May, G.D., Vance, C.P., Shoemaker, R.C. and Stupar, R.M. (2010) An integrative approach to genomic introgression mapping. *Plant Physiol.* **154**, 3–12.

Shalom, L., Samuels, S., Zur, N., Shlizerman, L., Doron-Faigenboim, A., Blumwald, E. and Sadka, A. (2014) Fruit load induces changes in global gene expression and in abscisic acid (ABA) and indole acetic acid (IAA) homeostasis in citrus buds. *J. Exp. Bot.* **65**(12): 3029–44.

Shin, L., Lo, J. and Yeh, K. (2012) Copper chaperone antioxidant protein1 is essential for copper homeostasis. *Plant Physiol.* **159**, 1099–1110.

Singh, V.K., Garg, R. and Jain, M. (2013) A global view of transcriptome dynamics during flower development in chickpea by deep sequencing. *Plant Biotechnol. J.* **11**, 691–701.

Smith, A.P., Nourizadeh, S.D., Peer, W.A., Xu, J., Bandyopadhyay, A., Murphy, A.S. and Goldsbrough, P.B. (2003) Arabidopsis AtGSTF2 is regulated by ethylene and auxin, and encodes a glutathione S-transferase that interacts with flavonoids. *Plant J.* **36**, 433–442.

Takahashi, N., Yamaguchi, I., Kono, T., Igoshi, M., Hirose, K. and Suzuki, K. (1975) Characterization of plant-growth substances in *Citrus unshiu* and their change in fruit development. *Plant Cell Physiol.* **16**, 1101–1111.

Taylor, L.P. and Grotewold, E. (2005) Flavonoids as developmental regulators. *Curr. Opin. Plant Biol.* **8**, 317–323.

Terol, J., Ibanez, V., Carbonell, J., Alonso, R., Estornell, L.H., Licciardello, C., Gut, I.G., Dopazo, J. and Talon, M. (2015) Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation. *BMC Genom.* **16**, 69.

Treutter, D. (2005) Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol (Stuttg)* **7**, 581–591.

Valverde, F. (2011) CONSTANS and the evolutionary origin of photoperiodic timing of flowering. *J. Exp. Bot.* **62**, 2453–2463.

Venu, R.C., Sreerekha, M.V., Nobuta, K., Belo, A., Ning, Y., An, G., Meyers, B. and Wang, G. (2011) Deep sequencing reveals the complex and coordinated transcriptional regulation of genes related to grain quality in rice cultivars. *BMC Genom.* **12**, 190.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13**, 1998–2004.

Wang, K.L., Li, H. and Ecker, J.R. (2002) Ethylene biosynthesis and signaling networks. *Plant Cell*, **14**(Suppl), S131–S151.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

Wang, L., Zhao, S., Gu, C., Zhou, Y., Zhou, H., Ma, J., Cheng, J. and Han, Y. (2013) Deep RNA-Seq uncovers the peach transcriptome landscape. *Plant Mol. Biol.* **83**, 365–377.

Wu, G.A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., Perrier, X., Ruiz, M., Scalabrin, S., Terol, J., Takita, M.A., Labadie, K., Poulain, J., Couloux, A., Jabbari, K., Cattonaro, F., Del Fabbro, C., Pinosio, S., Zuccolo, A., Chapman, J., Grimwood, J., Tadeo, F.R., Estornell, L.H., Munoz-Sanz, J., Ibanez, V., Herrero-Ortega, A., Aleza, P., Perez-Perez, J., Ramon, D., Brunel, D., Luro, F., Chen, C., Farmerie, W.G., Desany, B., Kodira, C., Mohiuddin, M., Harkins, T., Fredrikson, K., Burns, P., Lomsadze, A., Borodovsky, M., Reforgiato, G., Freitas-Astua, J., Quetier, F., Navarro, L., Roose, M., Wincker, P., Schmutz, J., Morgante, M., Machado, M.A., Talon, M., Jaillon, O., Ollitrault, P., Gmitter, F. and Rokhsar, D. (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotech.* **32**, 597–698.

Xu, Q., Chen, L.L., Ruan, X., Chen, D., Zhu, A., Chen, C., Bertrand, D., Jiao, W.B., Hao, B.H., Lyon, M.P., Chen, J., Gao, S., Xing, F., Lan, H., Chang, J.W., Ge, X., Lei, Y., Hu, Q., Miao, Y., Wang, L., Xiao, S., Biswas, M.K., Zeng, W., Guo, F., Cao, H., Yang, X., Xu, X.W., Cheng, Y.J., Xu, J., Liu, J.H., Luo, O.J., Tang, Z., Guo, W.W., Kuang, H., Zhang, H.Y., Roose, M.L., Nagarajan, N., Deng, X.X. and Ruan, Y. (2013) The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66.

Yates, S., Swain, M., Hegarty, M., Chernukin, I., Lowe, M., Allison, G., Ruttink, T., Abberton, M., Jenkins, G. and Skot, L. (2014) De novo assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. *BMC Genom.* **15**, 453.

Yu, K., Xu, Q., Da, X., Guo, F., Ding, Y. and Deng, X. (2012) Transcriptome changes during fruit development and ripening of sweet orange (*Citrus sinensis*). *BMC Genom.* **13**, 10.

Yun, Z., Jin, S., Ding, Y., Wang, Z., Gao, H., Pan, Z., Xu, J., Cheng, Y. and Deng, X. (2012) Comparative transcriptomics and proteomics analysis of citrus fruit, to improve understanding of the effect of low temperature on maintaining fruit quality during lengthy post-harvest storage. *J. Exp. Bot.* **63**, 2873–2893.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 PCA analysis carried out with all samples grouped by species.

Figure S2 qRT-PCR validation of 10 differentially expressed genes.

Figure S3 Phylogenetic tree showing the flavonol synthase putative homolog proteins in citrus.

Figure S4 Expression analysis of the flavonoid and flavonol biosynthesis pathway.

Figure S5 Expression analysis of the ethylene biosynthesis route.

Figure S6 Expression analysis of the IAA biosynthesis route.

Figure S7 Snapshot from the IGV genome browser showing a small region from chromosome 4 with RNA-Seq reads from *Citrus aurantium*, *Citrus clementina*, *Citrus maxima*, and *Citrus sinensis*, with reads in both strands, revealing mistakes in the Sanger sequencing. Positions 25 612 752 and 25 612 754 displaying G and A in all the RNA-Seq samples (including those not shown here), indicate that the reference sequence (shown below) with C and G is wrong.

Table S1 Read mapping summary.

Table S2 Increase of the number of genes associated to a GO term.

Table S3 Expression (RPKM) of Transcription Factor genes in different tissues.

Table S4 Genes used in qRT-PCR experiments to validate the differential expression analyses carried out with the RNA-Seq data.