Title
Annotation of the Corymbia terpene synthase gene family shows broad conservation but dynamic evolution of physical clusters relative to Eucalyptus.

Journal
Heredity, 121(1)

ISSN
0018-067X

Authors
Butler, Jakob B
Freeman, Jules S
Potts, Brad M
et al.

Publication Date
2018-07-01

DOI
10.1038/s41437-018-0058-1

Peer reviewed

Annotation of the *Corymbia* terpene synthase gene family shows broad conservation but dynamic evolution of physical clusters relative to *Eucalyptus*

Jakob B. Butler[1*], Jules S Freeman[1], Brad M. Potts[1,2], René E. Vaillancourt[1,2], Dario Grattapaglia[3], Orzenil B. Silva-Junior[3], Blake A. Simmons[4], Adam L. Healey[4], Jeremy Schmutz[5,6], Kerrie W. Barry[6], David J. Lee[7], Robert J. Henry[8], Graham J. King[9], Abdul Baten[9] and Mervyn Shepherd[9]

[1]School of Biological Sciences, University of Tasmania, Hobart, TAS 7001, Australia

[2]ARC Training Centre for Forest Value, University of Tasmania, Hobart, TAS 7001, Australia

[3]EMBRAPA Genetic Resources and Biotechnology, EPqB Final W5 Norte 70770-917, Brasilia, Brazil

[4]DOE Joint Bioenergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[5]Hudson-Alpha Institute for Biotechnology, Huntsville, AL, USA

[6]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA, USA

[7]Forest Industries Research Centre, University of the Sunshine Coast, Maroochydore DC, QLD 4558, Australia

[8]University of Queensland/QAAFI, Brisbane 4072, Australia

[9]Southern Cross Plant Science, Southern Cross University, Lismore, NSW 2480, Australia

*Author for correspondence: Jakob Butler, School of Biological Science, University of Tasmania, Private Bag 55, Hobart, TAS 7001, Australia. Ph. +61 (03) 6226 1826 Email: Jakob.Butler@utas.edu.au

Running title: Evolution of the *TPS* gene family in eucalypts

Word count (Introduction, Material and Methods, Results, Discussion): 7609

ABSTRACT

Terpenes are economically and ecologically important phytochemicals. Their synthesis is controlled by the terpene synthase (TPS) gene family, which is highly diversified throughout the plant kingdom. The plant family Myrtaceae are characterised by especially high terpene concentrations, and considerable variation in terpene profiles. Many Myrtaceae are grown commercially for terpene products including the eucalypts *Corymbia* and *Eucalyptus*. *Eucalyptus grandis* has the largest *TPS* gene family of plants currently sequenced, which is largely conserved in the closely related *E. globulus*. However, the *TPS* gene family has been well studied only in these two eucalypt species. The recent assembly of two *Corymbia citriodora* subsp. *variegata* genomes presents an opportunity to examine the conservation of this important gene family across more divergent eucalypt lineages**.** Manual annotation of the *TPS* gene family in *C. citriodora* subsp. *variegata* revealed a similar overall number, and relative subfamily representation, to that previously reported in *E. grandis* and *E. globulus.* Many of the *TPS* genes were in physical clusters that varied considerably between *Eucalyptus* and *Corymbia*, with several instances of translocation, expansion/contraction and loss. Notably, there was greater conservation in the subfamilies involved in primary metabolism than those involved in secondary metabolism, likely reflecting different selective constraints. The variation in cluster size within subfamilies and the broad conservation between the eucalypts in the face of this variation are discussed, highlighting the potential contribution of selection, concerted evolution and stochastic processes. These findings provide the foundation to better understand terpene evolution within the ecologically and economically important Myrtaceae.

Keywords: Terpene synthase, gene family, annotation, tandem duplications, *Corymbia*, *Eucalyptus*

INTRODUCTION

Terpenes are an extensive group of hydrocarbon-based compounds present in most plants, with thousands currently characterised (Padovan et al. 2014). While some terpenes are present in essentially all plants as primary metabolites, such as gibberellin or abscisic acid (Chen et al. 2011), many are secondary metabolites. Correspondingly, there is wide variation in the terpenes produced across different species, in line with their role in modulating diverse interactions between plants and their environment (Keszei et al. 2010a). Along with regulating growth and other developmental processes (Chen et al. 2011), terpenes play roles in pollinator attraction (Pichersky and Gershenzon 2002), chemical and physical barriers to herbivory (Lawler et al. 1999; O'Reilly-Wapstra et al. 2004; Heiling et al. 2010), and thermotolerance (Peñuelas et al. 2005), to name a few. Terpenes are also important economically due to their utilization as pharmaceuticals, industrial materials and biofuel precursors, as well as their direct impact on the fragrance and flavour of horticultural food products such as apples and wine (Schwab et al. 2013).

These varied terpenoid products are created by terpene synthase (TPS) enzymes. TPS enzymes synthesize terpenoid products from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), which are both created by the action of the mevalonic acid (MEV) pathway operating in the cytosol and the methylerythritol phosphate (MEP) pathway operating in the plastids (Chen et al. 2011). Extensive study in many plant species (Aubourg et al. 2002; Martin et al. 2010; Xiong et al. 2016; Hansen et al. 2017) has revealed that the *TPS* gene family is generally a mid-size family, with gene numbers ranging from 1 in *Physcomitrella patens* to 113 in *Eucalyptus grandis*. Previous phylogenetic analyses of the *TPS* gene family have revealed eight different subfamilies, designated *TPS-a* through to *TPS-h*. Each subfamily influences the synthesis of different terpenoid products, with the genes in the subfamilies *TPS-a* (sesqui-terpene), *TPS-b* and *TPS-g* (cyclic/acyclic mono-terpene), *TPS-c* and *TPS-e* (copalyl diphosphate, ent-kaurene, and di-, mono- and sesqui-terpene) and *TPS-f* (ent-kaurene and di-, mono- and sesqui-terpene) categorized by the structurally

70 distinct compounds they synthesize. Subfamilies *TPS-c*, *-e* and *-f* are predominantly involved in the
71 synthesis of primary metabolites such as gibberellin and abscisic acid, while subfamilies *TPS-a, TPS-b*
72 and *TPS-g* generally synthesize secondary metabolites including cineole and citronellal (Chen et al.
73 2011). The representation of *TPS* subfamilies is quite different across taxa; *TPS-d* and *TPS-h*
74 subfamilies, for example, are specific to gymnosperms and *Selaginella* spp., respectively (Chen et al.
75 2011). Given the large amount of variation in terpenoid profiles within and between taxa as well as
76 the economic and evolutionary importance of terpenoid products (Keszei et al. 2008; Schwab et al.
77 2013), it is no surprise there is an extensive body of literature exploring these compounds. Further
78 investigation of the gene family underlying terpenes in different taxa will greatly contribute to
79 understanding how this diversity arises.

80 The Myrtaceae, of the order Myrtales, are a family of plants that exhibit some of the highest
81 concentrations and diversity of foliar terpenes in plants. Across the Myrtaceae, hundreds of
82 compounds have been characterised (Padovan et al. 2014), with the foliage of individual trees often
83 containing over 40 identifiable compounds (Keszei et al. 2008). Due to these features many
84 Myrtaceous genera are key resources for commercial industries utilizing terpenes as essential oils
85 (Padovan et al. 2014) including *Melaleuca* (Keszei et al. 2010b), *Leptospermum* (Douglas et al. 2004),
86 *Eucalyptus and Corymbia* (Batish et al. 2008). Along with *Angophora*, *Eucalyptus* and *Corymbia* are
87 broadly classified as eucalypts (Slee et al. 2006). Eucalypts are the dominant trees in most Australian
88 native forest and the predominant hardwood plantation species in Australia and overseas, due to
89 their importance to the pulp, biofuel and timber industries (Rockwood et al. 2008; Shepherd et al.
90 2011). The characteristic smell of the eucalypts is due to their especially high concentration of foliar
91 terpenes, with the high diversity of compounds present in this foliage extensively studied (Ammon
92 et al. 1985; Lawler et al. 1998; Asante et al. 2001; Keszei et al. 2008). While the terpenoid profiles of
93 most eucalypts are dominated by α-pinene and 1,8-cineole (Keszei et al. 2010a; Padovan et al.
94 2014), chemotype variation is important both to plant ecology (O'Reilly-Wapstra et al. 2004; Keszei
95 et al. 2010a) and the essential oil industry. Analysis of the *Eucalyptus grandis* reference genome

(Myburg et al. 2014) revealed that this variability is accompanied by the largest number of *TPS* genes of any plant yet sequenced, closely followed by *Eucalyptus globulus* (Külheim et al. 2015). These genes often occur in duplicate arrays or physical clusters which are prone to relatively rapid expansion and contraction (Hanada et al. 2008). Given the most likely fate of duplicate genes is degeneration (Lynch and Conery 2000), the large number of genes present in these eucalypts suggests natural selection preserved these expansions, resulting in high variability of terpene products. Indeed, the combinations of terpenes present in eucalypts varies both between and within species (Keszei et al. 2008; O'Reilly-Wapstra et al. 2011) and within individuals (Padovan et al. 2012), in line with the diverse roles these compounds play in responding to ecological variation. Specific comparison of *E. grandis* and *E. globulus* also suggests that most *TPS* genes evolved prior to the divergence of these species, approximately 12 million years ago (MYA), but points to ongoing evolution as indicated by novel gene duplication, degeneration and gene loss (Külheim et al. 2015). Although this gene family has been well categorized in *E. globulus* and *E. grandis*, the extent to which it is conserved in other eucalypt lineages is currently unknown.

The genus *Corymbia* is predominantly endemic to the tropical, arid, and semi-arid zones of northern Australia (Hill and Johnson 1995; Ladiges et al. 2003), but is increasingly cultivated for forestry and essential oil production in Australia, India, Brazil, Fiji and South Africa (Asante et al. 2001; Vernin et al. 2004). It is a sister genus to *Eucalyptus* (Lee 2007), which diverged from a common ancestor approximately 52 MYA (Crisp et al. 2011; Thornhill et al. 2015). All eucalypts share the same haploid chromosome number (n = 11), which is highly conserved across most Myrtaceous species (Grattapaglia et al. 2012). However, in comparison to *E. grandis, Corymbia citriodora* subsp. *variegata* (hereafter referred to as CCV) has both a smaller genome size (370 MB vs 640 MB, Grattapaglia and Bradshaw Jr 1994) and several major differences in chromosome structure (Butler et al. 2017). The recent *de novo* genome assemblies for two CCV genotypes (Shepherd et al. 2015) provides the opportunity for comparison of individual loci and gene families.

In this study we annotate the *terpene synthase* gene family in the CCV reference genome and compare it to other plants (*Vitis, Populus, Arabidopsis*), but focus on the comparison of *Corymbia* with *E. grandis* and *E. globulus* (Külheim et al. 2015). We present evidence for broad conservation in this gene family across eucalypt lineages along with extensive variation within subfamilies in terms of the presence of specific clusters, and the number of genes contained within them. These results are discussed in the context of their evolutionary and ecological importance.

MATERIAL AND METHODS

Terpene synthase gene discovery

Initially, a CoGeBLAST (Lyons et al. 2008) search for *TPS* genes was performed on the CCV reference genome v1.1 (CCV18, Healey et al. 2017), based on conserved domains from all *TPS* subfamilies following Külheim et al. (2015). A preliminary list of putative *TPS* genes was created based on hits with high similarity (e-value < $1e^{-08}$). To identify if these preliminary hits were full length genes, the genomic regions surrounding each BLAST hit (+- 5 000 bp) were used in reverse BLAST searches against the non-redundant database at Genbank (http://www.ncbi.nlm.nih.gov, accessed 23/02/2017). The closest matching *TPS* gene from *E. grandis, E. globulus, Arabidopsis thaliana, Populus trichocarpa* or *Vitis vinifera* was compared to the putative *TPS* sequence using GeneWise (Birney and Durbin 2000), to determine exon-intron borders and reveal reading frame shifts or premature stop codons. A partial genome assembly from a different CCV individual (1CCV2-054) was also mined for *TPS* genes and, where possible, used to validate the results from the CCV18 genome assembly (Healey et al. 2017).

Phylogenetic analysis and annotation

The amino acid sequence of all putative CCV *TPS* genes were aligned using ClustalW along with those from *E. grandis, E. globulus, A. thaliana, P. trichocarpa* and *V. vinifera* (Külheim et al. 2015). Due to high levels of variation and variable exon counts between taxa the alignment was

trimmed to focus on regions conserved among all genes (positions in the alignment with >75% gap representation were removed), allowing a direct comparison with the results of Külheim et al. (2015). The phylogeny of the *TPS* family in these six organisms was determined using IQTREE (Nguyen et al. 2015) with 1 000 ultrafast bootstrap replicates (Minh et al. 2013). The JTT amino acid substitution model with estimation of invariable sites and gamma distribution was used as this model created the tree with the highest AICc value (corrected Akaike's information criterion) using the program SMS (Lefort et al. 2017). CCV putative *TPS* genes were sorted into the subfamilies *TPS-a, -b, -c, -e, -f* and *-g* based on sequence similarity to *TPS* genes previously classified in the other species. These genes were sorted by chromosome and by position within chromosome in the CCV reference genome, and annotated from the first *TPS-a* gene (*CorciTPS001*) to the final *TPS-g* gene (*CorciTPS102*).

Gene birth/death rates were estimated using the program Badirate (Librado et al. 2012). The BD-FR-CML model was used with the family option, which allows for a free turnover rate for each branch of the species tree, with the gain/loss events of internal nodes inferred by maximum likelihood and informed by the relative representation of each subfamily (Librado et al. 2012). This was performed for both the tree of the six species (with divergence times taken from Wikström et al. (2001)), and for the eucalypts alone. To improve the accuracy of the rate estimation in the latter analysis of the eucalypts, *TPS* subfamilies were further divided into their component orthologous groups before analysis, which were defined as the most inclusive clade of the gene tree compatible with the species tree.

RNA-Seq expression analysis

To examine the expression of putative functional genes, RNA sequencing was undertaken using mRNA isolated from 5 tissue types: flower initials, flower buds, bark, expanded leaf, and unexpanded leaf. Tissue was obtained from 1CCV2-054 (sequenced for the CCV54 assembly), and RNA extracted using Ambion RNAquenous kit with Ambion RNA Isolation aid and the standard

protocol (Life Technologies Australia, Mulgrave Vic). Total RNA was shipped to AGRF (Melbourne, Australia) for library preparation (TruSeq Stranded mRNA Sample, Illumina) and sequencing (HiSeq HT chemistry single read 50/100, Illumina). A total of 75 GB of sequence data was generated across all five libraries: 25 GB of 100 bp single-end reads, and 50 GB of 100 bp paired-end reads. Reads were quality controlled using BBMap tools (Bushnell 2016), and assembled into transcripts using Trinity de-novo RNA-Seq assembly pipeline (Haas et al. 2013). Transcripts were aligned to the CCV reference genome using CoGe's RNA-Seq analysis pipeline (Lyons and Freeling 2008). Detectable expression at the location of putative functional and pseudogenes was a criteria used to support the existence of putative genes. The clustering of gene expression was examined using the complete linkage method and Euclidean distance measures contained within the package 'gplots' (Warnes et al. 2016) in R (R Core Team 2017), allowing clusters to be identified based on dendogram structure.

Comparative analysis of the *TPS* gene family between species

To examine differences in genome organization and gene number in specific *TPS* clusters, the positions of *TPS* genes in the CCV and *E. grandis* genomes were collated and assigned to specific physical clusters. A physical cluster of *TPS* genes was defined as genes from the same subfamily occurring on the same chromosome, with further support for gene clusters based on close phylogenetic relationships. Homologous clusters were matched, requiring both close phylogenetic relationships between *TPS* genes and similar genomic position in each genome assembly. Homologous clusters that were both syntenic (located on the same chromosome) and matched the approximate position within that chromosome in both species were examined for copy number variation. *TPS* genes in the CCV54 assembly were also assigned to physical clusters and compared to the CCV18 reference genome to determine if there were any changes in copy number.

In cases where gene clusters in the CCV reference genome were placed on a different chromosome to their apparent homolog in *E. grandis* (evaluated by phylogenetic relatedness), verification of their position was undertaken in CCV54. The tool SYNFIND in CoGe (Lyons and

195 Freeling 2008) was used to determine the likely position of homologous genes in CCV54 taking into
196 account the synteny of the surrounding region. If gene position was conserved across both CCV
197 genome assemblies, movement of loci relative to *E grandis* was considered real, while
198 disagreements between the CCV assemblies were flagged as possible errors caused by misassembly,
199 with more weight given to the loci position mirroring that of *E. grandis*.

200 As the CCV genome assemblies (Healey et al. 2017) are anchored to linkage maps (Butler et
201 al. 2017), it is possible that markers on these maps may be mis-ordered, leading to incorrect contig
202 positioning and potentially incorrect conclusions on loci position and movement. To examine this,
203 the number of markers used to anchor and orient each contig housing *TPS* loci with putative
204 movement was used to determine the strength of contig placement.

205 **RESULTS**

206 **Discovery of *TPS* loci**

207 In the *Corymbia citriodora* subsp. *variegata* reference genome (CCV18) 127 loci were
208 discovered with high sequence similarity to terpene synthase (*TPS*) genes from other species. Using
209 a modified version of the classification method of Külheim et al. (2015), loci were classified into
210 three categories: (i) 64 were full length with no structural abnormalities and had evidence of
211 expression; (ii) 17 were full length, expressed but with up to two frame shifts or premature stop
212 codons; and (iii) 21 were full length, had no evidence of expression and up to two frame shifts or
213 premature stop codons. In accordance with Külheim et al. (2015), these were considered putatively
214 functional *TPS* genes, resulting in a total of 102 genes (Table 1, Table S1) used in further analysis. The
215 remaining 25 loci were classified as pseudogenes with more than two frame shifts or premature stop
216 codons, with no consideration given to expression (Table S2). Similar analysis of the partially
217 assembled CCV54 without expression data revealed 69 putative functional *TPS* genes and seven
218 pseudogenes (Table S3).

Table 1 Copy numbers of *TPS* genes by subfamily in various plant species

Subfamily	*C. citriodora* subsp. *variegata* (CCV)	*E. grandis*	*E. globulus*	*V. vinifera*	*A. thaliana*	*P. trichocarpa*
TPS-a	51 (2)	52	45	29	23	13
TPS-b1	26 (1)	27	28	8	6	10
TPS-b2	10 (1)	9	10	2	0	2
TPS-c	1 (1)	2	2	2	1	2
TPS-e	1 (1)	3	2	1	1	2
TPS-f	4 (0)	7	9	0	1	1
TPS-g	9 (0)	13	10	15	1	2
Total	102	113	106	57	33	32

Table adapted from (Külheim et al. 2015). Numbers in brackets indicate the number of orthologous pairs between *C. citriodora* subsp. *variegata* (CCV) and *E. grandis.* See Figure 1 for examples of orthologous pairs.

Phylogenetic analysis

The phylogenies presented show the relationship between the CCV18 *TPS* genes and those from *E. globulus, E. grandis, V. vinifera, P. trichocarpa,* and *A. thaliana*, divided into *TPS-a* (Figure 1), *TPS-b* and *TPS-g* (Figure 2) and *TPS-c, TPS-e* and *TPS-f* (Figure 3) subfamilies. The same *TPS* subfamilies were represented in each eucalypt species. Orthology (genes in different species directly descended from the same ancestral gene) between *TPS* genes in *E grandis* and *E. globulus* was common, with 60% of genes found in orthologous pairs (defined as a single gene in one species more closely related to a single gene in a different species than to a gene within its own genome, see Figure 1 for examples). However, only 9% of *TPS* genes in CCV were orthologous with pairs from the other eucalypts.

The *TPS-a* subfamily was represented by the most genes in CCV, as was the case in *E. grandis* and *E. globulus* (Table 1). However, specific *TPS-a* clades in CCV were expanded relative to the other eucalypts (for example, the clade containing *CorciTPS035* [Figure 1-a]), or missing entirely (for example, the clade containing *EgranTPS029* [Figure 1-b]). An interesting orthologous relationship

was seen between an *E. globulus TPS* gene (*EglobTPS022*) and a clade of CCV *TPS* genes with no specific *E. grandis* ortholog, suggesting this gene was lost or not found in *E. grandis* (Figure 1-c). While 31 of the *TPS-a* genes in *E. grandis* (60% of total *TPS-a* genes) and *E. globulus* (69%) were in orthologous pairs, greater divergence was evident in CCV as only two *TPS-a* genes (4%) were in orthologous pairs with other eucalypt genes (specifically *CorciTPS025* and *CorciTPS026*; Figure 1).

As seen in the *TPS-a* subfamily, the *TPS-b* and *TPS-g* subfamilies also provided evidence for expansion and contraction of physical clusters as well as loss of loci among the eucalypts (Figure 2). Only one *TPS-b1* gene (*CorciTPS053*) in CCV (4% of the total) was in an orthologous pair with the other eucalypts. In contrast, 19 of the *TPS-b1* genes in *E. grandis* (70%) and *E. globulus* (68%) occurred in orthologous pairs. Another potential gene loss in *E. grandis* was seen in the clade containing *EglobTPS077* and multiple CCV genes (Figure 2-a). Of the *TPS-b2* genes (Figure 2), five were in orthologous pairs between *E. grandis* (55%) and *E. globulus* (50%), while in CCV only one (10%) was orthologous to the other eucalypts. The remainder of the genes were arranged in clades specific to each eucalypt with no orthologous pairing (Figure 2-b, 2-c). In the *TPS-g* subfamily, six genes in *E. grandis* (46%) and *E. globulus* (60%) were found in orthologous pairs, but no orthologous pairs were found between CCV and the other eucalypts.

The *TPS-c* and *TPS-e* subfamilies, involved in the synthesis of primary metabolites, were generally conserved between the eucalypts (Figure 3). The single *TPS-c* gene in CCV was found in an orthologous pair with both other eucalypts, while a second orthologous pair was found between *E. grandis* and *E. globulus*. An identical situation was observed in the *TPS-e* subfamily, with the single gene in CCV paired with the two *Eucalyptus* species, and a second orthologous pair between *E. grandis* and *E. globulus.* In both cases, a second *TPS-c* and *TPS-e* gene was found in the CCV54 assembly in the minor scaffolds (contigs that were assembled into scaffolds but not anchored to the 11 chromosomes), suggesting the corresponding genes may be missing from the CCV18 assembly (although the possibility that the minor scaffolds represent alternate haplotypes which did not fuse

to the chromosomes cannot be dismissed). Both of these subfamilies are highly conserved in *A. thaliana, V. vinifera* and *P. trichocarpa*, as each only has 1-2 genes of each subfamily (Figure 3).

The *TPS-f* subfamily was more dynamic than the other subfamilies involved in primary metabolism (Figure 3). The orthologous pairings seen in this clade differed somewhat to those presented by Külheim et al. (2015), likely influenced by low bootstrap support in both studies, slight differences in methodology and the addition of CCV weakening support for previous clade structure. In our analysis, only two of the *E. grandis* (29%) and *E. globulus* (22%) *TPS-f* loci were in orthologous pairs, while a single *TPS-f* loci was directly orthologous between CCV (25%) and *E. globulus* (*CorciTPS092* and *EglobTPS121*), with no gene from *E. grandis* present. In contrast to *TPS-c* and *TPS-e*, *A. thaliana* and *P. trichocarpa* only have a single *TPS-f* gene, while no *TPS-f* was found in *V. vinifera* (Table 1).

The estimated gene birth rate in the *TPS* gene family was negligible ($\leqslant$ 0.0002 events/gene/million years [e/g/my]) for *A. thaliana, V. vinifera* and *P. trichocarpa*, while the death rate ranged from 0.0016 - 0.0031 e/g/my (Figure S1-a). In contrast, the eucalypt lineage was estimated to have experienced a magnitude higher rate of gene birth (0.0282 e/g/my). Within the eucalypt lineages, death rate was similar in both *Eucalyptus* and *Corymbia* (0.0063 - 0.0071 e/g/my, Figure S1-b). However, the gene birth rate in *E. grandis* (0.0125 e/g/my, since divergence from *E. globulus*) was seven times higher than the estimated birth rate in CCV (0.0018 e/g/my).

Proportional representation and genome organisation of *TPS* genes

There were no significant differences in subfamily representation (the proportion of genes in each subfamily) between *E. grandis* and CCV (χ^2_4=3.69, P>0.05 [combining *TPS-c*, *-e*, and *-f* due to sample size]), or the number of genes involved in primary *versus* secondary metabolism (χ^2_1=2.41, P>0.05). A similar lack of significant difference was observed between *E. grandis* and *E. globulus* in the number of loci at the subfamily (χ^2_4=1.53, P>0.05) or primary *versus* secondary metabolite

286 (χ^2_1=0.3, P>0.05) levels, providing evidence that the broad features of this gene family are conserved
287 between *Eucalyptus* and *Corymbia*.

288 Seventy-five putative functional *TPS* genes were found across all 11 chromosomes (74% of
289 the total) in the CCV18 genome assembly, with 27 genes found within minor scaffolds (26%, Table
290 S1). In comparison, 97 and 16 *TPS* genes were found on chromosomes (86%) and minor scaffolds
291 (14%), respectively, in *E. grandis.* The relative proportion of genes located on the main
292 chromosomes in each species is consistent with the estimated completeness of each assembly
293 (Myburg et al. 2014; Healey et al. 2017). In the CCV genome, *TPS* genes were often arranged in
294 physical clusters with genes from only one subfamily, as was seen in *E. grandis* (Külheim et al. 2015).
295 On average, there were 3.7 *TPS* genes per cluster in CCV (only considering those on chromosomes),
296 while *E. grandis* averaged 5.1 per cluster. This difference may reflect the greater proportion of *TPS*
297 genes on minor scaffolds in CCV compared to *E. grandis*. In CCV these clusters occurred in true
298 tandem arrays (no intervening genes between putative *TPS* genes), localised clusters with other
299 genes contained within and combinations of the two (Table 2, Table S4).

300
301
302
303
304
305
306
307
308
309

Table 2 Structure of the *Corymbia citriodora* subsp. *variegata* (CCV) terpene synthase physical clusters and the *Eucalyptus grandis* clusters which are syntenic to CCV

Species[a]	Subfamily	Chromosome	*TPS* genes	Position (bp)	Span (bp)	Intervening genes	Internal clustering[b]
CCV	*TPS-a*	Chr3	9	29,395,680	3,598,941	202	1(9)2(1)1(22)3(2)1(168)1
Egr	*TPS-a*	Chr3	3	47,928,331	1,606,854	64	1(2)1(62)1
CCV	*TPS-a*	Chr4	3	11,823,925	14,705	0	N/A
CCV	*TPS-a*	Chr4	2	18,100,282	23,876	0	N/A
Egr	*TPS-a*	Chr4	5	19,896,024	246,197	7	1(2)3(5)1
CCV	*TPS-a*	Chr5	4	16,525,672	1,219,850	68	1(68)3
CCV	*TPS-a*	Chr6	6	33,461,887	106,929	0	N/A
Egr	*TPS-a*	Chr6	10	42,991,263	321,452	0	N/A
CCV	*TPS-a*	Chr7	3	2,183,056	809,251	59	1(58)1(1)1
CCV	*TPS-a*	Chr7	2	14,463,251	1,761,379	105	1(105)1
CCV	*TPS-b1*	Chr1	8	22,285,470	1,866,767	103	1(64)2(1)2(36)1(1)1(1)1
Egr	*TPS-b1*	Chr1	7	17,720,921	1,286,126	50	1(3)4(1)1(46)1
CCV	*TPS-b1*	Chr2	2	2,027,307	13,692	0	N/A
CCV	*TPS-b1*	Chr4	2	8,861,360	169,636	5	1(5)1
Egr	*TPS-b1*	Chr4	8	16,009,931	217,347	7	1(1)1(1)2(4)2(1)2
CCV	*TPS-b1*	Chr5	3	41,952,144	3,787,929	215	1(192)1(23)1
CCV	*TPS-b1*	Chr8	3	29,799,505	47,398	3	1(1)1(2)1
CCV	*TPS-b2*	Chr10	3	13,605,402	38,144	1	2(1)1
CCV	*TPS-b2*	Chr11	5	23,810,145	420,788	25	2(25)3
Egr	*TPS-b2*	Chr11	9	10,288,308	1,164,219	33	3(31)1(1)4(1)1
CCV	*TPS-f*	Chr4	3	7,135,889	385,574	31	1(31)2
Egr	*TPS-f*	Chr4	7	12,270,273	287,241	2	3(2)4
CCV	*TPS-g*	Chr2	2	19,657,471	29,412	2	1(2)1
CCV	*TPS-g*	Chr5	3	44,341,209	100,371	2	1(2)2
Egr	*TPS-g*	Chr5	12	62,540,499	1,272,677	25	3(2)3(2)1(19)1(1)1(1)3

[a]Egr indicate *TPS* clusters from *E. grandis* syntenic with the CCV *TPS* cluster directly above. Loci with only a single *TPS* gene are not shown.

[b]Structure of the gene cluster, with non-*TPS* genes indicated by brackets. N/A indicates no intervening genes.

Across all *TPS* subfamilies, 10 physical clusters were both syntenic and phylogenetically similar between *E. grandis* and CCV18 (Figure 4). These clusters were assumed to be homologous between these species and were examined for copy number variation. *E globulus* was not examined due to the lack of an assembled genome. There was no significant correlation between gene number in syntenic homologous clusters between species (Spearman's r_8=0.29, P>0.10), suggesting independent expansion or contraction has occurred between *E. grandis* and CCV. Seven clusters

were homologous but non-syntenic, with the chromosome assignment of three non-syntenic clusters in the CCV reference genome supported by the second CCV genome assembly CCV54 (Figure 4, Table S5). The position of the single *TPC-c* gene conflicted between the CCV assemblies (despite both CCV18 and CCV54 [not shown] having contig-marker support for placement), potentially due to assembly error in one or the other. The general placement of clusters was supported by examining the markers in the linkage maps used to aid genome assembly. Contigs were anchored to their map position by an average of ten markers, with only three contigs not supported by at least three markers (Table S6), providing support for their correct placement and therefore the non-syntenic nature of the *TPS* clusters.

Gene structure in the *TPS-a*, *-b* and *-g* subfamilies (involved in secondary metabolite synthesis) was highly conserved (Figure 5), with most having seven exons, and only a small proportion departing from this structure with between four and six exons. The conserved catalytic motif DDxxD (Hosfield et al. 2004; Gao et al. 2012) was generally located on the fourth exon, similar to *E. grandis* (Külheim et al. 2015). The placement of this motif on different exons was always associated with uncommon exon number. The genes from *TPS* subfamilies *-c*, *-e*, and *-f* (involved in primary metabolite synthesis) had between 10 and 13 exons, with the exception of *CorciTPS092* with six exons. The DDxxD motif in these subfamilies, when present, was not found in a consistent position. High variability was noted in the size of the first intron across all subfamilies, similar to that observed in *E. grandis* (Külheim et al. 2015). Genes ranged in size from 1 564 - 7 747 bp, with final products ranging from 337 - 739 amino acids in length (Table S1).

TPS gene expression

A heat map showing relative transcript abundance in five tissues is shown in Figure 6. Several expression clusters were observed, with the first expressed in both unexpanded and expanded leaves. This cluster mostly comprised genes from the *TPS-a* and *TPS-b2* subfamilies. The next cluster was characterised by expression of *TPS-a* and *TPS-b1* genes in leaves and flowers. A final cluster

consisted of *TPS-a* and *TPS-b1* genes expressed in flower initials and flower buds. Of the genes involved in primary metabolism, *CorciTPS088* (*TPS-c*) was moderately expressed in bark, while CorciTPS089 (*TPS-e*) was moderately expressed across all five libraries examined. No expression was detected in the *TPS-f* subfamily.

DISCUSSION

Broad conservation in the eucalypt *TPS* family

Our analyses indicate broad conservation in gene numbers, subfamily representation, physical position and structure of clusters in the *TPS* gene family in *Corymbia citriodora* subsp. *variegata* (CCV) when compared to its divergent sister eucalypts *Eucalyptus grandis* and *E. globulus*. These eucalypts all have the same *TPS* subfamilies, which is expected given the evolution of these subfamilies is believed to pre-date the formation of the Myrtaceae (Keszei et al. 2010a). However, their similar gene numbers and subfamily representation was unexpected given (i) their divergence time from one another (Crisp et al. 2011; Thornhill et al. 2015) relative to their divergence time from the other species studied (*V. vinifera, P. trichocarpa* and *A. thaliana*) and (ii) the instability generally found in large gene families (Lynch 2007; Demuth and Hahn 2009).

We found 102 putative functional *TPS* genes in CCV, which is similar to the numbers found in *Eucalyptus grandis* (113) and *E. globulus* (106) (Külheim et al. 2015). The low variation in total number of *TPS* genes and proportional representation of each subfamily between *E. globulus, E. grandis* and CCV provides evidence for broad conservation of this gene family across these eucalypt lineages. This is in contrast to the other taxa examined in this study (*V. vinifera, P. trichocarpa* and *A. thaliana*), which varied extensively in the *TPS* family in gene number, subfamily presence and proportional representation (Aubourg et al. 2002; Martin et al. 2010; Irmisch et al. 2014). Few instances of gene orthology were detected between these three species or to the eucalypts, especially in the subfamilies involved in secondary metabolite synthesis. All these species are thought to have shared a common ancestor approximately 115 MYA (Wikström et al. 2001; Chaw et

al. 2004), which, when considering the divergence of *Eucalyptus* and *Corymbia* at approximately 52 MYA (Crisp et al. 2011; Thornhill et al. 2015), makes the conservation observed between these divergent eucalypts notable. This leads us to suggest the *TPS* family size and structure observed is representative of eucalypts in general.

The number of *TPS* genes in all three eucalypts currently studied is notably high compared to other plants. Previous studies have revealed *TPS* gene family sizes ranging from one in the bryophyte *Physcomitrella patens* (Hayashi et al. 2006) to 57 in *V. vinifera* (Martin et al. 2010). Consistent with our relatively high estimates of gene birth in eucalypts compared with other taxa (Figure S1), *Eucalyptus grandis* appears to have a gene duplication rate 3 - 5 times that of *Arabidopsis* and *Populus* but comparable rates of gene loss (Myburg et al. 2014), which may contribute to the higher *TPS* gene numbers in the eucalypts. Factors such as physiology and longevity of these plants may play a role in determining the optimal *TPS* gene family size. For instance, plants that emit or store few terpenes generally have few *TPS* loci, such as *A. thaliana* and *P. trichocarpa*, while those that emit and store a more varied range of terpenes often contain more *TPS* genes (Külheim et al. 2015). Overabundance of terpenes can cause autotoxicity (Goodger et al. 2013), but plants able to store terpenes in trichomes or other glandular structures (Carr and Carr 1970) may escape this autotoxic effect. Indeed, eucalypts and *V. vinifera*, both characterised by diverse terpene profiles and the highest numbers of *TPS* loci in plants studied to date, have specialised storage structures such as oil glands. Longevity may also be a contributing factor. Due to their long generation time, more elaborate stress response mechanisms are required in perennial plants compared to herbaceous species (Soler et al. 2015). This may account for the expansion of gene families involved in stress responses in many perennials, as large numbers of genes provide an advantage in inducible responses such as pathogen resistance and other stressors and allow for rapid evolution in response to environmental change (Żmieńko et al. 2014; Sharma and Pandey 2015). For example, the *MYB* gene family, known to be involved in responses to biotic and abiotic stressors, is often expanded into

large duplicate arrays in woody species but not in herbs (Soler et al. 2015), mirroring the discrepancy seen in *TPS* numbers between herbaceous species such as *A. thaliana* and the eucalypts.

Variation in the *TPS* genes specific to each eucalypt lineage

The conservatism at the subfamily level masks the variable expansion and contraction of gene numbers in orthologous clusters within subfamilies of *TPS* genes which, along with the much higher birth and death rate relative to the other taxa studied (Figure S1), signals an evolutionarily dynamic gene family. While the importance of whole genome duplications in plant evolution is often emphasised (Soltis et al. 2014), equally as important are smaller scale duplications at the level of individual genes or gene families (Żmieńko et al. 2014). These smaller scale gene duplications (broadly defined as segmental duplications) occur when errors in DNA replication, recombination or repair generate a copy of a DNA segment containing one or more genes (Lynch and Conery 2000). Many duplicate genes are tandemly associated with their parent copy (tandem duplicates) or occur in 'localised' (within a few Mb) regions of the genome, although non-localised inter/intra-chromosomal duplicates are found at lower frequency (Leister 2004; Myburg et al. 2014). Two copies of a gene are often superfluous and thus either may begin to accumulate mutations, resulting in one of several fates: neo-functionalization, where the mutated gene develops a new function; sub-functionalization, where the two copies of the gene split the function of the original gene; or degeneration, where the gene is deactivated through mutations causing loss of function, often resulting in a pseudogene (Lynch and Conery 2000). The varied structures of *TPS* clusters in both *E. grandis* and CCV is indicative of the complex evolutionary history of this gene family (explored further in the section Physical structure of *TPS* gene clusters).

Many clades throughout the phylogenies show orthologous pairing between genes from *E. grandis* and *E. globulus.* In contrast CCV *TPS* genes are more divergent with the most closely related genes to the *Eucalyptus* species often in separate clades within subfamilies; consistent with the more recent divergence of *E. grandis* and *E. globulus* compared to the divergence of *Corymbia* and

Eucalyptus (Crisp et al. 2011; Thornhill et al. 2015). Külheim et al. (2015) suggest that the similarities in *TPS* genes observed between *E. grandis* and *E. globulus* are a result of much of the evolution of this gene family occurring prior to their divergence. In contrast, the differences exhibited between *Corymbia* and *Eucalyptus* may result from the expansion or contraction of these gene clusters after their divergence. This is likely the case in clades with a large disparity in *TPS* gene number between CCV and the other eucalypts, for instance the *TPS-a* clade with 16 genes in CCV compared to six in *E. grandis* (Figure 1-a). Concerted evolution may have played a role in the differentiation of some members of the *TPS* gene family, obscuring the orthology between *TPS* genes from *E. grandis* and *E. globulus* compared with the related *Corymbia*. Concerted evolution is a process by which copies of genes separated by speciation grow to resemble neighbouring gene copies rather than their true orthologs from other species, through mechanisms such as ectopic gene conversion (Chen et al. 2007). This process may be acting throughout the *TPS* gene family and is probably the most parsimonious explanation for cases where a cluster is of similar size in all three species such as shown in Figure 1-d and Figure 2-c/2-d, as opposed to multiple instances of lineage specific expansion. For example, evidence for gene conversion was found between CCV *TPS-b2* genes in a clade with similar numbers of genes in each species (Figure 2-d, Table S7), lending support to this hypothesis. However, sequencing and annotation of the *TPS* gene family in a sister taxa of the eucalypts [e.g. *Arillastrum, Allosyncarpia, Stockwellia,* or *Eucalyptopsis* (Macphail and Thornhill 2016)] is needed to provide a suitable outgroup to elucidate which mode of evolution affected specific clades as well as whether expansion/contraction of clusters occurred in the *Corymbia* or *Eucalyptus* lineage.

Variation in the *TPS* subfamilies involved in secondary metabolite synthesis

While the overall proportional representation of each *TPS* subfamily is not significantly different, CCV and *Eucalyptus* exhibit marked differences in gene number within several physical clusters in subfamilies *TPS-a*, *-b* and *-g* (Figure 1, Figure 2). These *TPS* subfamilies are involved in the

synthesis of secondary metabolites, which play roles in biotic/abiotic stress responses (Chen et al. 2011). Differential expansion/contraction of gene clusters between species has been often observed, including in the receptor kinase gene family across Brassicaceae (Hofberger et al. 2015) and the *MYB* family across various taxa (Wilkins et al. 2009; Soler et al. 2015); specifically, *R2R3-MYB* gene number varies from 118 in *V. vinifera* to 192 in *P. trichocarpa* (Wilkins et al. 2009). There is potential that localised duplication of these genes facilitates the gain of new function while keeping new copies under similar regulatory control, either through directly copying the original regulatory elements or through other controls such as shared promoters (Williams and Bowles 2004). This mechanism is thought to provide a selective advantage in inducible responses such as biotic resistance, as their shared regulatory control will express both the original and this new potentially advantageous gene when a response is induced (Leister 2004; Hanada et al. 2008). If advantageous, these duplicate genes will be maintained, leading to the expansion of clusters as seen in the *TPS* genes presented here.

Conservation in the *TPS* subfamilies involved in primary metabolite synthesis

In contrast to the other subfamilies, those involved in the synthesis of primary metabolites (*TPS-c*, *-e,* and to a lesser extent *-f*) are more conserved in cluster copy number across eucalypt species (Figure 3), likely reflecting stronger selective constraints on primary *versus* secondary metabolites (Chen et al. 2011). Conservation within a selectively constrained section of an expanded gene family has been previously observed in families such as *MYB* (Wilkins et al. 2009) and *SBP-box* in plants (Zhang et al. 2015), consistent with our findings. As well as greater conservation of gene numbers within clusters, there was also greater conservation of synteny in the subfamilies involved in the synthesis of primary metabolites than those involved in secondary metabolite synthesis across the eucalypts. All *TPS* loci involved in primary metabolite synthesis (*TPS-c*, *-e* and *-f*) were syntenic between *E. grandis* and CCV with no evidence of transposition between chromosomes (aside from a single *TPS-c* gene for which there is evidence of misassembly). The hypothesis of 'gene balance'

suggests that duplicate genes that act in dosage-dependant manners are usually only retained after polyploidy events (Veitia 2004). In the event of a small scale duplication the other parts of the metabolic pathway are often unchanged, which may cause unused product to accumulate and result in detrimental dosage effects (Freeling 2009; Tang and Amon 2013). The conservation seen in *TPS* gene families involved in primary metabolism is consistent with this hypothesis. It is interesting to note that *A. thaliana* has only a single copy of *TPS-c*, *-e* and *-f* genes, while the eucalypts generally have two or more (Table 1, Figure 3). Whole genome duplications specific to each lineage have been detected in both *Arabidopsis* and the plant order Myrtales to which the family Myrtaceae belongs (Arabidopsis Genome Initiative 2000; Myburg et al. 2014), suggesting the persistence of *TPS* duplicates in these subfamilies was not advantageous for *Arabidopsis*.

Contributions of stochastic and selective pressures to the variation in the *TPS* gene family

The balance observed in total subfamily representation may be due to the stochastic nature of mechanisms driving gene duplication and loss (Lynch 2007). While selection will act to fix or purify (inactivate) a beneficial or detrimental duplicate gene, these duplicates can also be selectively neutral, leading to their maintenance and subsequent cluster expansion with very minor impact on the fitness of the organism (Iskow et al. 2012). Maintaining a large library of neutral genes can be selectively advantageous in areas of environmental volatility, allowing the organism potentially to be 'pre-adapted' to stressors (Hurles 2004; Hanada et al. 2008; Kondrashov 2012). Genes in large families have also been shown to be gained and lost at very similar rates through analysis of gene 'birth and death' across gene families in multiple genomes (Demuth and Hahn 2009; Szöllősi and Daubin 2012) and specifically in *A. thaliana* (Cannon et al. 2004). If changes in cluster number are occurring across the entire *TPS* gene family within a species, expansion in one cluster may be countered by degeneration in another, contributing to the overall balance in subfamily representation despite the apparent species specific gain and loss of loci observed between the eucalypts.

The conservation of high *TPS* numbers and subfamily proportional representation across the eucalypts despite the extensive variation in some subfamilies may be a signal that selection is involved. While selection may be acting on the phenotype to drive duplicated genes to fixation or degeneration, the combined effect of these large gene families is also likely to be influenced by selection. The maintenance of a large library of genes, while advantageous in some situations (Żmieńko et al. 2014; Sharma and Pandey 2015), may have associated costs. These include increasing expression and regulation requirements with increasing number (Schiffer et al. 2016) and the possibility of 'runaway expansion' contributing to genome instability (Gijzen 2009; Schiffer et al. 2016), which may be detrimental enough to select against further expansion of *TPS* clusters. Given that increased gene copies often result in increased expression of the subsequent product, there may also be a maximum amount of *TPS* genes that eucalypts can support without experiencing specific deleterious gene dosage effects. For example, the overexpression of particular *TPS-a* genes has been shown to retard growth in tomato (Fray et al. 1995), tobacco (Busch et al. 2002) and *A. thaliana* (Aharoni et al. 2003; Ee et al. 2014) (though not without exception, see Schnee et al. (2006)). This is thought to be the result of under-expression of primary metabolites, such as gibberellin and abscisic acid, due to terpenoid precursor reserves becoming exhausted by the over-synthesis of secondary metabolites. This theory, along with the autotoxicity explored earlier, may select against unregulated expansion in *TPS* clusters and contribute to the stability in *TPS* gene numbers and subfamily representation across the eucalypts.

Pseudogenes and expression of *TPS* loci

The most likely fate of duplicate genes is to be released from selection and acquire mutations which render them non-functional, resulting in pseudogenes. We found 25 *TPS* pseudogenes in the CCV genome, which is 24.5% of the total putative *TPS* gene family size (Table S2). Fifteen of these occurred in the main chromosome assemblies of which all but one was within existing TPS clusters, providing further evidence for the extensive history of local gene duplications

522 in this lineage. Of the 25 pseudogenes, 15 showed evidence of expression, which is an interesting
523 finding. While pseudogenes are thought to play at most a passive role in the genome, such as being
524 sequence donor/receptors for proximal genes (Zheng and Gerstein 2007), it has been shown that
525 some pseudogene transcripts in humans can bind to mRNA from related functional genes and affect
526 their expression (Vinckenbosch et al. 2006). If this is the case in *Corymbia,* these pseudogenes may
527 be part of a mechanism for modulating gene expression.

528 Expression analysis of putative functional genes (Figure 6) revealed several distinct
529 expression clusters each of which involved multiple subfamilies. Secondary metabolite subfamilies
530 were represented across most tissues, consistent with the broad applications of these terpenoids
531 (Keszei et al. 2010a). An interesting pattern was detected in the *TPS-b2*, which were highly expressed
532 in the unexpanded and expanded leaf tissue libraries, with low expression in other libraries. This
533 subfamily is involved in the synthesis of isoprene, a terpenoid hypothesized to confer
534 thermotolerance (Peñuelas et al. 2005). Isoprene is known to lower tissue surface temperature
535 when emitted (Sasaki et al. 2007) and also improves the stability of plant membranes (Singsaas et al.
536 1997). As these both affect photosynthetic rate, the higher expression of isoprene in leaf tissue is
537 consistent with these modes of action. The analysis also showed the *TPS-f* subfamily in CCV was not
538 expressed in the five tissues examined (flower buds and initials, unexpanded and expanded leaf, and
539 bark). This is consistent with similar analysis in *E. grandis,* which revealed that most *TPS-f* genes
540 were solely expressed in root tissue (Külheim et al. 2015), a tissue not covered by our analysis. This
541 subfamily also showed higher divergence than the other primary metabolite subfamilies in all three
542 eucalypt species. Due to this non-typical expression pattern, Külheim et al. (2015) suggest *TPS-f* play
543 a role mediating interactions with herbivores and other soil organisms (Wenke et al. 2010), or
544 influencing allelopathic effects (del Moral and Muller 1970). Indeed, the divergence of the *TPS-f*
545 subfamily across the eucalypts could signal the potential environmental specificity of these
546 interactions.

Physical structure of *TPS* gene clusters

In both *E. grandis* and CCV, most *TPS* genes were clustered in localised regions of the genome (spanning up to 3.5 MB). The fact that each cluster contained only *TPS* genes from the same subfamily that are also closely related in sequence (with the exception of one *TPS-a* gene located within a dispersed *TPS-b* cluster on chromosome five of CCV) suggests they were generated by localised or tandem gene duplication. Indeed, *E. grandis* is characterised by a high rate of tandem duplication relative to other plants (Myburg et al. 2014), which has been proposed as the main reason for the extensive *TPS* family in the eucalypts (Myburg et al. 2014; Külheim et al. 2015), as well as gene families in many other species (Kliebenstein et al. 2001; Leister 2004; Hofberger et al. 2013; Hofberger et al. 2015; Li et al. 2015). In some cases the eucalypt *TPS* genes were in true tandem arrays with no genes interspersed (the largest being a syntenic cluster with six *TPS-a* genes in CCV and 10 in *E. grandis*), while in most cases several non-*TPS* genes were present within these clusters, ranging from 1 - 192 genes separating the closest *TPS* in CCV (Table 2). The varying spans and intervening gene number of *TPS* clusters in *E. grandis* and CCV likely reflect the many ways clusters can form and be subsequently rearranged (Leister 2004; Lynch 2007; Field et al. 2011). For example, segmental duplications range in size and can result in partial genes to large-scale genome segments being copied and translocated to new inter/intra-chromosomal positions (Flagel and Wendel 2009; Wang et al. 2012), or positions local to the origin (Cannon et al. 2004). Hence, the duplication process may initially result in tandem, localised or dispersed gene pairs. Superimposed on this variation, localised (and tandem) duplications can be subsequently dispersed by various mechanisms of genome rearrangement (Lynch 2007; Field et al. 2011); including inversions, insertion/deletions, translocations and further segmental duplications (the tandem expansion of *NB-LRR* genes contained within several *TPS* clusters may be an example of the latter, see Table S4). Differentiating between these various processes requires determining the relative age of duplications, which due to the inherent difficulties introduced by concerted evolution obfuscating mutations is beyond the scope of this study (Mendivil-Ramos and Ferrier 2012).

The physical clustering of non-homologous, but functionally related genes is an emerging theme in plant genomics, particularly in the case of secondary metabolite pathways (Chu et al. 2011; Field et al. 2011; Takos and Rook 2012). Many non-*TPS* genes within *TPS* clusters have putative functions that may interact with or complement the function of *TPS* genes (Table S4). Genes potentially involved in the synthesis of terpene precursors, such as prenyl transferases, along with those involved in post-translational modification of terpenes such as cytochrome c oxidases and NAD-dependant dehydrogenases (Keszei et al. 2008) were found within *TPS* clusters in both *E. grandis* (Külheim et al. 2015) and CCV. Also found were genes from the *NB-LRR*, *MYB* and *WRKY* families, which among other things are involved in pest resistance (Liu et al. 2004; Eitas and Dangl 2010), much like *TPS* genes. The location of these genes within *TPS* clusters may be advantageous, as genes involved in the same biosynthetic pathway or in similar responses can be regulated together at the chromatin level (Field and Osbourn 2008; Chu et al. 2011). This arrangement may also be beneficial for inheritance, as a collection of beneficial alleles from a single metabolic pathway are less likely to be separated by recombination when in close proximity (Chu et al. 2011).

Conclusions

This study contributes to a greater understanding of the terpene synthase gene family through detailed annotation in the recently assembled *C. citriodora* subsp. *variegata* genome and comparative analysis with the previously studied *E. grandis* and *E. globulus*. These *Eucalyptus* species have the most *TPS* loci discovered in any plant to date, and our results show the large size of this gene family is conserved in the sister genus *Corymbia*, suggesting this may be a characteristic of the eucalypts. Both the proportional representation of subfamilies and the syntenic physical position of gene clusters indicated a high degree of conservation in the *TPS* gene family between CCV and *E. grandis.* Despite this conservation, cluster specific variation within subfamilies involved in secondary metabolite synthesis were observed, and we discuss the potential contributions of selection,

concerted evolution and stochastic processes to this observation. The higher degree of conservation of *TPS* genes involved in primary metabolite synthesis is likely due to greater selective constraints.

AKNOWLEDGEMENTS

The research presented here is part of a larger project working towards the creation of a reference genome for *Corymbia*, please see http://scu.edu.au/scps/index.php/137 for more information. The germplasm used to create the genomes referenced in this study was provided by the Queensland Department of Agriculture and Fisheries (DAF). The authors thank John Oostenbrink (DAF) for his work to produce the hybrid families, Valerie Hecht (University of Tasmania) for valuable advice regarding the creation of the phylogenies and Agnelo Furtado (University of Queensland) for consultation regarding the *Corymbia* genome project. This work was supported by the Australian Research Council [grant numbers DP140102552, DP110101621], and an Australian Government Research Training Program Scholarship. Sequencing and assembly data carried out by EMBRAPA as part of the *Corymbia* genome project was supported by FAPDF grant "Nextree" 193.000.570/2009. For the portion of the work conducted by the Joint Genome Institute and the Joint BioEnergy Institute, support was provided by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA ARCHIVING

New sequences used for alignment are presented in the Supplementary Material, available at Heredity's website. The *Corymbia citriodora* subsp. *variegata* genome assemblies will be published in the Comparative Genomics database (https://genomevolution.org/coge/) when publication is completed.

SUPPLEMENTARY MATERIAL

Table S1 - *C. citriodora* subsp. *variegata* putative functional terpene synthase genes from the CCV18 genome assembly

Table S2 - *C. citriodora* subsp. *variegata* terpene synthase pseudogenes from the CCV18 genome assembly

Table S3 - *C. citriodora* subsp. *variegata* terpene synthase genes from the CCV54 genome assembly

Table S4 - Intervening genes in the *C. citriodora* subsp. *variegata* terpene synthase physical clusters

Table S5 - *TPS* cluster copy number comparison between *C. citriodora* subsp. *variegata* assemblies CCV18 and CCV54

Table S6 - Markers anchoring contigs containing putative translocated *TPS* gene clusters in CCV18

Table S7 - Gene conversion events in the *TPS-b2* subfamily of *Corymbia citriodora* subsp. *variegata*

Figure S1 - Gene birth and death rates in the *TPS* gene family across a) multiple taxa and b) the eucalypt lineages

642 **REFERENCES**

643 Aharoni A, Giri AP, Deuerlein S, Griepink F, de Kogel W-J, Verstappen FWA et al. (2003) Terpenoid
644 metabolism in wild-type and transgenic *Arabidopsis* plants. The Plant Cell 15:2866-2884

645 Ammon DG, Barton AF, Clarke DA, Tjandra J (1985) Rapid and accurate determination of terpenes in
646 the leaves of *Eucalyptus* species. Analyst 110:921-924

647 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant
648 *Arabidopsis thaliana*. Nature 408:796-815

649 Asante KS, Brophy JJ, Doran JC, Goldsack RJ, Hibbert DB, Larmour JS (2001) A comparative study of
650 the seedling leaf oils of the spotted gums: species of the *Corymbia* (Myrtaceae), section
651 Politaria. Australian Journal of Botany 49:55-66

652 Aubourg S, Lecharny A, Bohlmann J (2002) Genomic analysis of the terpenoid synthase (AtTPS) gene
653 family of *Arabidopsis thaliana*. Molecular Genetics and Genomics 267:730-745

654 Batish DR, Singh HP, Kohli RK, Kaur S (2008) *Eucalyptus* essential oil as a natural pesticide. Forest
655 Ecology and Management 256:2166-2174

656 Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome
657 Research 10:547-548

658 Busch M, Seuter A, Hain R (2002) Functional analysis of the early steps of carotenoid biosynthesis in
659 tobacco. Plant Physiology 128:439-453

660 Bushnell B (2016) BBMap. http://www.sourceforge.net/projects/bbmap/

661 Butler JB, Vaillancourt RE, Potts BM, Lee DJ, King GJ, Baten A et al. (2017) Comparative genomics of
662 *Eucalyptus* and *Corymbia* reveals low rates of genome structural rearrangement. BMC
663 Genomics 18:397

664 Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem
665 gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant
666 Biology 4:10

667 Carr D, Carr S (1970) Oil glands and ducts in *Eucalyptus* L'Héit. II. Development and structure of oil
668 glands in the embryo. Australian Journal of Botany 18:191-212
669 Chaw S-M, Chang C-C, Chen H-L, Li W-H (2004) Dating the monocot–dicot divergence and the origin
670 of core eudicots using whole chloroplast genomes. Journal of Molecular Evolution 58:424-
671 441
672 Chen F, Tholl D, Bohlmann J, Pichersky E (2011) The family of terpene synthases in plants: a mid‐
673 size family of genes for specialized metabolism that is highly diversified throughout the
674 kingdom. The Plant Journal 66:212-229
675 Chen J-M, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms,
676 evolution and human disease. Nature Reviews: Genetics 8:762-775
677 Chu HY, Wegel E, Osbourn A (2011) From hormones to secondary metabolism: the emergence of
678 metabolic gene clusters in plants. The Plant Journal 66:66-79
679 Crisp MD, Burrows GE, Cook LG, Thornhill AH, Bowman DMJS (2011) Flammable biomes dominated
680 by eucalypts originated at the Cretaceous-Palaeogene boundary. Nature Communications
681 2:193
682 del Moral R, Muller CH (1970) The allelopathic effects of *Eucalyptus camaldulensis*. The American
683 Midland Naturalist 83:254-282
684 Demuth JP, Hahn MW (2009) The life and death of gene families. Bioessays 31:29-39
685 Douglas MH, van Klink JW, Smallfield BM, Perry NB, Anderson RE, Johnstone P et al. (2004) Essential
686 oils from New Zealand manuka: triketone and other chemotypes of *Leptospermum*
687 *scoparium*. Phytochemistry 65:1255-1264
688 Ee S-F, Mohamed-Hussein Z-A, Othman R, Shaharuddin NA, Ismail I, Zainal Z (2014) Functional
689 characterization of sesquiterpene synthase from *Polygonum minus*. The Scientific World
690 Journal 2014:11
691 Eitas TK, Dangl JL (2010) NB-LRR proteins: pairs, pieces, perception, partners, and pathways. Current
692 Opinion in Plant Biology 13:472-477

693 Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE (2011) Formation of plant
694 metabolic gene clusters within dynamic chromosomal regions. Proceedings of the National
695 Academy of Sciences 108:16116-16121
696 Field B, Osbourn AE (2008) Metabolic diversification - independent assembly of operon-like gene
697 clusters in different plants. Science 320:543-547
698 Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytologist
699 183:557-564
700 Fray RG, Wallace A, Fraser PD, Valero D, Hedden P, Bramley PM et al. (1995) Constitutive expression
701 of a fruit phytoene synthase gene in transgenic tomatoes causes dwarfism by redirecting
702 metabolites from the gibberellin pathway. The Plant Journal 8:693-701
703 Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-
704 genome, segmental, or by transposition. Annual Review of Plant Biology 60:433-453
705 Gao Y, Honzatko RB, Peters RJ (2012) Terpenoid synthase structures: a so far incomplete view of
706 complex catalysis. Natural Product Reports 29:1153-1175
707 Gijzen M (2009) Runaway repeats force expansion of the *Phytophthora infestans* genome. Genome
708 Biology 10:241
709 Goodger JQ, Heskes AM, Woodrow IE (2013) Contrasting ontogenetic trajectories for phenolic and
710 terpenoid defences in *Eucalyptus froggattii*. Annals of Botany 112:651-659
711 Grattapaglia D, Bradshaw Jr HD (1994) Nuclear DNA content of commercially important *Eucalyptus*
712 species and hybrids. Canadian Journal of Forest Research 24:1074-1078
713 Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C et al. (2012) Progress
714 in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. Tree Genetics &
715 Genomes 8:463-508
716 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al. (2013) De novo
717 transcript sequence reconstruction from RNA-Seq: reference generation and analysis with
718 Trinity. Nature Protocols 8:10.1038/nprot.2013.1084

719 Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific
720 expansion of plant tandem duplicates in the adaptive response to environmental stimuli.
721 Plant Physiology 148:993-1003
722 Hansen NL, Heskes AM, Hamberger B, Olsen CE, Hallström BM, Andersen-Ranberg J et al. (2017) The
723 terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene
724 synthase among the monoterpene synthase TPS-b subfamily. The Plant Journal 89:429-441
725 Hayashi K-i, Kawaide H, Notomi M, Sakigi Y, Matsuo A, Nozaki H (2006) Identification and functional
726 analysis of bifunctional ent-kaurene synthase from the moss *Physcomitrella patens*. FEBS
727 Letters 580:6175-6181
728 Healey A, Shepherd M, Baten A, King GJ, Lee DJ, Furtado A et al. (2017) Sequencing the branches of
729 the eucalypt tree: comparison between *Eucalyptus* and *Corymbia g*enomes. In: Plant &
730 Animal Genome Conference XXV, San Diego, United States of America.
731 Heiling S, Schuman MC, Schoettner M, Mukerjee P, Berger B, Schneider B et al. (2010) Jasmonate
732 and ppHsystemin regulate key malonylation steps in the biosynthesis of 17-
733 hydroxygeranyllinalool diterpene glycosides, an abundant and effective direct defense
734 against herbivores in *Nicotiana attenuata*. The Plant Cell 22:273-292
735 Hill KD, Johnson LA (1995) Systematic studies in the eucalypts 7. A revision of the bloodwoods, genus
736 *Corymbia* (Myrtaceae). Telopea 6:185-504
737 Hofberger JA, Lyons E, Edger PP, Chris Pires J, Eric Schranz M (2013) Whole genome and tandem
738 duplicate retention facilitated glucosinolate pathway diversification in the mustard family.
739 Genome Biology and Evolution 5:2155-2173
740 Hofberger JA, Nsibo DL, Govers F, Bouwmeester K, Schranz ME (2015) A complex interplay of
741 tandem- and whole-genome duplication drives expansion of the L-type lectin receptor
742 kinase gene family in the Brassicaceae. Genome Biology and Evolution 7:720-734

743 Hosfield DJ, Zhang Y, Dougan DR, Broun A, Tari LW, Swanson RV et al. (2004) Structural basis for
744 bisphosphonate-mediated inhibition of isoprenoid biosynthesis. Journal of Biological
745 Chemistry 279:8526-8529
746 Hurles M (2004) Gene duplication: the genomic trade in spare parts. PLoS Biology 2:e206
747 Irmisch S, Jiang Y, Chen F, Gershenzon J, Köllner TG (2014) Terpene synthases and their contribution
748 to herbivore-induced volatile emission in western balsam poplar (*Populus trichocarpa*). BMC
749 Plant Biology 14:270
750 Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human
751 adaptation. Trends in Genetics 28:245-257
752 Keszei A, Brubaker CL, Carter R, Köllner T, Degenhardt J, Foley WJ (2010a) Functional and
753 evolutionary relationships between terpene synthases from Australian Myrtaceae.
754 Phytochemistry 71:844-852
755 Keszei A, Brubaker CL, Foley WJ (2008) A molecular perspective on terpene variation in Australian
756 Myrtaceae. Australian Journal of Botany 56:197-213
757 Keszei A, Hassan Y, Foley WJ (2010b) A biochemical interpretation of terpene chemotypes in
758 *Melaleuca alternifolia*. Journal of Chemical Ecology 36:652-661
759 Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T (2001) Gene duplication in
760 the diversification of secondary metabolism: tandem 2-oxoglutarate–dependent
761 dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. The Plant Cell 13:681-694
762 Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing
763 environment. Proceedings of the Royal Society B: Biological Sciences 279:5048-5057
764 Külheim C, Padovan A, Hefer C, Krause ST, Köllner TG, Myburg AA et al. (2015) The *Eucalyptus*
765 terpene synthase gene family. BMC Genomics 16:1-18
766 Ladiges PY, Udovicic F, Nelson G (2003) Australian biogeographical connections and the phylogeny of
767 large genera in the plant family Myrtaceae. Journal of Biogeography 30:989-998

768 Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K (1998) Intraspecific variation in *Eucalyptus*
769 secondary metabolites determines food intake by folivorous marsupials. Oecologia 116:160-
770 169
771 Lawler IR, Stapley J, Foley WJ, Eschler BM (1999) Ecological example of conditioned flavor aversion in
772 plant–herbivore interactions: effect of terpenes of *Eucalyptus l*eaves on feeding by common
773 ringtail and brushtail possums. Journal of Chemical Ecology 25:401-415
774 Lee DJ (2007) Achievements in forest tree genetic improvement in Australia and New Zealand 2:
775 Development of *Corymbia* species and hybrids for plantations in eastern Australia. Australian
776 Forestry 70:11-16
777 Lefort V, Longueville J-E, Gascuel O (2017) SMS: Smart Model Selection in PhyML. Molecular Biology
778 and Evolution 34:2422-2424
779 Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of
780 plant disease resistance genes. Trends in Genetics 20:116-122
781 Li F, Zhou C, Weng Q, Li M, Yu X, Guo Y et al. (2015) Comparative genomics analyses reveal extensive
782 chromosome colinearity and novel quantitative trait loci in *Eucalyptus*. PLoS ONE
783 10:e0145144
784 Librado P, Vieira FG, Rozas J (2012) BadiRate: estimating family turnover rates by likelihood-based
785 methods. Bioinformatics 28:279-281
786 Liu Y, Schiff M, Dinesh-Kumar S (2004) Involvement of MEK1 MAPKK, NTF6 MAPK, WRKY/MYB
787 transcription factors, COI1 and CTR1 in N-mediated resistance to tobacco mosaic virus. The
788 Plant Journal 38:800-809
789 Lynch M (2007) The origins of genome architecture. Sinauer Associates, Inc., Sunderland, MA
790 Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science
791 290:1151-1155
792 Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as
793 DNA sequences. The Plant Journal 53:661-673

794 Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H et al. (2008) Finding and comparing syntenic
795 regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids.
796 Plant Physiology 148:1772-1781
797 Macphail M, Thornhill AH (2016) How old are the eucalypts? A review of the microfossil and
798 phylogenetic evidence. Australian Journal of Botany 64:579-599
799 Martin DM, Aubourg S, Schouwey MB, Daviet L, Schalk M, Toub O et al. (2010) Functional
800 annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene
801 synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. BMC
802 Plant Biology 10:226
803 Mendivil-Ramos O, Ferrier DEK (2012) Mechanisms of gene duplication and translocation and
804 progress towards understanding their relative contributions to animal genome evolution.
805 International Journal of Evolutionary Biology 2012:10
806 Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap.
807 Molecular Biology and Evolution 30:1188-1195
808 Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J et al. (2014) The genome
809 of *Eucalyptus grandis*. Nature 510:356-362
810 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic
811 algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution
812 32:268-274
813 O'Reilly-Wapstra JM, McArthur C, Potts BM (2004) Linking plant genotype, plant defensive chemistry
814 and mammal browsing in a *Eucalyptus* species. Functional Ecology 18:677-684
815 O'Reilly-Wapstra JM, Freeman JS, Davies NW, Vaillancourt RE, Fitzgerald H, Potts BM (2011)
816 Quantitative trait loci for foliar terpenes in a global eucalypt species. Tree Genetics &
817 Genomes 7:485-498
818 Padovan A, Keszei A, Külheim C, Foley WJ (2014) The evolution of foliar terpene diversity in
819 Myrtaceae. Phytochemistry Reviews 13:695-716

820 Padovan A, Keszei A, Wallis IR, Foley WJ (2012) Mosaic eucalypt trees suggest genetic control at a
821 point that influences several metabolic pathways. Journal of Chemical Ecology 38:914-923
822 Peñuelas J, Llusià J, Asensio D, Munné-Bosch S (2005) Linking isoprene with plant thermotolerance,
823 antioxidants and monoterpene emissions. Plant, Cell & Environment 28:278-286
824 Pichersky E, Gershenzon J (2002) The formation and function of plant volatiles: perfumes for
825 pollinator attraction and defense. Current Opinion in Plant Biology 5:237-243
826 R Core Team (2017) R: A language and environment for statistical computing. R Foundation for
827 Statistical Computing, Vienna, Austria
828 Rockwood D, Rudie A, Ralph S, Zhu J, Winandy J (2008) Energy product options for *Eucalyptus*
829 *s*pecies grown as short rotation woody crops. International Journal of Molecular Sciences
830 9:1361
831 Sasaki K, Saito T, Lämsä M, Oksman-Caldentey K-M, Suzuki M, Ohyama K et al. (2007) Plants utilize
832 isoprene emission as a thermotolerance mechanism. Plant and Cell Physiology 48:1254-1262
833 Schiffer PH, Gravemeyer J, Rauscher M, Wiehe T (2016) Ultra large gene families: a matter of
834 adaptation or genomic parasites? Life 6:32
835 Schnee C, Köllner TG, Held M, Turlings TCJ, Gershenzon J, Degenhardt J (2006) The products of a
836 single maize sesquiterpene synthase form a volatile defense signal that attracts natural
837 enemies of maize herbivores. Proceedings of the National Academy of Sciences of the
838 United States of America 103:1129-1134
839 Schwab W, Fuchs C, Huang F-C (2013) Transformation of terpenes into fine chemicals. European
840 Journal of Lipid Science and Technology 115:3-8
841 Sharma M, Pandey GK (2015) Expansion and function of repeat domain proteins during stress and
842 development in plants. Frontiers in Plant Science 6:1218
843 Shepherd M, Bartle J, Lee DJ, Brawner J, Bush D, Turnbull P et al. (2011) Eucalypts as a biofuel
844 feedstock. Biofuels 2:639-657

Shepherd M, Baten A, Junior OBdS, Lee DJ, Butler JB, Freeman J et al. (2015) Towards a *Corymbia* reference genome: comparative efficiencies of Illumina, PacBio and hybrid de novo assemblies of a complex heterozygous genome. In: Vettori C, Vendramin GG, Paffetti D, Travaglini D (eds) Proceedings of the IUFRO Tree Biotechnology 2015 Conference: "Forests: the importance to the planet and society", Florence, Italy.

Singsaas EL, Lerdau M, Winter K, Sharkey TD (1997) Isoprene increases thermotolerance of isoprene-emitting species. Plant Physiology 115:1413-1420

Slee A, Brooker M, Duffy S, West J (2006) EUCLID eucalypts of Australia. 3rd edn. Centre for Plant Biodiversity Research - CSIRO Publishing Canberra

Soler M, Camargo ELO, Carocha V, Cassan-Wang H, San Clemente H, Savelli B et al. (2015) The *Eucalyptus grandis* R2R3-MYB transcription factor family: evidence for woody growth-related evolution and function. New Phytologist 206:1364-1377

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then…and now: Stebbins revisited. American Journal of Botany 101:1057-1078

Szöllősi GJ, Daubin V (2012) Modeling gene family evolution and reconciling phylogenetic discord. Evolutionary Genomics: Statistical and Computational Methods 2:29-51

Takos AM, Rook F (2012) Why biosynthetic genes for chemical defense compounds cluster. Trends in Plant Science 17:383-388

Tang Y-C, Amon A (2013) Gene copy-number alterations: a cost-benefit analysis. Cell 152:394-405

Thornhill AH, Ho SYW, Külheim C, Crisp MD (2015) Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. Molecular Phylogenetics and Evolution 93:29-43

Veitia RA (2004) Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics 168:569

Vernin GA, Parkanyi C, Cozzolino F, Fellous R (2004) GC/MS analysis of the volatile constituents of *Corymbia citriodora* Hook. from Réunion Island. Journal of Essential Oil Research 16:560-565

871 Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in
872 the human genome. Proceedings of the National Academy of Sciences of the United States
873 of America 103:3220-3225
874 Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression
875 divergence: a view from plants. Annals of the New York Academy of Sciences 1256:1-14
876 Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T et al. (2016) gplots: various R
877 programming tools for plotting data. https://CRAN.R-project.org/package=gplots
878 Wenke K, Kai M, Piechulla B (2010) Belowground volatiles facilitate interactions between plant roots
879 and soil organisms. Planta 231:499-506
880 Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family
881 tree. Proceedings of the Royal Society of London Series B: Biological Sciences 268:2211-2220
882 Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM (2009) Expansion and diversification of the
883 *Populus* R2R3-MYB family of transcription factors. Plant Physiology 149:981-993
884 Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis*
885 *thaliana*. Genome Research 14:1060-1067
886 Xiong W, Wu P, Jia Y, Wei X, Xu L, Yang Y et al. (2016) Genome-wide analysis of the terpene synthase
887 gene family in physic nut (*Jatropha curcas* L.) and functional identification of six terpene
888 synthases. Tree Genetics & Genomes 12:97
889 Zhang S-D, Ling L-Z, Yi T-S (2015) Evolution and divergence of SBP-box genes in land plants. BMC
890 Genomics 16:787
891 Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead
892 rise up, or do they? Trends in Genetics 23:219-224
893 Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant
894 genomes. Theoretical and Applied Genetics 127:1-18

895

TITLES AND LEGENDS TO FIGURES

Figure 1. Phylogeny of the *TPS-a* subfamily. This tree was created through maximum likelihood analysis comparing the *TPS-a* subfamily from *C. citriodora* subsp. *variegata* (Corci) with those from *E. grandis* (Egran)*, E. globulus* (Eglob)*, P. trichocarpa* (Pt), *V. vinifera* (Vv) and *A. thaliana* (At). Bootstrap values supported by <80% are noted by number, while those with bootstrap values between 80-94% are indicated by the symbol *. All others have values >95%. Scale represents amino acid substitutions per site. Letters indicate specific clades referred to in the text. A *TPS-b* gene from *C. citriodora* subsp. *variegata* was used as the outgroup. a-d refers to results discussed in the text. Examples of orthologous pairings are given by numbers 1 & 2. 3 is not considered an orthologous pairing as *EglobTPS004* shares its most recent ancestral gene with two genes from *E. grandis* rather than one. 4 is not considered an orthologous pairing as *EglobTPS027* and *EgranTPS021* do not share the same most recent ancestral gene. b gives an example of genes in orthologous pairings, with the exception of *EgranTPS029*, which is does not pair to a single gene from another species. c shows an example of a non-orthologous pairing, as *EglobTPS022* is closely related to several genes from CCV rather than a specific one.

Figure 2. Phylogeny of the *TPS-b* and *TPS-g* subfamilies. This tree was created through maximum likelihood analysis comparing the *TPS-b* and *TPS-g* subfamilies from *C. citriodora* subsp. *variegata* (Corci) with those from *E. grandis* (Egran)*, E. globulus* (Eglob)*, P. trichocarpa* (Pt), *V. vinifera* (Vv) and *A. thaliana* (At). Bootstrap values supported by <80% are noted by number, while those with bootstrap values between 80-94% are indicated by the symbol *. All others have values >95%. Scale represents amino acid substitutions per site. Letters indicate specific clades referred to in the text. A *TPS-a* gene from *C. citriodora* subsp. *variegata* was used as the outgroup. a-c refers to results discussed in the text.

Figure 3. Phylogeny of the *TPS-c, TPS-e* and *TPS-f* subfamilies. This tree was created through maximum likelihood analysis comparing the *TPS-c*, *TPS-e* and *TPS-f* subfamilies from *C. citriodora*

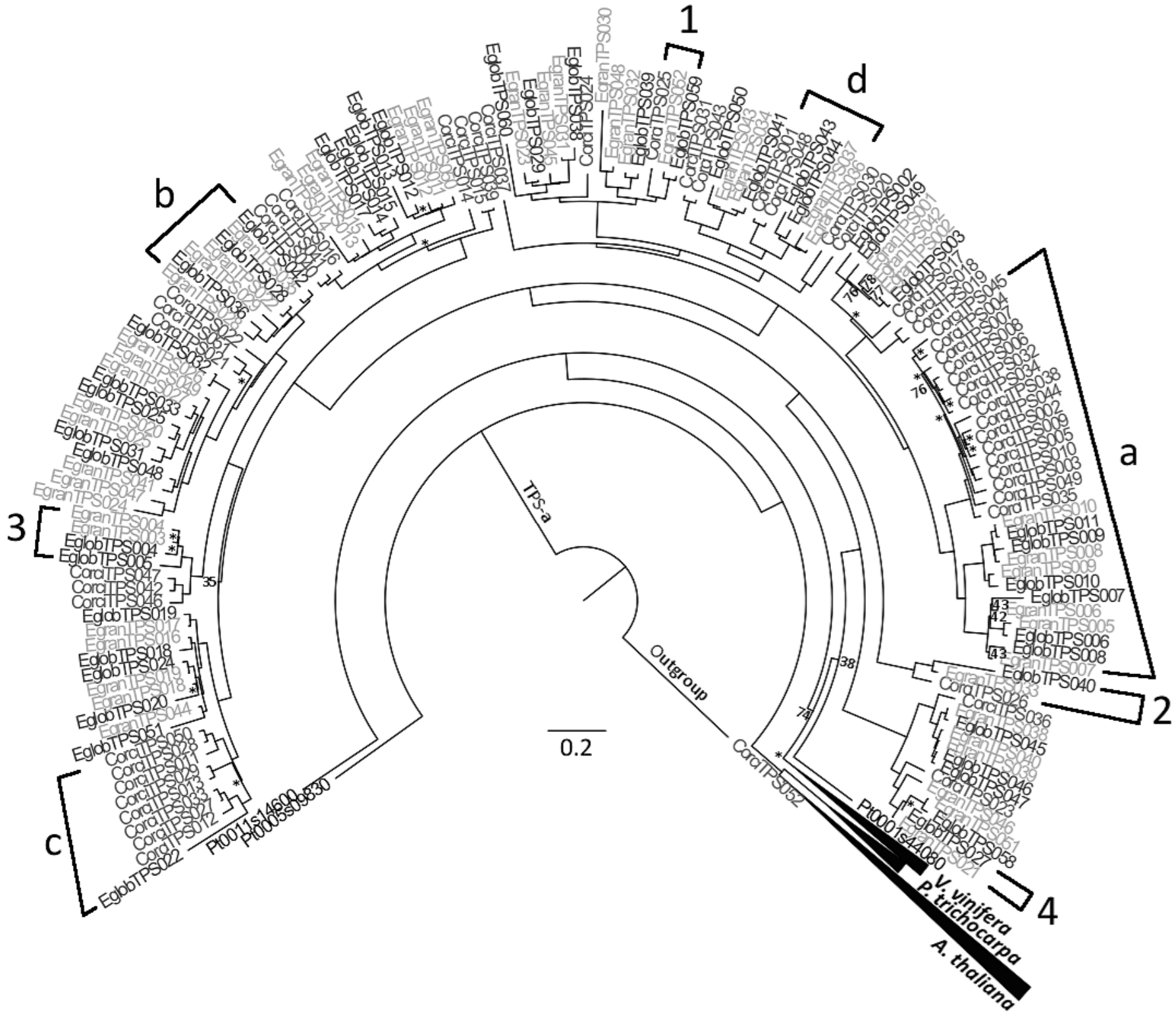
subsp. *variegata* (Corci) with those from *E. grandis* (Egran), *E. globulus* (Eglob), *P. trichocarpa* (Pt), *V. vinifera* (Vv) and *A. thaliana* (At). Bootstrap values supported by <80% are noted by number, while those with bootstrap values between 80-94% are indicated by the symbol *. All others have values >95%. Scale represents amino acid substitutions per site. A *TPS-b* gene from *C. citriodora* subsp. *variegata* was used as the outgroup.

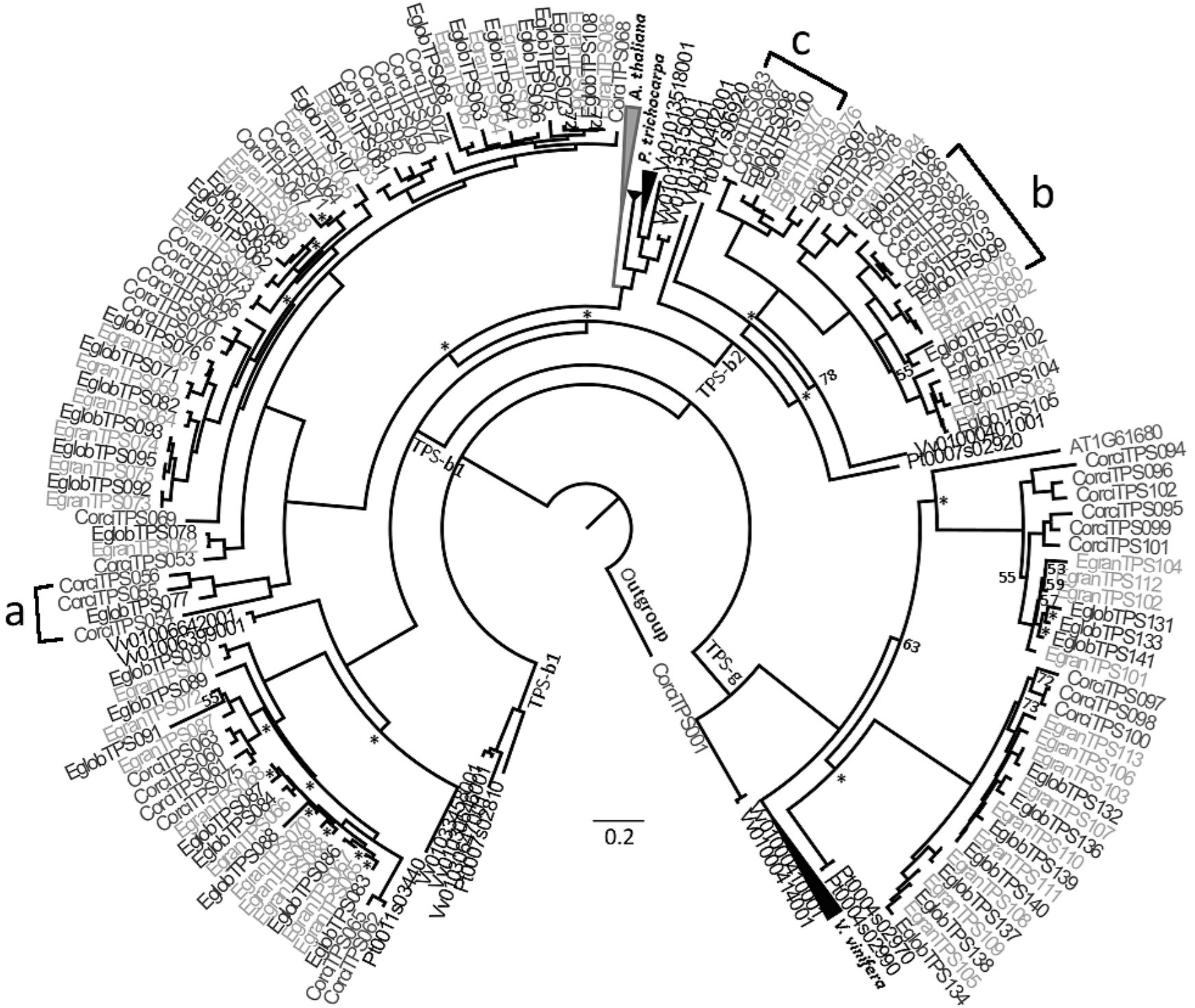
Figure 4. Comparison of copy number and genomic location of *TPS* physical clusters between *E. grandis* (Egr) and *C. citriodora* subsp. *variegata* (CCV). Chromosomes are scaled by physical size. Locus names show the number of *TPS* genes and the subfamily they belong to. Separate clusters on the same chromosome were defined based on both physical distance and phylogenetic relatedness (see Table S4). Solid lines indicate clusters that are both homologous and syntenic between the two species, while broken lines indicate homologous clusters that are present on different chromosomes in each species. For example, in the *TPS-b* subfamily, a cluster of eight *TPS* genes are present on chromosome 4 in *E. grandis*, in contrast to the syntenic and non-syntenic homologous clusters present in *C. citriodora* subsp. *variegata* on chromosome 4 and 2, respectively. Non-syntenic loci between *C. citriodora* subsp. *variegata* and *E. grandis* are circled (only on CCV) to indicate support for this placement based on the CCV54 genome assembly. Similarly, loci are tagged with an asterisk on CCV to indicate disagreement (see Table S5). *TPS* clusters without lines indicate that their homolog is present in the minor scaffolds of the other species and cannot be examined for synteny. Homology of singleton *TPS* genes is not shown.

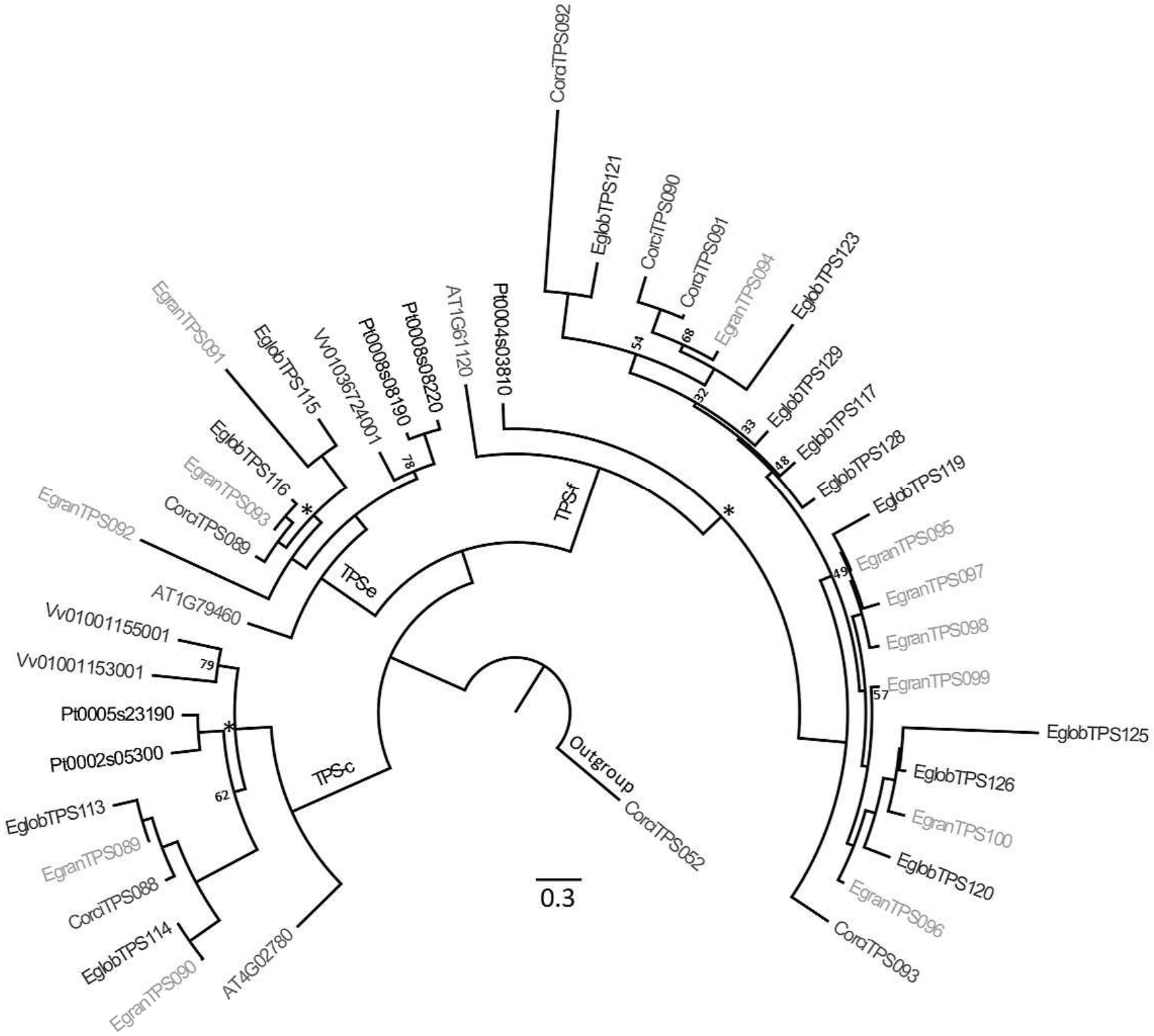
Figure 5. Gene structure of the 102 putative functional *TPS* genes from *Corymbia citriodora* subsp. *variegata*. Exons are shown as boxes, while introns are shown as lines. The arrow indicates the position of the conserved DDxxD motif.

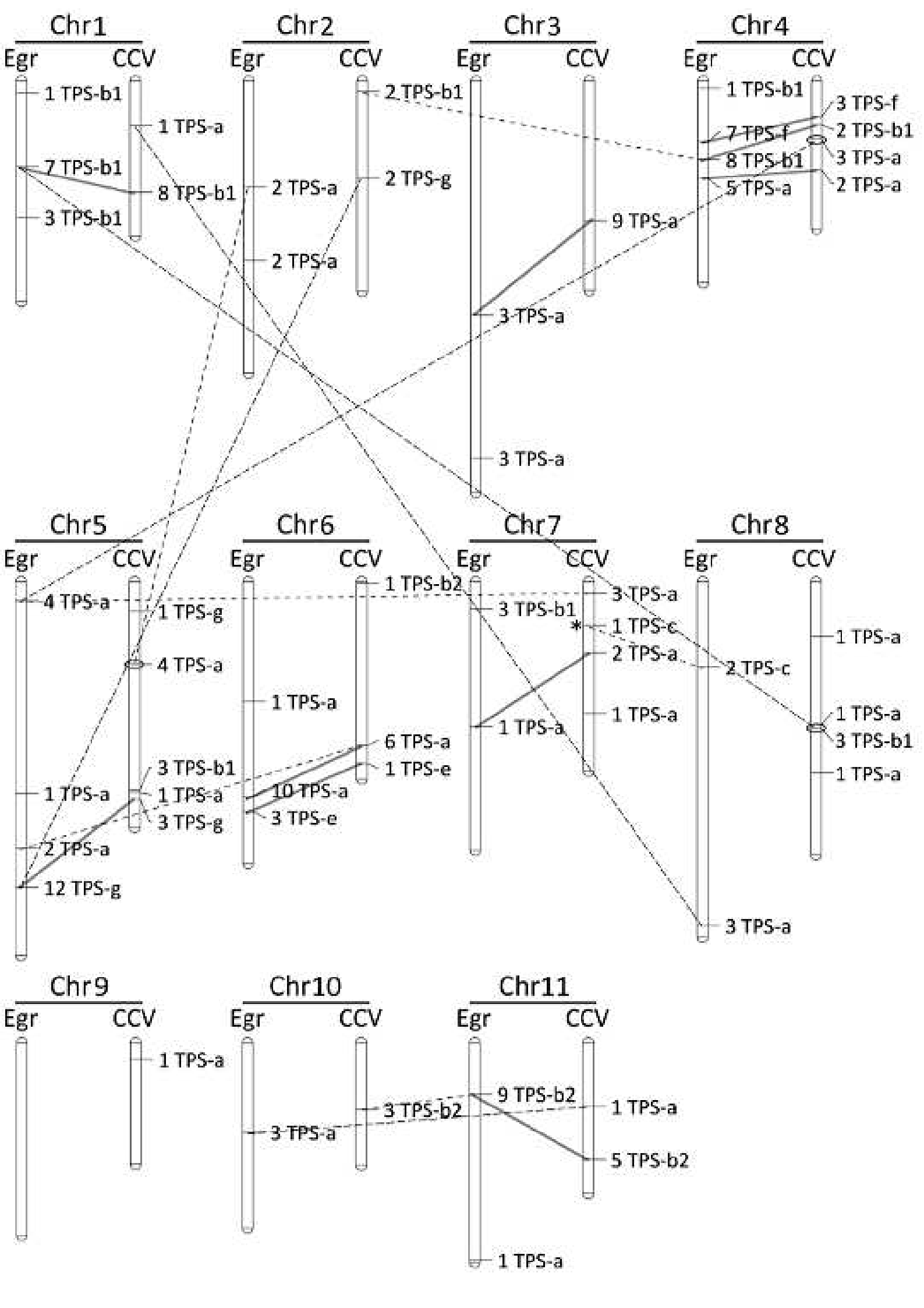
Figure 6. Gene expression clustering of 102 *TPS* genes from *Corymbia citriodora* subsp. *variegata* expressed in five tissues. RNAseq data is shown as fragments per kilobase of transcript per million mapped reads (FPKM), with FPKM values normalised within libraries (largest FPKM value set to 1,

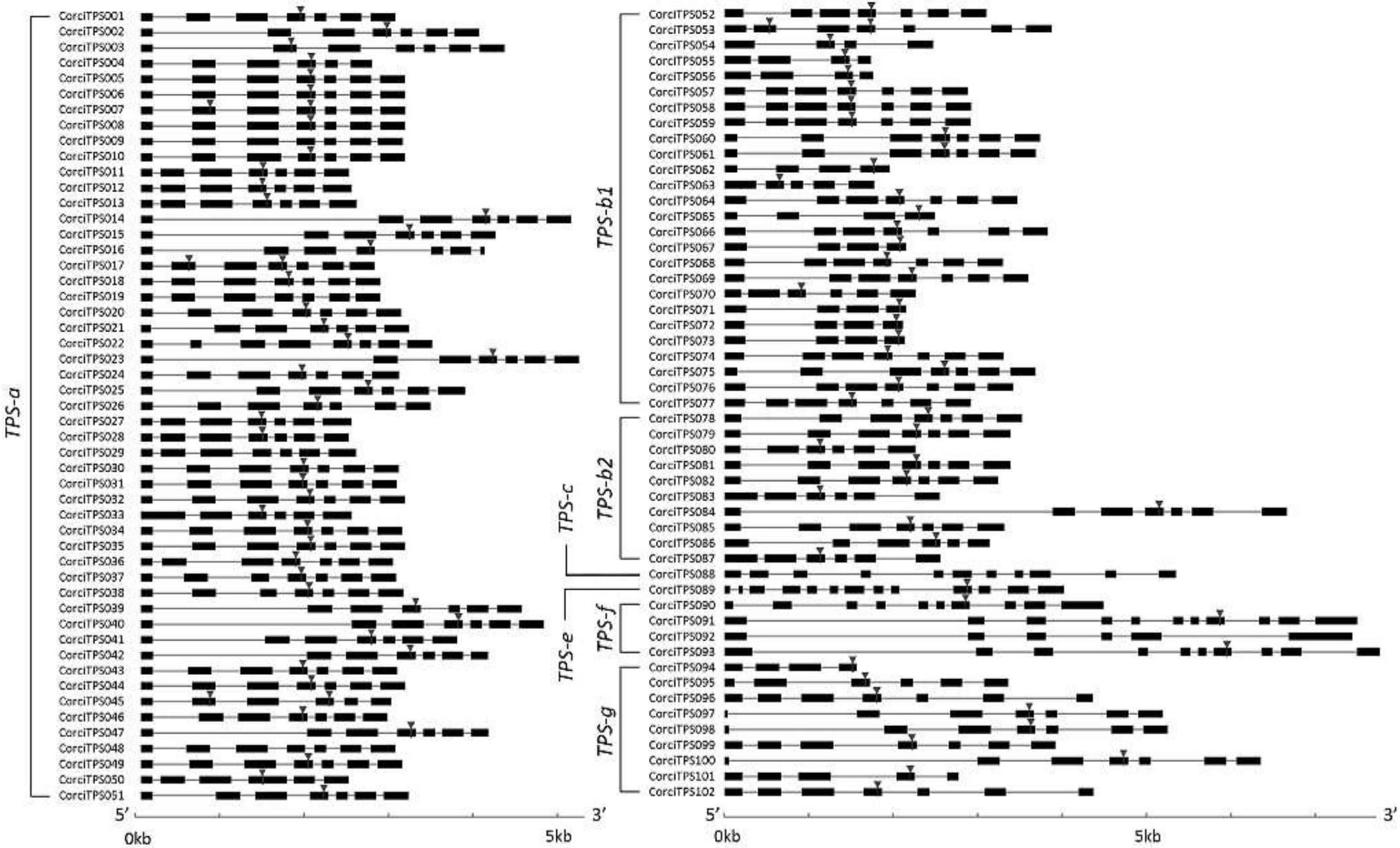
946 with other scores scaled accordingly). The *TPS* subfamily is indicated by suffix after the gene name.
947 The sampled tissues are: flower initials (FI), flower buds (FB), bark (BA), expanded leaf (LE), and
948 unexpanded leaf (LU).


1
d
b
a
3
2
c
4
TPS-a
Outgroup
0.2
V. vinifera
P. trichocarpa
A. thaliana


a
b
c
TPS-b1
TPS-b2
TPS-b3
TPS-g
Outgroup
0.2
A. thaliana
P. trichocarpa
V. vinifera


CorcTPS092
EglobTPS121
CorcTPS090
CorcTPS091
EgranTPS094
EglobTPS123
EglobTPS129
EglobTPS117
EglobTPS128
EglobTPS119
EgranTPS095
EgranTPS097
EgranTPS098
EgranTPS099
EglobTPS125
EglobTPS126
EgranTPS100
EglobTPS120
EgranTPS096
CorcTPS093
Pt0004s03810
AT1G61120
Pt0008s08220
Pt0008s08190
Vv01036724001
EglobTPS115
EgranTPS091
EglobTPS116
EgranTPS093
CorcTPS089
EgranTPS092
AT1G79460
Vv01001155001
Vv01001153001
Pt0005s23190
Pt0002s05300
EglobTPS113
EgranTPS089
CorcTPS088
EglobTPS114
EgranTPS090
AT4G02780
TPSf
TPSe
TPSc
Outgroup
CorcTPS052
0.3
54
68
32
33
48
49
57
78
79
62
*
*
*


Chr1
Egr
CCV
1 TPS-b1
7 TPS-b1
3 TPS-b1
1 TPS-a
8 TPS-b1
Chr2
Egr
CCV
2 TPS-b1
2 TPS-g
2 TPS-a
2 TPS-a
Chr3
Egr
CCV
9 TPS-a
3 TPS-a
3 TPS-a
Chr4
Egr
CCV
1 TPS-b1
7 TPS-f
8 TPS-b1
5 TPS-a
3 TPS-f
2 TPS-b1
3 TPS-a
2 TPS-a
Chr5
Egr
CCV
4 TPS-a
1 TPS-a
2 TPS-a
12 TPS-g
1 TPS-g
4 TPS-a
3 TPS-b1
1 TPS-a
3 TPS-g
Chr6
Egr
CCV
1 TPS-b2
1 TPS-a
6 TPS-a
1 TPS-e
10 TPS-a
3 TPS-e
Chr7
Egr
CCV
3 TPS-b1
1 TPS-a
3 TPS-a
1 TPS-c
2 TPS-a
1 TPS-a
Chr8
Egr
CCV
2 TPS-c
3 TPS-a
1 TPS-a
1 TPS-a
3 TPS-b1
1 TPS-a
Chr9
Egr
CCV
1 TPS-a
Chr10
Egr
CCV
3 TPS-b2
3 TPS-a
Chr11
Egr
CCV
9 TPS-b2
3 TPS-b2
1 TPS-a
5 TPS-b2
1 TPS-a


CorciTPS001
CorciTPS002
CorciTPS003
CorciTPS004
CorciTPS005
CorciTPS006
CorciTPS007
CorciTPS008
CorciTPS009
CorciTPS010
CorciTPS011
CorciTPS012
CorciTPS013
CorciTPS014
CorciTPS015
CorciTPS016
CorciTPS017
CorciTPS018
CorciTPS019
CorciTPS020
CorciTPS021
CorciTPS022
CorciTPS023
CorciTPS024
CorciTPS025
CorciTPS026
CorciTPS027
CorciTPS028
CorciTPS029
CorciTPS030
CorciTPS031
CorciTPS032
CorciTPS033
CorciTPS034
CorciTPS035
CorciTPS036
CorciTPS037
CorciTPS038
CorciTPS039
CorciTPS040
CorciTPS041
CorciTPS042
CorciTPS043
CorciTPS044
CorciTPS045
CorciTPS046
CorciTPS047
CorciTPS048
CorciTPS049
CorciTPS050
CorciTPS051
TPS-a
5′
3′
0kb
5kb
CorciTPS052
CorciTPS053
CorciTPS054
CorciTPS055
CorciTPS056
CorciTPS057
CorciTPS058
CorciTPS059
CorciTPS060
CorciTPS061
CorciTPS062
CorciTPS063
CorciTPS064
CorciTPS065
CorciTPS066
CorciTPS067
CorciTPS068
CorciTPS069
CorciTPS070
CorciTPS071
CorciTPS072
CorciTPS073
CorciTPS074
CorciTPS075
CorciTPS076
CorciTPS077
CorciTPS078
CorciTPS079
CorciTPS080
CorciTPS081
CorciTPS082
CorciTPS083
CorciTPS084
CorciTPS085
CorciTPS086
CorciTPS087
CorciTPS088
CorciTPS089
CorciTPS090
CorciTPS091
CorciTPS092
CorciTPS093
CorciTPS094
CorciTPS095
CorciTPS096
CorciTPS097
CorciTPS098
CorciTPS099
CorciTPS100
CorciTPS101
CorciTPS102
TPS-b1
TPS-b2
TPS-c
TPS-e
TPS-f
TPS-g
5′
3′
0kb
5kb

0.0 0.2 0.4 0.6
FPKM
FI
FB
BA
LE
LU
CorciTPS080-b2
CorciTPS079-b2
CorciTPS081-b2
CorciTPS023-a
CorciTPS011-a
CorciTPS013-a
CorciTPS012-a
CorciTPS028-a
CorciTPS050-a
CorciTPS086-b2
CorciTPS082-b2
CorciTPS087-b2
CorciTPS069-b1
CorciTPS078-b2
CorciTPS022-a
CorciTPS021-a
CorciTPS068-b1
CorciTPS051-a
CorciTPS074-b1
CorciTPS062-b1
CorciTPS040-a
CorciTPS005-a
CorciTPS004-a
CorciTPS024-a
CorciTPS076-b1
CorciTPS052-b1
CorciTPS045-a
CorciTPS044-a
CorciTPS002-a
CorciTPS096-g
CorciTPS027-a
CorciTPS043-a
CorciTPS031-a
CorciTPS097-g
CorciTPS007-a
CorciTPS036-a
CorciTPS017-a
CorciTPS006-a
CorciTPS018-a
CorciTPS019-a
CorciTPS020-a
CorciTPS029-a
CorciTPS030-a
CorciTPS033-a
CorciTPS037-a
CorciTPS049-a
CorciTPS053-b1
CorciTPS054-b1
CorciTPS060-b1
CorciTPS061-b1
CorciTPS067-b1
CorciTPS075-b1
CorciTPS090-f
CorciTPS091-f
CorciTPS092-f
CorciTPS093-f
CorciTPS100-g
CorciTPS102-g
CorciTPS098-g
CorciTPS099-g
CorciTPS088-c
CorciTPS048-a
CorciTPS003-a
CorciTPS085-b2
CorciTPS008-a
CorciTPS073-b1
CorciTPS072-b1
CorciTPS058-b1
CorciTPS064-b1
CorciTPS066-b1
CorciTPS014-a
CorciTPS055-b1
CorciTPS056-b1
CorciTPS001-a
CorciTPS094-g
CorciTPS095-g
CorciTPS065-b1
CorciTPS035-a
CorciTPS089-e
CorciTPS034-a
CorciTPS032-a
CorciTPS010-a
CorciTPS009-a
CorciTPS084-b2
CorciTPS063-b1
CorciTPS083-b2
CorciTPS026-a
CorciTPS025-a
CorciTPS046-a
CorciTPS042-a
CorciTPS047-a
CorciTPS070-b1
CorciTPS071-b1
CorciTPS016-a
CorciTPS059-b1
CorciTPS039-a
CorciTPS038-a
CorciTPS015-a
CorciTPS041-a
CorciTPS077-b1
CorciTPS057-b1
CorciTPS101-g