Title

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

Permalink

https://escholarship.org/uc/item/43x013sh

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

Heredity (2018) 121:87–104
https://doi.org/10.1038/s41437-018-0058-1

ARTICLE

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] · Mervyn Shepherd[9]

Received: 23 October 2017 / Revised: 17 December 2017 / Accepted: 12 January 2018 / Published online: 10 March 2018

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.

Introduction

Terpenes are an extensive group of hydrocarbon-based compounds present in most plants, with thousands currently characterised (Padovan et al. 2014). Although 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,

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-018-0058-1) contains supplementary material, which is available to authorized users.

✉ Jakob B. Butler
Jakob.Butler@utas.edu.au

[1] School of Natural 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, Brasilia 70770-917, 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, QLD 4072, Australia

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

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

The Myrtaceae, of the order Myrtales, is a family of plants that exhibit some of the highest concentrations and diversity of foliar terpenes in plants. Across the Myrtaceae, hundreds of compounds have been characterised (Padovan et al. 2014), with the foliage of individual trees often containing over 40 identifiable compounds (Keszei et al. 2008). Due to these features many Myrtaceous genera are key resources for commercial industries utilising terpenes as essential oils (Padovan et al. 2014) including *Melaleuca* (Keszei et al. 2010b), *Leptospermum* (Douglas et al. 2004), *Eucalyptus* and *Corymbia* (Batish et al. 2008). Along with *Angophora*, *Eucalyptus* and *Corymbia* are broadly classified as eucalypts (Slee et al. 2006). Eucalypts are the dominant trees in most Australian native forest and the predominant hardwood plantation species in Australia and overseas, due to their importance to the pulp, biofuel and timber industries (Rockwood et al. 2008; Shepherd et al. 2011). The characteristic smell of the eucalypts is due to their especially high concentration of foliar terpenes, with the high diversity of compounds present in this foliage extensively studied (Ammon et al. 1985; Lawler et al. 1998; Asante et al. 2001; Keszei et al. 2008). While the terpenoid profiles of most eucalypts are dominated by α-pinene and 1,8-cineole (Keszei et al. 2010a; Padovan et al. 2014), chemotype variation is important both to plant ecology (O'Reilly-Wapstra et al. 2004; Keszei 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 categorised 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 vs 640 MB, Grattapaglia and Bradshaw 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 five 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 RNAqueous 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,

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*.

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 organisation 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 homologue in *E. grandis* (evaluated by phylogenetic relatedness), verification of their position was undertaken in CCV54. The tool SYNFIND in CoGe (Lyons and Freeling 2008) was used to determine the likely position of homologous genes in CCV54 taking into account the synteny of the surrounding region. If gene position was conserved across both CCV genome assemblies, movement of loci relative to *E grandis* was considered real, while disagreements between the CCV assemblies were flagged as possible errors caused by misassembly, with more weight given to the loci position mirroring that of *E. grandis*.

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

RESULTS

Discovery of *TPS* loci

In the *Corymbia citriodora* subsp. *variegata* reference genome (CCV18) 127 loci were discovered with high sequence similarity to *terpene synthase* (*TPS*) genes from other species. Using a modified version of the classification method of Külheim et al. (2015), loci were classified into three categories: (i) 64 were full length with no structural abnormalities and had evidence of expression; (ii) 17 were full length, expressed but with up to two frame shifts or premature stop codons; and (iii) 21 were full length, had no evidence of expression and up to two frame shifts or premature stop codons. In accordance with Külheim et al. (2015), these were considered putatively functional *TPS* genes, resulting in a total of 102 genes (Table 1, Table S1)

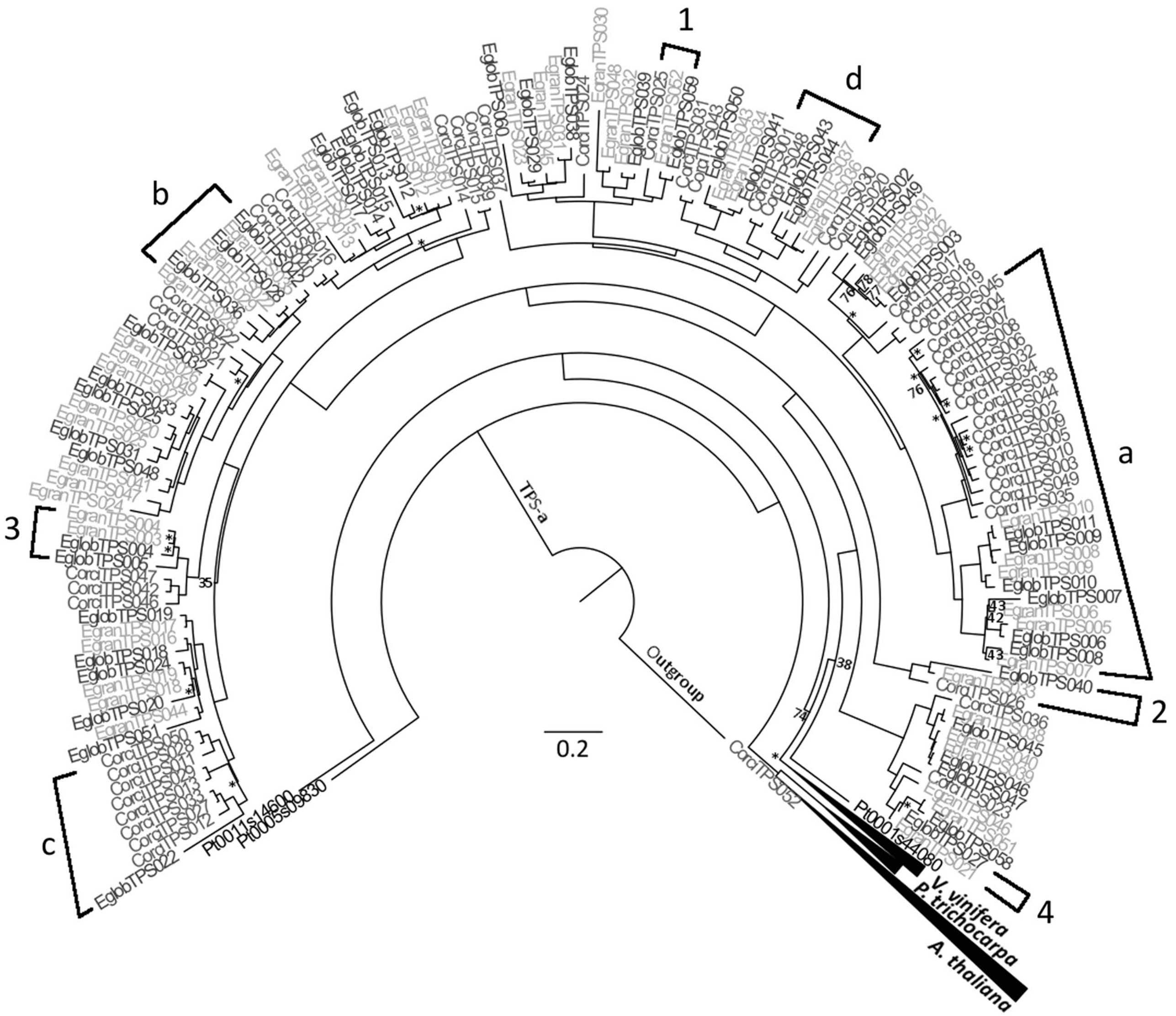

Fig. 1 Phylogeny of the *TPS-a* subfamily. This tree was created through maximum likelihood analysis comparing the *TPS-a* subfamily from *C. citriodora* subsp. *variegata* (Corci) with those from *E. grandis* (Egran), *E. globulus* (Eglob), *P. trichocarpa* (Pt), *V. vinifera* (Vv) and *A. thaliana* (At). Bootstrap values supported by <80% are noted by number, while those with bootstrap values between 80 and 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. **a**–**d** refers to results discussed further 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. **a** shows an example of a clade that is expanded in CCV relative to the other eucalypts. **b** gives an example of genes in orthologous pairings, with the exception of *EgranTPS029*, which 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. **d** shows a clade structure possibly resulting from concerted evolution

used in further analysis. The remaining 25 loci were classified as pseudogenes with more than two frame shifts or premature stop codons, with no consideration given to expression (Table S2). Similar analysis of the partially assembled CCV54 without expression data revealed 69 putative functional *TPS* genes and seven pseudogenes (Table S3). See Fig. 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* (Fig. 1), *TPS-b* and *TPS-g* (Fig. 2) and *TPS-c, TPS-e* and *TPS-f* (Fig. 3) subfamilies. The same *TPS* subfamilies were represented in each eucalypt species. Orthology (genes in different species directly descended

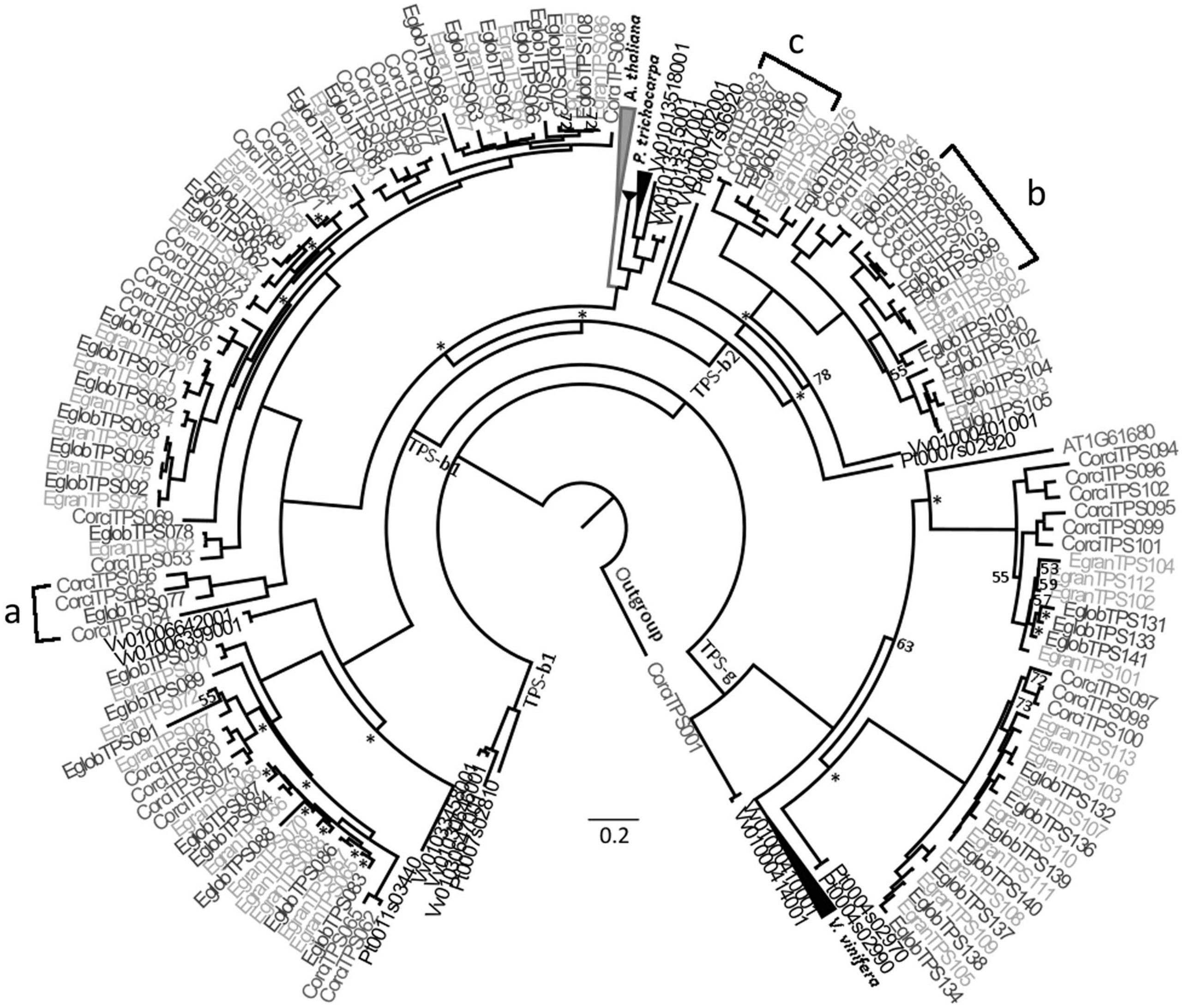

Fig. 2 Phylogeny of the *TPS-b* and *TPS-g* subfamilies. This tree was created through maximum likelihood analysis comparing the *TPS-b* and *TPS-g* subfamilies from *C. citriodora* subsp. *variegata* (Corci) with those from *E. grandis* (Egran), *E. globulus* (Eglob), *P. trichocarpa* (Pt), *V. vinifera* (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. A *TPS-a* gene from *C. citriodora* subsp. *variegata* was used as the outgroup. **a**–**c** refers to results discussed in the text

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 Fig. 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* [Fig. 1a]), or missing entirely (for example, the clade containing *EgranTPS029* [Fig. 1b]). 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* (Fig. 1c). 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*; Fig. 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 (Fig. 2). Only one *TPS-b1* gene (*CorciTPS053*) in CCV (4% of the total) was in an orthologous

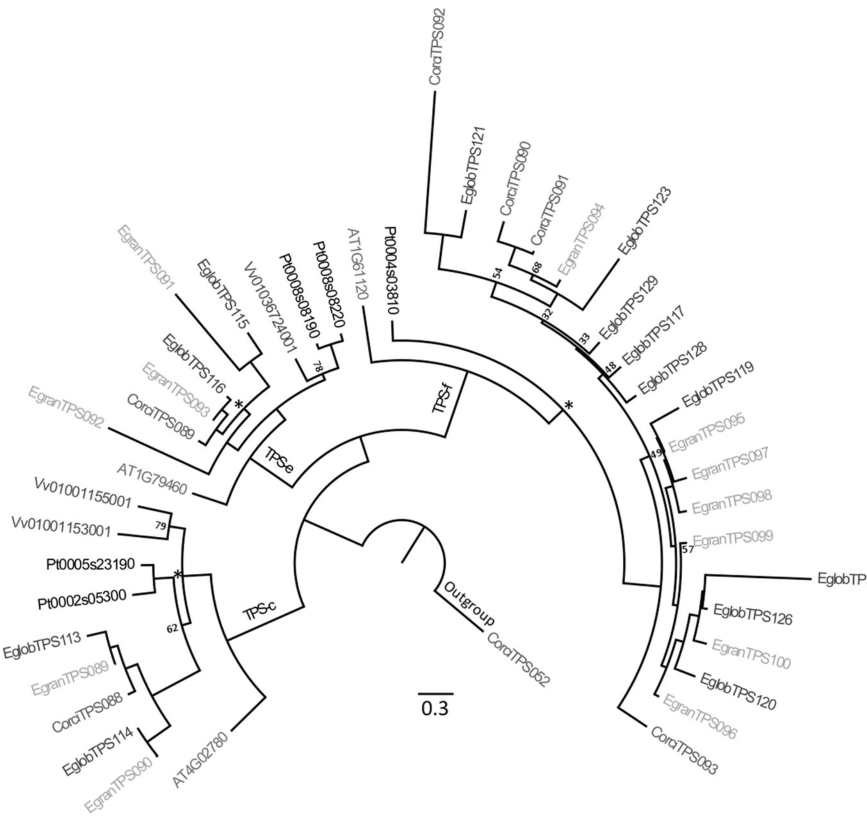

Fig. 3 Phylogeny of the *TPS-c, TPS-e* and *TPS-f* subfamilies. This tree was created through maximum likelihood analysis comparing the *TPS-c*, *TPS-e* and *TPS-f* subfamilies from *C. citriodora* subsp. *variegata* (Corci) with those from *E. grandis* (Egran), *E. globulus* (Eglob), *P. trichocarpa* (Pt), *V. vinifera* (Vv), and *A. thaliana* (At). Bootstrap values supported by <80% are noted by number, while those with bootstrap values between 80 and 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

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 (Fig. 2a). Of the *TPS-b2* genes (Fig. 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 (Fig. 2b, 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 (Fig. 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 (Fig. 3).

The *TPS-f* subfamily was more dynamic than the other subfamilies involved in primary metabolism (Fig. 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 (≤ 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 to 0.0031 e/g/my (Figure S1-a). In contrast, the eucalypt lineage was estimated to have experienced a two orders of magnitude higher rate of gene birth (0.0282 e/g/my). Within the eucalypt lineages, death rate was similar in both *Eucalyptus* and *Corymbia*

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

(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 ($\chi^2_4 = 3.69$, $P > 0.05$ [combining *TPS-c*, *-e*, and *-f* due to sample size]), or the number of genes involved in primary versus secondary metabolism ($\chi^2_1 = 2.41$, $P > 0.05$). A similar lack of significant difference was observed between *E. grandis* and *E. globulus* in the number of loci at the subfamily ($\chi^2_4 = 1.53$, $P > 0.05$) or primary *versus* secondary metabolite ($\chi^2_1 = 0.3$, $P > 0.05$) levels, providing evidence that the broad features of this gene family are conserved between *Eucalyptus* and *Corymbia*.

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

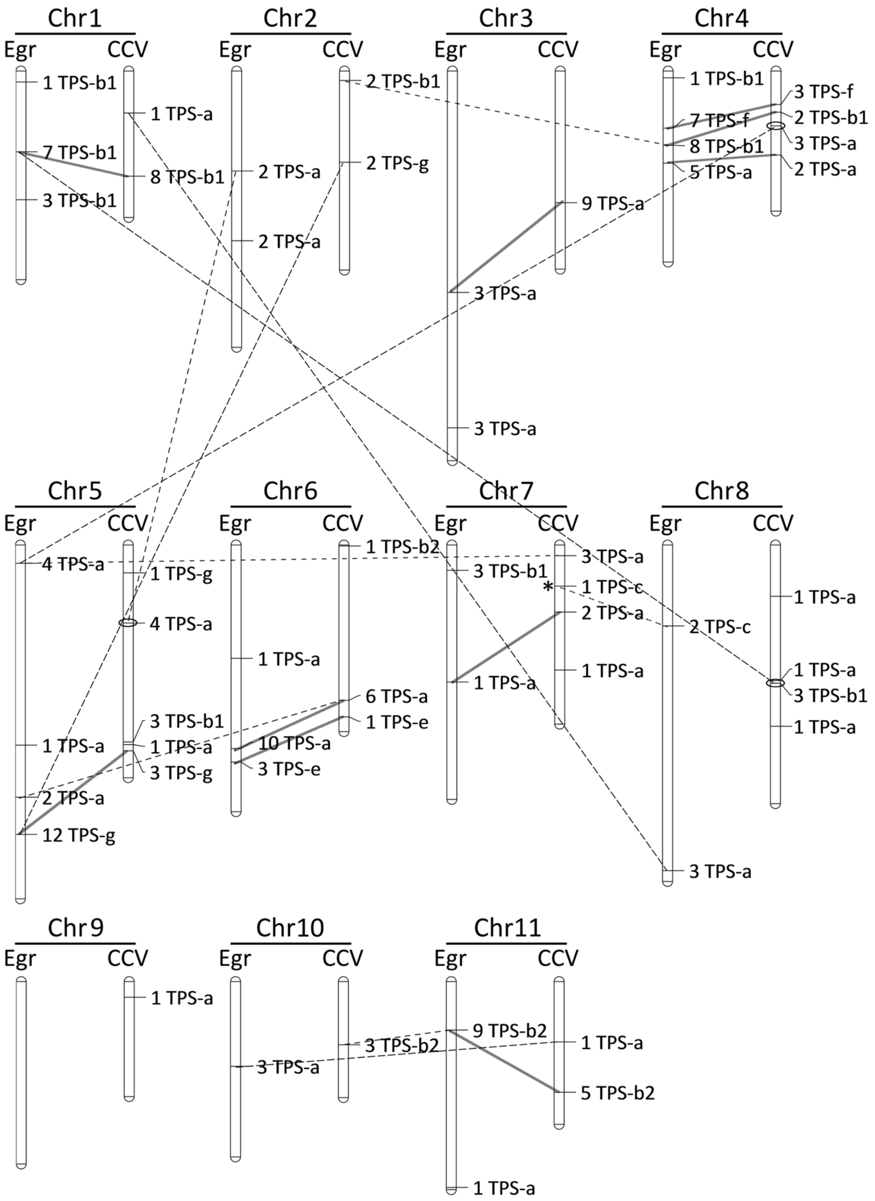

Fig. 4 Comparison of copy number and genomic location of *TPS* physical clusters between *E. grandis* (Egr) and *C. citriodora* subsp. *variegata* (CCV). Chromosomes are scaled by physical size. Locus names show the number of *TPS* genes and the subfamily they belong to. Separate clusters on the same 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 homologue is present in the minor scaffolds of the other species and cannot be examined for synteny. Homology of singleton *TPS* genes is not shown

Across all *TPS* subfamilies, 10 physical clusters were both syntenic and phylogenetically similar between *E. grandis* and CCV18 (Fig. 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

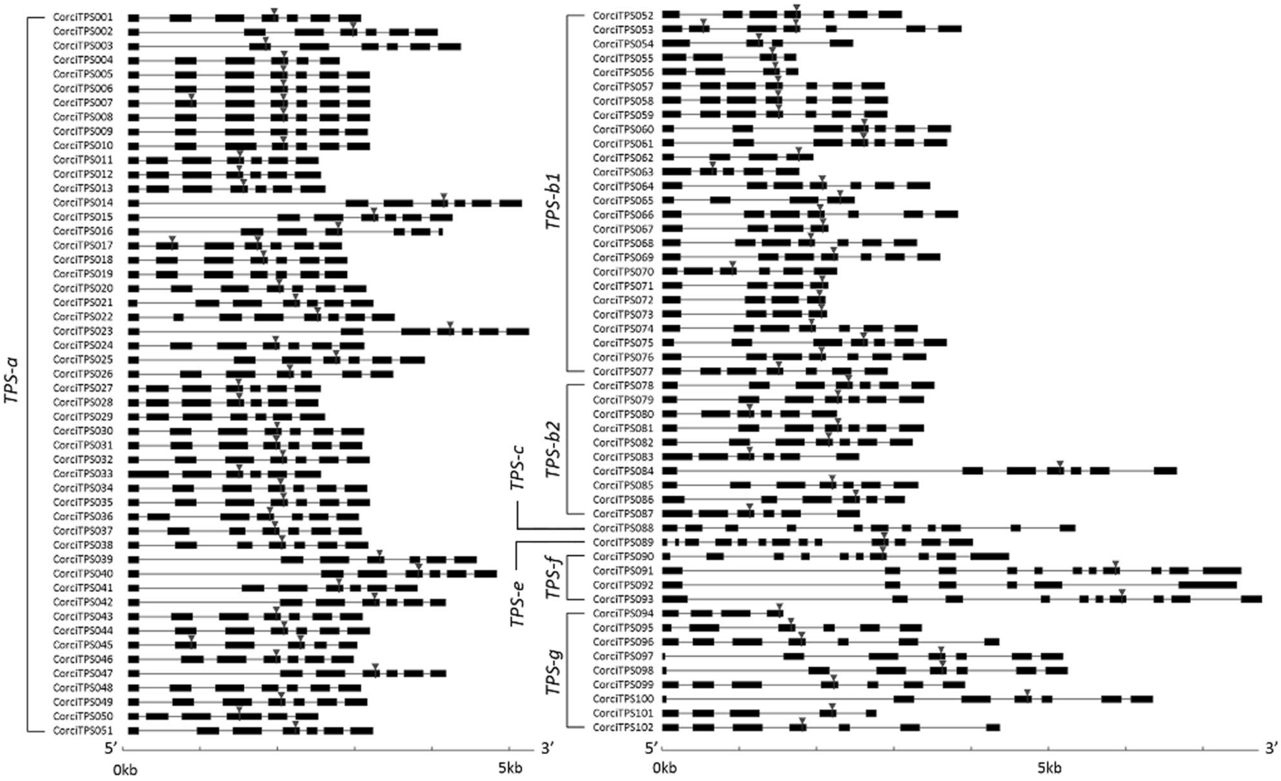

Fig. 5 Gene structure of the 102 putative functional *TPS* genes from *Corymbia citriodora* subsp. *variegata*. Exons are shown as boxes, while introns are shown as lines. The arrow indicates the position of the conserved DDxxD motif

the CCV reference genome supported by the second CCV genome assembly CCV54 (Fig. 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 (Fig. 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 1564 to 7747 bp, with final products ranging from 337 to 739 amino acids in length (Table S1).

TPS gene expression

A heat map showing relative transcript abundance in five tissues is shown in Fig. 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.

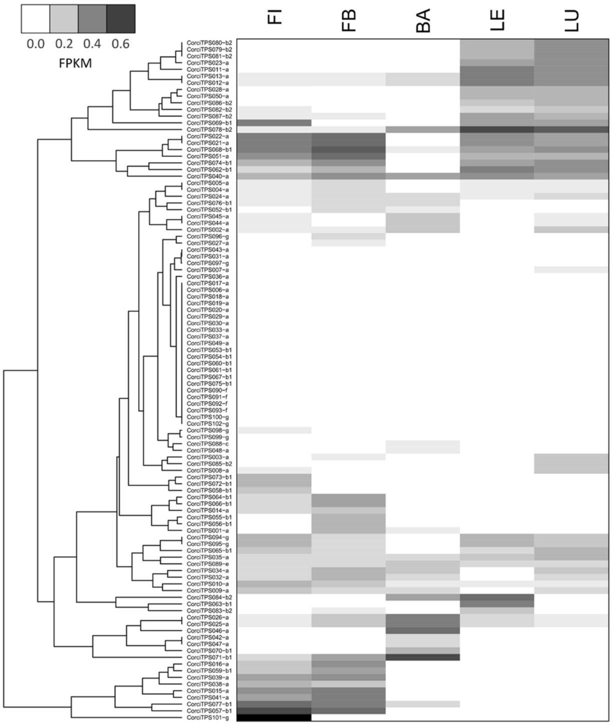

Fig. 6 Gene expression clustering of 102 *TPS* genes from *Corymbia citriodora* subsp. *variegata* expressed in five tissues. RNAseq data is shown as fragments per kilobase of transcript per million mapped reads (FPKM), with FPKM values normalised within libraries (largest FPKM value set to 1, with other scores scaled accordingly). The *TPS* subfamily is indicated by suffix after the gene name. The sampled tissues are: flower initials (FI), flower buds (FB), bark (BA), expanded leaf (LE), and unexpanded leaf (LU)

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 ~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: neofunctionalization, 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* (Fig. 1a). 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 Figs. 1d and 2c, 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 (Fig. 2d, 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* (Figs. 1 and 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 (Fig. 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, Fig. 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 in this lineage. Of the 25 pseudogenes, 15 showed evidence of expression, which is an interesting finding. While pseudogenes are thought to play at most a passive role in the genome, such as being sequence donor/receptors for proximal genes (Zheng and Gerstein 2007), it has been shown that some pseudogene transcripts in humans can bind to mRNA from related functional genes and affect their expression (Vinckenbosch et al. 2006). If this is the case in *Corymbia*, these pseudogenes may be part of a mechanism for modulating gene expression.

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

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.

Acknowledgements 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.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

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

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

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

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

Aubourg S, Lecharny A, Bohlmann J (2002) Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana* Mol Genet Genomics 267:730–745

Batish DR, Singh HP, Kohli RK, Kaur S (2008) *Eucalyptus* essential oil as a natural pesticide. Forest Ecol Manag 256:2166–2174

Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome Res 10:547–548

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

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

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

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

Carr D, Carr S (1970) Oil glands and ducts in *Eucalyptus* L'Héit. II. Development and structure of oil glands in the embryo. Aust J Botany 18:191–212

Chaw S-M, Chang C-C, Chen H-L, Li W-H (2004) Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424–441

Chen F, Tholl D, Bohlmann J, Pichersky E (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J 66:212–229

Chen J-M, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease Nature Rev Genet 8:762–775

Chu HY, Wegel E, Osbourn A (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants Plant J 66:66–79

Crisp MD, Burrows GE, Cook LG, Thornhill AH, Bowman DMJS (2011) Flammable biomes dominated by eucalypts originated at the Cretaceous-Palaeogene boundary Nat Commun 2:193

del Moral R, Muller CH (1970) The allelopathic effects of *Eucalyptus camaldulensis* Am Midl Nat 83:254–282

Demuth JP, Hahn MW (2009) The life and death of gene families. Bioessays 31:29–39

Douglas MH, van Klink JW, Smallfield BM, Perry NB, Anderson RE, Johnstone P et al. (2004) Essential oils from New Zealand manuka: triketone and other chemotypes of *Leptospermum scoparium*. Phytochemistry 65:1255–1264

Ee S-F, Mohamed-Hussein Z-A, Othman R, Shaharuddin NA, Ismail I, Zainal Z (2014) Functional characterization of sesquiterpene synthase from *Polygonum minus* Sci World J 2014:11

Eitas TK, Dangl JL (2010) NB-LRR proteins: pairs, pieces, perception, partners, and pathways. Curr Opin Plant Biol 13:472–477

Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions Proc Natl Acad Sci USA 108:16116–16121

Field B, Osbourn AE (2008) Metabolic diversification - independent assembly of operon-like gene clusters in different plants. Science 320:543–547

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants New Phytol 183:557–564

Fray RG, Wallace A, Fraser PD, Valero D, Hedden P, Bramley PM et al. (1995) Constitutive expression of a fruit phytoene synthase gene in transgenic tomatoes causes dwarfism by redirecting metabolites from the gibberellin pathway Plant J 8:693–701

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60:433–453

Gao Y, Honzatko RB, Peters RJ (2012) Terpenoid synthase structures: a so far incomplete view of complex catalysis. Nat Prod Rep 29:1153–1175

Gijzen M (2009) Runaway repeats force expansion of the *Phytophthora infestans* genome. Genome Biol 10:241

Goodger JQ, Heskes AM, Woodrow IE (2013) Contrasting ontogenetic trajectories for phenolic and terpenoid defences in *Eucalyptus froggattii*. Ann Bot (Lond) 112:651–659

Grattapaglia D, Bradshaw Jr HD (1994) Nuclear DNA content of commercially important *Eucalyptus* species and hybrids Can J For Res 24:1074–1078

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C et al. (2012) Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus Tree Genet Genomes 8:463–508

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al. (2013) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc 8:https://doi.org/10.1038/nprot.2013.1084

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148:993–1003

Hansen NL, Heskes AM, Hamberger B, Olsen CE, Hallström BM, Andersen-Ranberg J et al. (2017) The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily Plant J 89:429–441

Hayashi K-i, Kawaide H, Notomi M, Sakigi Y, Matsuo A, Nozaki H (2006) Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *Physcomitrella patens*. FEBS Lett 580:6175–6181

Healey A, Shepherd M, Baten A, King GJ, Lee DJ, Furtado A et al. (2017) Sequencing the branches of the eucalypt tree: comparison between *Eucalyptus* and *Corymbia* genomes. In: Plant & Animal Genome Conference XXV, San Diego, United States of America.

Heiling S, Schuman MC, Schoettner M, Mukerjee P, Berger B, Schneider B et al. (2010) Jasmonate and ppHsystemin regulate key malonylation steps in the biosynthesis of 17-hydroxygeranyllinalool diterpene glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. Plant Cell 22:273–292

Hill KD, Johnson LA (1995) Systematic studies in the eucalypts 7. A revision of the bloodwoods, genus *Corymbia* (Myrtaceae). Telopea 6:185–504

Hofberger JA, Lyons E, Edger PP, Chris Pires J, Eric Schranz M (2013) Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family Genome Biol Evol 5:2155–2173

Hofberger JA, Nsibo DL, Govers F, Bouwmeester K, Schranz ME (2015) A complex interplay of tandem- and whole-genome duplication drives expansion of the L-type lectin receptor kinase gene family in the Brassicaceae Genome Biol Evol 7:720–734

Hosfield DJ, Zhang Y, Dougan DR, Broun A, Tari LW, Swanson RV et al. (2004) Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis J Biol Chem 279:8526–8529

Hurles M (2004) Gene duplication: the genomic trade in spare parts. PLoS Biol 2:e206

Irmisch S, Jiang Y, Chen F, Gershenzon J, Köllner TG (2014) Terpene synthases and their contribution to herbivore-induced volatile emission in western balsam poplar (*Populus trichocarpa*). BMC Plant Biol 14:270

Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human adaptation Trends Genet 28:245–257

Keszei A, Brubaker CL, Foley WJ (2008) A molecular perspective on terpene variation in Australian Myrtaceae Aust J Bot 56:197–213

Keszei A, Brubaker CL, Carter R, Köllner T, Degenhardt J, Foley WJ (2010a) Functional and evolutionary relationships between terpene synthases from Australian Myrtaceae Phytochemistry 71:844–852

Keszei A, Hassan Y, Foley WJ (2010b) A biochemical interpretation of terpene chemotypes in *Melaleuca alternifolia* J Chem Ecol 36:652–661

Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T (2001) Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate–dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell 13:681–694

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc B: Biol Sci 279:5048–5057

Külheim C, Padovan A, Hefer C, Krause ST, Köllner TG, Myburg AA et al. (2015) The *Eucalyptus* terpene synthase gene family. BMC Genomics 16:1–18

Ladiges PY, Udovicic F, Nelson G (2003) Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. J Biogeogr 30:989–998

Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K (1998) Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials. Oecologia 116:160–169

Lawler IR, Stapley J, Foley WJ, Eschler BM (1999) Ecological example of conditioned flavor aversion in plant–herbivore interactions: effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums. J Chem Ecol 25:401–415

Lee DJ (2007) Achievements in forest tree genetic improvement in Australia and New Zealand 2: Development of *Corymbia* species and hybrids for plantations in eastern Australia. Aust Forestry 70:11–16

Lefort V, Longueville J-E, Gascuel O (2017) SMS: Smart Model Selection in PhyML. Mol Biol Evol 34:2422–2424

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes Trends Genet 20:116–122

Li F, Zhou C, Weng Q, Li M, Yu X, Guo Y et al. (2015) Comparative genomics analyses reveal extensive chromosome colinearity and novel quantitative trait loci in *Eucalyptus*. PLoS ONE 10: e0145144

Librado P, Vieira FG, Rozas J (2012) BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281

Liu Y, Schiff M, Dinesh-Kumar S (2004) Involvement of MEK1 MAPKK, NTF6 MAPK, WRKY/MYB transcription factors, COI1 and CTR1 in N-mediated resistance to tobacco mosaic virus. Plant J 38:800 –809

Lynch M (2007) The origins of genome architecture. Sinauer Associates, Inc, Sunderland, MA

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53:661–673

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H et al. (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol 148:1772–1781

Macphail M, Thornhill AH (2016) How old are the eucalypts? A review of the microfossil and phylogenetic evidence Aust J Bot 64:579–599

Martin DM, Aubourg S, Schouwey MB, Daviet L, Schalk M, Toub O et al. (2010) Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. BMC Plant Biol 10:226

Mendivil-Ramos O, Ferrier DEK (2012) Mechanisms of gene duplication and translocation and progress towards understanding their relative contributions to animal genome evolution. Int J Evolut Biol 2012:10

Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J et al. (2014) The genome of *Eucalyptus grandis*. Nature 510:356–362

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274

O'Reilly-Wapstra JM, McArthur C, Potts BM (2004) Linking plant genotype, plant defensive chemistry and mammal browsing in a *Eucalyptus* species. Funct Ecol 18:677–684

O'Reilly-Wapstra JM, Freeman JS, Davies NW, Vaillancourt RE, Fitzgerald H, Potts BM (2011) Quantitative trait loci for foliar terpenes in a global eucalypt species. Tree Genet Genomes 7:485–498

Padovan A, Keszei A, Külheim C, Foley WJ (2014) The evolution of foliar terpene diversity in Myrtaceae. Phytochem Rev 13:695–716

Padovan A, Keszei A, Wallis IR, Foley WJ (2012) Mosaic eucalypt trees suggest genetic control at a point that influences several metabolic pathways. J Chem Ecol 38:914–923

Peñuelas J, Llusià J, Asensio D, Munné-Bosch S (2005) Linking isoprene with plant thermotolerance, antioxidants and monoterpene emissions. Plant Cell Environ 28:278–286

Pichersky E, Gershenzon J (2002) The formation and function of plant volatiles: perfumes for pollinator attraction and defense. Curr Opin Plant Biol 5:237–243

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rockwood D, Rudie A, Ralph S, Zhu J, Winandy J (2008) Energy product options for *Eucalyptus s*pecies grown as short rotation woody crops. Int J Mol Sci 9:1361

Sasaki K, Saito T, Lämsä M, Oksman-Caldentey K-M, Suzuki M, Ohyama K et al. (2007) Plants utilize isoprene emission as a thermotolerance mechanism. Plant Cell Physiol 48:1254–1262

Schiffer PH, Gravemeyer J, Rauscher M, Wiehe T (2016) Ultra large gene families: a matter of adaptation or genomic parasites? Life 6:32

Schnee C, Köllner TG, Held M, Turlings TCJ, Gershenzon J, Degenhardt J (2006) The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. Proc Natl Acad Sci USA 103:1129–1134

Schwab W, Fuchs C, Huang F-C (2013) Transformation of terpenes into fine chemicals. Eur J Lipid Sci Technol 115:3–8

Sharma M, Pandey GK (2015) Expansion and function of repeat domain proteins during stress and development in plants. Front Plant Sci 6:1218

Shepherd M, Bartle J, Lee DJ, Brawner J, Bush D, Turnbull P et al. (2011) Eucalypts as a biofuel 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 Physiol 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 Phytol 206:1364–1377

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then… and now: Stebbins revisited. Am J Bot 101:1057–1078

Szöllősi GJ, Daubin V (2012) Modeling gene family evolution and reconciling phylogenetic discord. Evolut Genomics: Stat Comput Methods 2:29–51

Takos AM, Rook F (2012) Why biosynthetic genes for chemical defense compounds cluster. Trends Plant Sci 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. Mol Phylogenet Evol 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. J Essential Oil Res 16:560–565

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103:3220–3225

Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. Ann N Y Acad Sci 1256:1–14

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T et al. (2016) gplots: various R programming tools for plotting data. https://CRAN.R-project.org/package=gplots

Wenke K, Kai M, Piechulla B (2010) Belowground volatiles facilitate interactions between plant roots and soil organisms. Planta 231:499–506

Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc Royal Soc B 268:2211–2220

Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM (2009) Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors. Plant Physiol 149:981–993

Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res 14:1060–1067

Xiong W, Wu P, Jia Y, Wei X, Xu L, Yang Y et al. (2016) Genome-wide analysis of the terpene synthase gene family in physic nut (*Jatropha curcas* L.) and functional identification of six terpene synthases. Tree Genet Genomes 12:97

Zhang S-D, Ling L-Z, Yi T-S (2015) Evolution and divergence of SBP-box genes in land plants. BMC Genomics 16:787

Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 23:219–224

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127:1–18