















## Pests, diseases, and aridity have shaped the genome of *Corymbia citriodora*

Adam L. Healey <sup>1,2✉</sup>, Mervyn Shepherd <sup>3</sup>, Graham J. King <sup>3</sup>, Jakob B. Butler <sup>4</sup>, Jules S. Freeman <sup>4,5,6</sup>, David J. Lee <sup>7</sup>, Brad M. Potts<sup>4,5</sup>, Orzenil B. Silva-Junior<sup>8</sup>, Abdul Baten <sup>3,9</sup>, Jerry Jenkins <sup>1</sup>, Shengqiang Shu <sup>10</sup>, John T. Lovell <sup>1</sup>, Avinash Sreedasyam<sup>1</sup>, Jane Grimwood <sup>1</sup>, Agnelo Furtado<sup>2</sup>, Dario Grattapaglia<sup>8,11</sup>, Kerrie W. Barry<sup>10</sup>, Hope Hundley<sup>10</sup>, Blake A. Simmons <sup>2,12</sup>, Jeremy Schmutz <sup>1,10</sup>, René E. Vaillancourt<sup>4,5</sup> & Robert J. Henry <sup>2</sup>

*Corymbia citriodora* is a member of the predominantly Southern Hemisphere Myrtaceae family, which includes the eucalypts (*Eucalyptus*, *Corymbia* and *Angophora*; ~800 species). *Corymbia* is grown for timber, pulp and paper, and essential oils in Australia, South Africa, Asia, and Brazil, maintaining a high-growth rate under marginal conditions due to drought, poor-quality soil, and biotic stresses. To dissect the genetic basis of these desirable traits, we sequenced and assembled the 408 Mb genome of *Corymbia citriodora*, anchored into eleven chromosomes. Comparative analysis with *Eucalyptus grandis* reveals high synteny, although the two diverged approximately 60 million years ago and have different genome sizes (408 vs 641 Mb), with few large intra-chromosomal rearrangements. *C. citriodora* shares an ancient whole-genome duplication event with *E. grandis* but has undergone tandem gene family expansions related to terpene biosynthesis, innate pathogen resistance, and leaf wax formation, enabling their successful adaptation to biotic/abiotic stresses and arid conditions of the Australian continent.

<sup>1</sup> HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. <sup>2</sup> University of Queensland/QAAFI, Brisbane, QLD, Australia. <sup>3</sup> Southern Cross Plant Science, Southern Cross University, Lismore, NSW, Australia. <sup>4</sup> School of Natural Sciences, University of Tasmania, Hobart, TAS, Australia. <sup>5</sup> ARC Training Centre for Forest Value, University of Tasmania, Hobart, TAS, Australia. <sup>6</sup> Scion, Rotorua, New Zealand. <sup>7</sup> Forest Industries Research Centre, University of the Sunshine Coast, Sippy Downs, QLD, Australia. <sup>8</sup> EMBRAPA Genetic Resources and Biotechnology, Brasília, Brazil. <sup>9</sup> Institute of Precision Medicine & Bioinformatics, Camperdown, NSW, Australia. <sup>10</sup> Department of Energy Joint Genome Institute, Berkeley, CA, USA. <sup>11</sup> Genomic Science Program, Universidade Católica de Brasília, Taguatinga, Brazil. <sup>12</sup> Joint BioEnergy Institute, Emeryville, CA, USA. ✉email: [ahealey@hudsonalpha.org](mailto:ahealey@hudsonalpha.org)

Embedded within genomes are the footprints of climatic and evolutionary history in which progenitor lineages have undergone selection<sup>1</sup>. The detection of these footprints can provide insight into the historic conditions experienced by organisms of interest, with plant genomes in particular often exhibiting distinct adaptive signatures due to their sessile nature<sup>2</sup>. Forest trees, as some of the longest-lived plants and therefore exhibiting strong local adaptation, are important for the renewable delivery of materials and energy worldwide, play a key role in carbon cycling and storage, and affect rainfall patterns<sup>3</sup>. Angiosperms (flowering plants) abound with tree species that occur in most taxonomic orders<sup>4</sup>. However, our current insights into the evolution of their genomes are primarily based on comparative analysis of Northern Hemisphere deciduous taxa, within families such as the Rosaceae<sup>5–7</sup>, Salicaceae<sup>8</sup>, Fagaceae<sup>9,10</sup>, and Oleaceae<sup>11,12</sup>. Although *Eucalyptus grandis* was the second forest tree genome to be assembled<sup>13</sup>, there has been little progress in unravelling key aspects of genome organization and evolution within the predominantly Southern Hemisphere family Myrtaceae to which it belongs. The Myrtaceae is a diverse, ecologically and economically important plant lineage (~5,700 species; 132 genera<sup>14</sup>) that includes tree species such as clove (*Syzygium*), guava (*Psidium*), tea-trees (*Melaleuca* and *Leptospermum*) and mangroves (*Osbornia*)<sup>15</sup>. It also includes the globally grown eucalypts, which are endemic to Australia and islands to its north<sup>16</sup>.

Eucalypts comprise over 800 species, belonging to three closely related genera—*Angophora*, *Corymbia*, and *Eucalyptus*<sup>17</sup>. Eucalypts diverged from their closest Myrtaceae relative, Syncarpia, approximately 65–68 million years ago (MYA)<sup>18</sup> and radiated into diverse environments undergoing rapid expansion immediately after the Cretaceous, followed by domination during the Paleocene-Eocene thermal maximum (~55 MYA) and climate aridification in the mid-late Miocene (~15 MYA)<sup>19–21</sup>. *Eucalyptus*, the largest genus, diverged from the *Angophora-Corymbia* lineage ~60 MYA, which roughly corresponds to the separation of the Australian continent from Antarctica [83–45 MYA], at which time the two had long separated from other Gondwana land masses<sup>16,22</sup>. *Eucalyptus* is widely distributed across the Australian continent but is largely consolidated in the more southern bioregions<sup>23</sup>. In contrast, *Corymbia* and *Angophora* are largely absent from most southern forests, having radiated through coastal and sub-coastal regions of eastern Australia, with *Corymbia* also extending across the northern, tropical ‘top-end’ of Australia<sup>24</sup>. These differences in the geographic range likely reflect evolutionary, adaptive differences between the *Angophora-Corymbia* and *Eucalyptus* lineages. Climate niche adaptation is signaled by field trials showing *Corymbia* species are typically more cold-averse and drought tolerant than *Eucalyptus*<sup>25,26</sup> and thrive in a wide range of rainfall conditions (0.6–2.0 m/year)<sup>27</sup> and marginal soils<sup>28</sup>. In terms of biotic environment, insects and fungal diseases have represented the primary pest challenge for both lineages, with genus-level differences in susceptibility often evident<sup>29,30</sup>. The emblematic defensive strategy taken within the Myrtaceae has been the generation of a diverse range of terpenoids<sup>31</sup>, with complex profiles matching their diversity and a corresponding expansion of the terpene synthase gene family<sup>32,33</sup>.

Here we present the genome assembly of *C. citriodora* subsp. *variegata* (CCV), a new Myrtaceae reference sequence for a taxa important for timber, pulp and paper, carbon sequestration, and essential oil production in areas considered too marginal for other productive species due to pests, diseases, and drought. Despite their divergence and adaptive radiation across different biomes in Australia, the genome structure among *Eucalyptus grandis* and CCV is highly conserved ( $2n = 22$ ). *Corymbia* has retained evidence of an ancient (109 MYA) Myrtales whole-genome

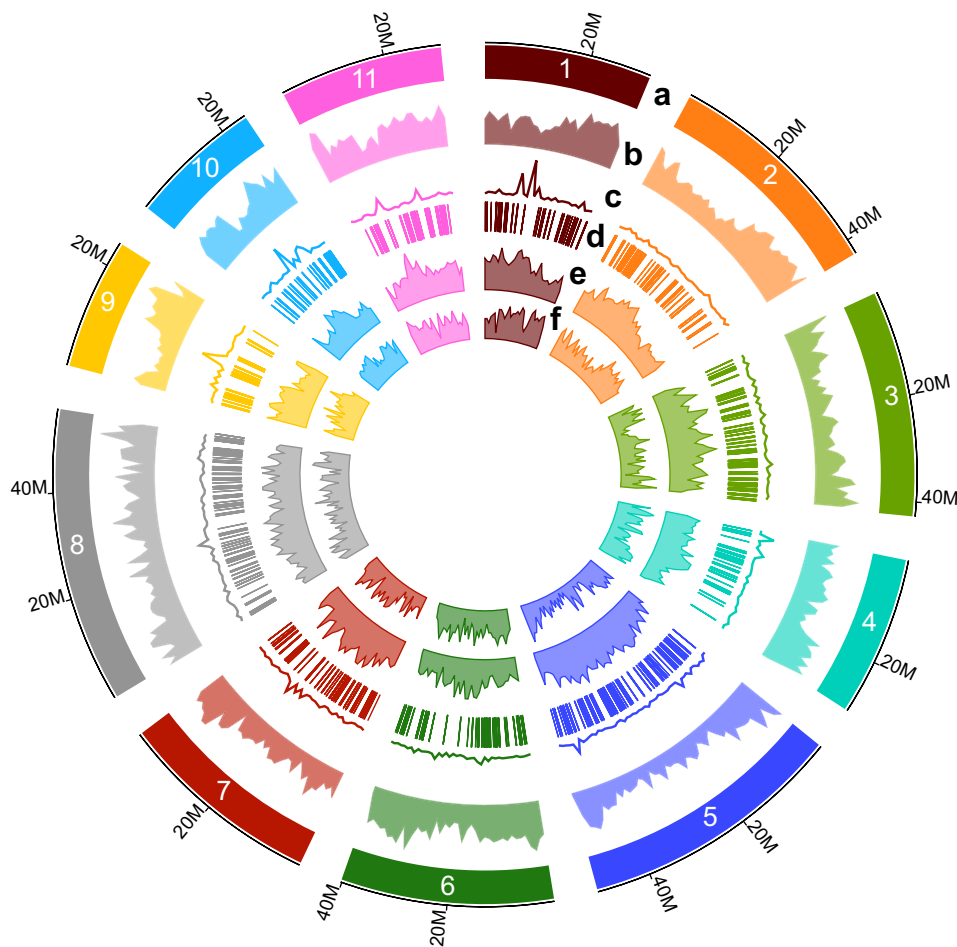
duplication (WGD) event and exhibits post-divergence gene family expansions related to terpene synthesis and biotic/abiotic stress resistance. This new genome sequence will enable comparative genomic studies for the dominant hardwood taxa in the Southern Hemisphere and will serve as a valuable resource for further development of this strategic woody biomass resource for manufacturing and bioenergy sectors.

## Results

**Genome assembly and annotation.** *Corymbia citriodora* subsp. *variegata* genotype CCV2-018 was selected for reference sequencing due to its wide use as a parent in the spotted gum breeding program of the Queensland Department of Agriculture and Fisheries, and its use for the generation of interspecific hybrids for investigating pulp and bioenergy production<sup>34</sup>. In brief, 129 Gb of raw data was generated from two Illumina HiSeq2500 libraries (2 × 150 bp paired end; insert sizes: 400 and 800 bp), representing ~320× sequencing coverage of the genome. The genome assembly was generated using a modified version of Arachne (v.20071016)<sup>35</sup>. Contig assembly and initial scaffolding steps produced 37,263 contigs in 32,740 scaffolds (N50 length: 132.6 Kb), totaling 563.0 Mb. Gap patching on the scaffolds was performed using ~25× PacBio reads (N50 length: 17,094 bp) and QUIVER ([www.github.com/PacificBiosciences/GenomicConsensus](http://www.github.com/PacificBiosciences/GenomicConsensus)). Final scaffolding was completed using SSPACE-Standard<sup>36</sup> (Version 2.0) with Nextera long mate pair libraries (insert size 4 Kb and 8 Kb), resulting in a 537.9 Mb assembly (16,786 scaffolds; 20,979 contigs) with a scaffold N50 of 312 Kb.

To anchor the scaffolds into chromosomes, the sequences were ordered and oriented into 11 pseudomolecules (Fig. 1; Supplementary Data File 1) using *Corymbia* genetic maps<sup>37</sup>. Three high-density linkage maps were generated from two *C. torelliana* × *C. citriodora* subsp. *variegata* hybrid crosses (CT2-050 × CCV2-054, CT2-018 × CCV2-054) genotyped with Diversity Arrays DArTseq technology<sup>38</sup>, and contigs were anchored to the marker sequences using ALLMAPS<sup>39</sup> (Supplementary Fig. 1). The average Spearman correlation coefficient of centimorgan (cM) positions for genetic map markers from all three linkage maps and physical locations on scaffolds was 0.96. The pseudomolecules range in size from 24.8 Mb (Chromosome 9) to 55.7 Mb (Chromosome 8). The total genome size of chromosome anchored scaffolds ( $n = 4,033$ ) was 412 Mb (408 Mb in contigs), which is close to the estimated genome size of 370–390 Mb, based on flow cytometry<sup>40</sup>. Global genome heterozygosity was estimated at ~0.5% through calling heterozygous SNPs against repeat-masked bases in chromosomes. The evaluation of the protein-coding annotation completeness with single-copy orthologs on the assembly was undertaken using Benchmarking Universal Single-Copy Orthologs (BUSCO; Supplementary Table 1)<sup>41</sup>, receiving a 95.1% score, suggesting high quality and completeness. In addition, 90% of derived single-copy genes in *E. grandis* were also single copy on CCV chromosomes, suggesting that pseudomolecule construction was complete and alternative haplotypes had not been introduced into the main genome assembly (Supplementary Fig. 2).

To annotate the genome, RNA was collected from five separate tissues (expanded leaves [EL], unexpanded leaves [UL], flower buds [FB], flower initials [FI], photosynthetic bark cortex [BA]) (Supplementary Figs. 3, 4) and was used for de novo gene model prediction. The final annotation of protein-coding gene products comprised 35,632 primary transcripts and 10,019 alternative transcripts for a total of 45,651 transcript models. The set of primary transcripts had a mean length of 3.4 Kb, a mean of 4.8 exons, with a median exon length of 176 bp and a median intron length of 202 bp. The total amount of repetitive content captured



**Fig. 1 Main genome features of *Corymbia citriodora* subspecies *variegata* (CCV).** **a** The eleven chromosomes of CCV. Numbers along the outside track denote chromosome length in megabases. **b** Gene density (gene number per megabase; 5–143). **c** Average expression (rpkm; 0.5–150) among collected tissues for RNASeq per megabase. **d** Tandem gene arrays. **e** Percent repetitive content per megabase (5–61%). **f** Number of heterozygous SNPs per megabase (230–7,282).

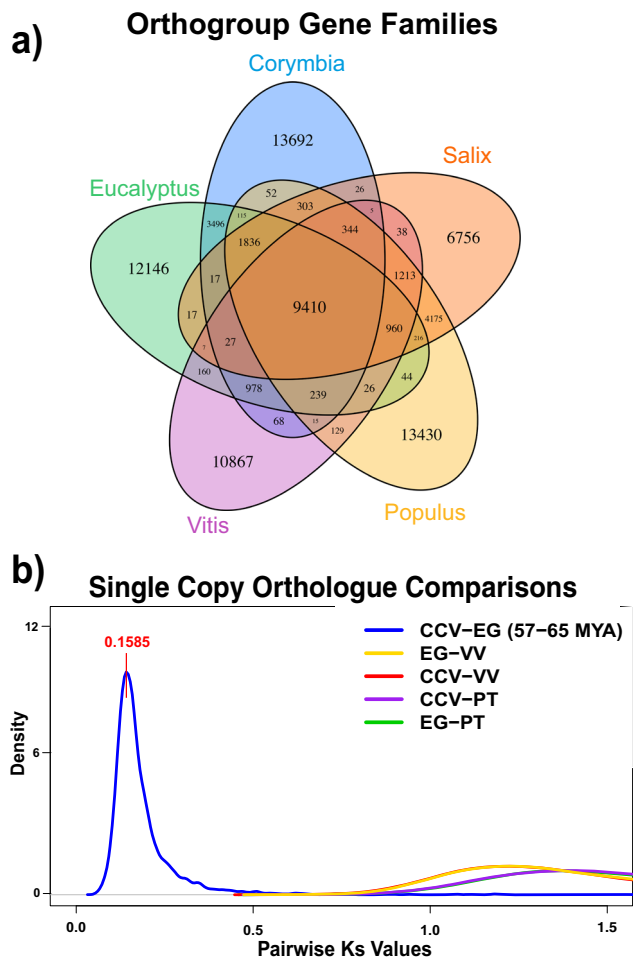
**Table 1 Assembly and genome statistics for *Corymbia citriodora* subspecies *variegata* (v2.1- Phytozome v13) and *Eucalyptus grandis* (v2.0- Phytozome v13).**

|  | <i>C. citriodora</i> ssp. <i>variegata</i> | <i>E. grandis</i> |
|--|--|-------------------|
| Bases in chromosomes (Megabases [Mb])      | 408  | 641               |
| Number of scaffolds in chromosomes         | 4,033                                      | 4,952             |
| Number of chromosomes                      | 11   | 11                |
| Contig N50 length                          | 185.5 Kb                                   | 67.2 Kb           |
| Scaffold N50 length                        | 31.5 Mb                                    | 57.5 Mb           |
| GC content                                 | 39.1%                                      | 39.3%             |
| Repetitive content                         | 35.78%                                     | 43.96%            |
| Retro transposable elements (RNA; Class I) | 19.48%                                     | 22.06%            |
| DNA transposable elements (Class II)       | 5.42%                                      | 7.22%             |
| Total number of primary gene models        | 35,632                                     | 36,349            |
| Total number of transcript models          | 45,651                                     | 46,280            |

in the pseudomolecules was 146.5 Mb, which represents ~35.8% of the genome (Table 1). The repetitive content was primarily comprised of Class I Retro transposable elements (19.48%) and Class II DNA transposable elements (5.42%).

**Comparative genome analysis.** Divergence between *Corymbia* and other woody angiosperm genomes<sup>42</sup> (*Eucalyptus grandis*<sup>13</sup>,

*Salix purpurea* [willow]<sup>43</sup>, *Populus trichocarpa* [poplar]<sup>8</sup>, *Vitis vinifera* [grape]<sup>44</sup>) was investigated using the synonymous mutation rate (Ks) among single-copy orthologous genes. All-on-all Diamond alignment hits were filtered based on syntenic blocks (as a complimentary measure to orthology), finding 9,410 syntenic orthogroups among the five genomes and 3,496 eucalypt-specific (shared among *E. grandis* and CCV) gene families (Fig. 2a; Supplementary Data File 2). Among the syntenic



**Fig. 2 Orthologous gene groups among woody plant genomes. a** Shared orthogroups among *Corymbia citriodora* ssp. *variegata*, *Eucalyptus grandis*, *Populus trichocarpa*, *Salix purpurea*, and *Vitis vinifera*. **b** Single-copy ortholog synonymous substitution rate (Ks) comparisons to date the *Corymbia/Eucalyptus* divergence. Blue line- Synonymous substitution rates among CCV and EG orthologs. Yellow line- Synonymous substitution rates among EG and VV orthologs. Red line- Synonymous substitution rates among CCV and VV orthologs. Purple line- Synonymous substitution rates among CCV and PT orthologs. Green line- Synonymous substitution rates among EG and PT orthologs. PT *Populus trichocarpa*, CCV *Corymbia citriodora* subsp. *variegata*, EG *Eucalyptus grandis*, VV *Vitis vinifera*.

orthogroups, 2,942 contained single-copy orthologs and were used to estimate the synonymous substitution rate (Ks) within *Corymbia*. Based on the median Ks peak among CCV and *E. grandis* (0.1585) (Fig. 2b; Supplementary Data File 3) and estimated divergence times based on the fossil record (57.2, 58.5, 64.6 MYA)<sup>16</sup>, the synonymous mutation rate (site/year) among orthologs was estimated between  $1.385 \times 10^{-9}$  and  $1.227 \times 10^{-9}$ .

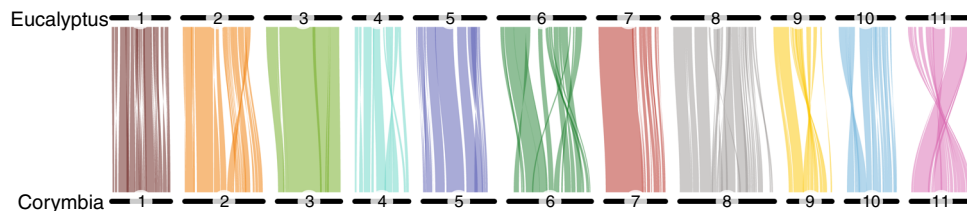
While these mutation rates are consistent with those observed within Salicaceae<sup>45</sup>, they are substantially slower (3.8–4.0 fold) than SNP-based population estimates from *E. grandis* ( $4.93 \times 10^{-9}$ )<sup>46</sup>. To investigate this result further, we calculated estimates of the population mutation rate parameter for *Corymbia* from re-sequencing data of the parental genotype (CCV2-018), as well as four unrelated CCV genotypes (CCV2-019, CCV2-025, CCV2-045, CCV2-046). The measures of population mutation rate ( $4N_e\mu$ ) obtained from the genotypes following maximum likelihood estimators based on the shotgun sequence data ranged between  $7.12 \times 10^{-3}$  and  $8.32 \times 10^{-3}$  (average  $7.86 \times 10^{-3}$ ). For the purpose

of the comparison with *E. grandis*, we assumed an ancestral population size of 112,421, which is consistent with the past demographic history of that species<sup>46</sup>. On this basis, the mutation rate per site per generation in CCV is estimated between  $1.59 \times 10^{-8}$  and  $1.85 \times 10^{-8}$ , which is consistent with the Ks mutation result if a generation time of about 15 years is assumed for CCV ( $1.85 \times 10^{-8} / 15 \text{ y} = 1.23 \times 10^{-9}$  site/year). However, it is worth noting that the true generation time of CCV is unknown, as CCV undergoes mass flowering<sup>47</sup>, and (unlike *E. grandis*) forms lignotubers through which they can regenerate<sup>48</sup>.

Assuming the above assumptions are plausible, the nucleotide diversity seems to be lower in CCV than in *E. grandis*, while the overall chromosome recombination rates appear to be consistent between both species (CCV = 2.85 cM/Mb; *E. grandis* = 2.98 cM/Mb)<sup>49</sup> (Supplementary Table 2). Other than recombination, one factor that negatively correlates with diversity in genomes is the density of targets for purifying selection, which has been often approximated by the density of coding sequence<sup>50</sup>. In a scenario in which long-term effective population size and recombination remain equal between the species, the higher density of coding sequences in CCV due to its smaller genome could be a contributing factor for the reduction of its diversity. It might be important to carry out a further detailed investigation if the apparent reduction of diversity in CCV is predominantly an effect of the increase in density of targets for selection due to changes in chromosome size. An alternative explanation may relate to differences in the ancestral population sizes or its patterns of variation in comparison to *E. grandis* demographics.

Despite differences in their genome size and ~60 MYA divergence, CCV and *E. grandis* have few large-scale intra-chromosomal rearrangements and have retained large syntenic blocks (Fig. 3; Supplementary Data File 4). Both species have 11 chromosomes ( $2n = 22$ ), with chromosomes 1, 3, 5, and 7 being largely 1:1 syntenic. Chromosomes 4, 8, 9, 10, and 11 contain major inversions, and chromosomes 2 and 6 harbor inverted intra-chromosomal translocations. These major chromosomal rearrangements have been previously described<sup>37</sup>, but chromosome 11 in the genome assembly was inverted relative to the genetic map to maximize synteny with *E. grandis*. Within Myrtaceae, a chromosome number of 11 ( $2n = 22$ ) has been widely conserved across most major clades, with some exceptions of polyploidy ( $2n = 33, 44, 66$  [3x, 4x, 6x]) occurring within *Leptospermum*, *Psidium*, and *Eugenia*<sup>15</sup>. Globally, 71% ( $n = 25,357$ ) of CCV genes were retained in large intra-chromosomal syntenic blocks with *E. grandis* and 86% average identity between protein sequences (top hit among CCV and *E. grandis* primary proteins). An average of 17 syntenic blocks were detected on each chromosome. Chromosome 3 was the most syntenic with 98% of genes captured in six blocks and the largest block containing 89% of all chromosome 3 genes (Table 2; Fig. 3; Supplementary Data File 4). Chromosome 6, despite multiple inverted translocations, maintained 75.3% of genes in 19 syntenic blocks. The overall correlation ( $r$ ) between CCV and *E. grandis* chromosome sizes is 0.88 ( $n = 11$ ;  $p = 0.003$ ).

Despite high synteny, Myrtaceae genomes vary considerably in size<sup>13,15</sup>. Across the entire Myrtaceae family, the 1n genome size range of diploids (based on flow cytometry) is approximately fivefold, from 234 Mb (*Myrciaria glazioviana*; pg/1C = 0.239) to 710 Mb (*Eucalyptus saligna*; largest eucalypt; pg/1C = 0.735) and 1.1 Gb (*Melaleuca leucadendra*; pg/1C = 1.100)<sup>15,51,52</sup>. The difference in genome size between CCV (408 Mb) and *E. grandis* (641 Mb) was 233 Mb, of which ~139 Mb could be attributed to repetitive content (35.8% vs 43.9%, respectively when compared using the same pipeline). While this observation is consistent with other plant genomes where repeat content contributes to genome size differences<sup>53</sup>, comparisons between *E. grandis* and *E. globulus*



**Fig. 3** Intra-chromosomal syntenic blocks among *Eucalyptus grandis* and *Corymbia citriodora* subsp. *variegata*. Numbers represent individual chromosomes. Minimum number of genes per block is 25.

**Table 2** Comparison of chromosome gene content and synteny between *Corymbia citriodora* subspecies *variegata* and *Eucalyptus grandis*.

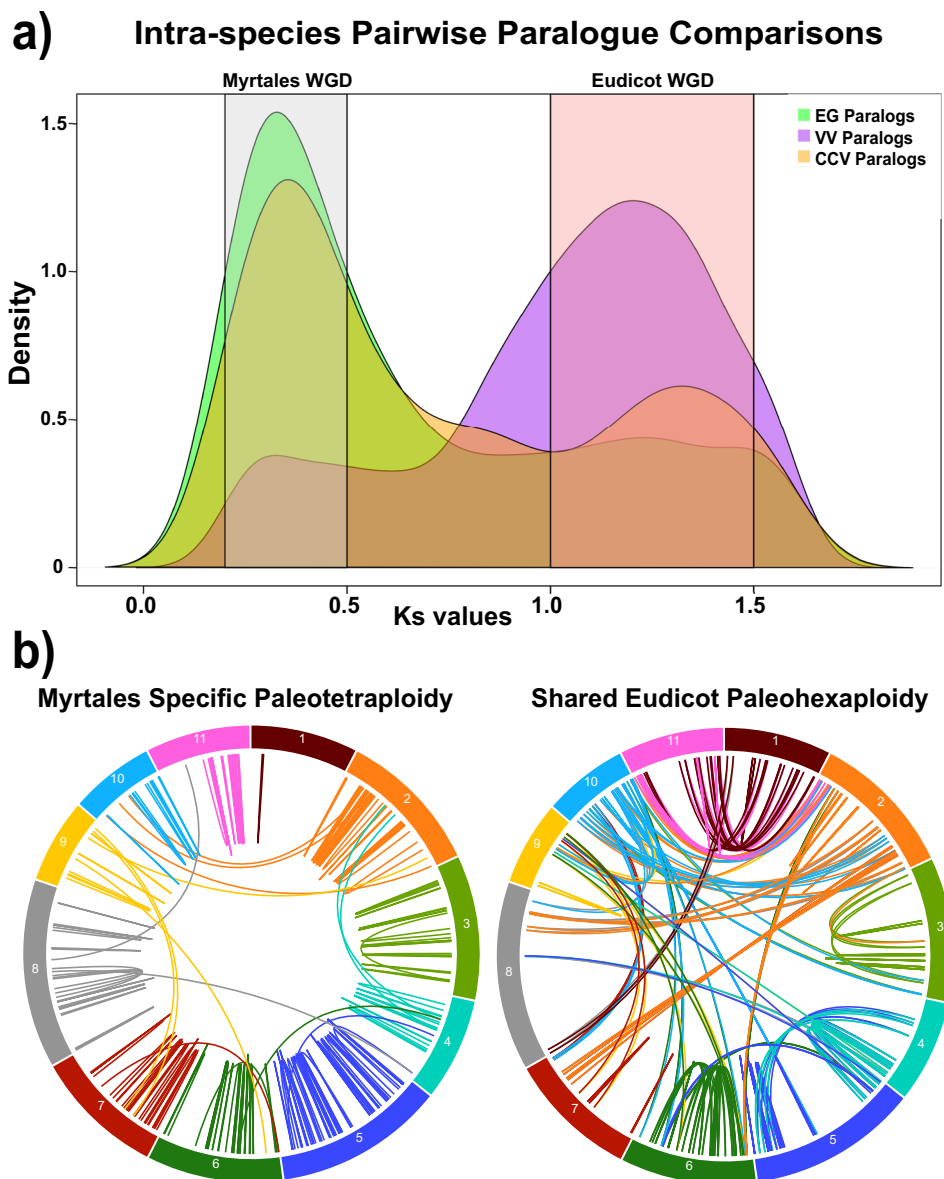
| Chromosome | Total # of genes on Chr | Genes in syntenic blocks | Percent genes in syntenic blocks | Total number of syntenic blocks | Number of genes in largest syntenic block | Percent of genes captured by largest block |
|------------|-------------------------|--------------------------|----------------------------------|---------------------------------|---|--|
| 1          | 2,436                   | 2,235                    | 91.7                             | 25                              | 426                                       | 17.5                                       |
| 2          | 3,246                   | 2,810                    | 86.6                             | 23                              | 677                                       | 20.9                                       |
| 3          | 2,618                   | 2,574                    | 98.3                             | 6                               | 2,297                                     | 87.7                                       |
| 4          | 2,234                   | 1,684                    | 75.4                             | 11                              | 403                                       | 18.0                                       |
| 5          | 2,958                   | 2,373                    | 80.2                             | 13                              | 746                                       | 25.2                                       |
| 6          | 3,427                   | 2,582                    | 75.3                             | 19                              | 906                                       | 26.4                                       |
| 7          | 2,478                   | 2,206                    | 89.0                             | 15                              | 1,523                                     | 61.5                                       |
| 8          | 3,923                   | 3,298                    | 84.1                             | 29                              | 600                                       | 15.3                                       |
| 9          | 1,792                   | 1,463                    | 81.6                             | 13                              | 422                                       | 23.5                                       |
| 10         | 2,099                   | 1,874                    | 89.3                             | 19                              | 335                                       | 16.0                                       |
| 11         | 2,497                   | 2,258                    | 90.4                             | 17                              | 710                                       | 28.4                                       |

(section *Symphyomyrtus*; ~36 MYA divergence)<sup>13</sup> showed a genome size difference of 111 Mb (i.e., 641 Mb and 530 Mb respectively), the majority of which was attributable to non-repetitive *E. grandis* specific sequences distributed throughout the genome (88.7 Mb). *Eucalyptus grandis* compared to the more distantly related, recently sequenced draft assembly of *E. pauciflora* (subgenus *Eucalyptus*)<sup>53</sup> found that while the genome of *E. pauciflora* (594 Mb) was 16% smaller than *E. grandis*, its repeat content was greater (44.77% versus 41.22%).

**Whole-genome duplications.** Following the eudicot gamma WGD event (~140 MYA-paleohexaploidy)<sup>54</sup>, eucalypts underwent a lineage-specific paleotetraploidy event (~109 MYA), which coincides with the Myrtales divergence from other Rosids<sup>13,55</sup> and is considerably older than other WGD events that have occurred in poplar, *Arabidopsis*, and soybean<sup>45,56,57</sup>. Evidence of this event is present in *Corymbia*, based on Ks values of syntenic paralogous sequences. Ks values among intra-specific paralogs of CCV, *E. grandis* and *V. vinifera* revealed a clear signal of the shared eudicot paleohexaploidy event<sup>54</sup> (Ks ~1.2) and the Myrtales specific paleotetraploidy event (Ks ~0.4) (Fig. 4a; Supplementary Data File 5). Illustration of these paralogous pairs shows a similar intra-chromosomal dispersion pattern within CCV (Fig. 4b; Supplementary Data File 5) as *E. grandis* (Myburg et al.<sup>13</sup>; Fig. 2b). KEGG pathway enrichments among *Corymbia* paralogs within the Myrtales WGD peak ( $n = 528$ ) showed significant enrichment for the biosynthesis of unsaturated fatty acids ( $P = 0.02$ ), nitrogen metabolism ( $P = 0.03$ ), plant-pathogen interaction ( $P = 0.03$ ), and sesquiterpenoid and triterpenoid biosynthesis ( $P = 0.02$ ) (Supplementary Data File 5; Supplementary Data File 6). The importance and prevalence of terpene biosynthesis is well documented in the eucalypts as a mechanism for mediation of abiotic/biotic stresses<sup>33</sup>. Unsaturated fatty acids are a key component in the waxy leaf cuticle of *Eucalyptus*, which not only protect against temperature stress through osmo-

regulation of water, but are also implicated in resistance to fungal pathogens such as Myrtle rust (*Austropuccinia psidii*)<sup>58</sup>, which threatens 1,285 species of Myrtaceae<sup>59</sup>.

*Eucalyptus grandis* chromosome 3 has previously been identified as the most conserved, maintaining synteny with *P. trichocarpa* chromosome XVIII despite being more than 100 million years diverged<sup>13</sup>. CCV chromosome 3 also maintains this pattern, while other conserved syntenic blocks among CCV and *E. grandis* were dispersed among multiple *P. trichocarpa* chromosomes (e.g., CCV chromosome 5; Supplementary Data File 6). Myburg et al.<sup>13</sup> postulated that chromosome 3 (and chromosome XVIII in *P. trichocarpa*) each represent a single copy of the ancestral eudicot chromosome A4. After the Myrtales-specific WGD, chromosome 3 homologs fused, as evidenced by all retention of paralogs from the WGD residing on chromosome 3; a pattern shared by CCV (Fig. 4b; Supplementary Data File 5). Myburg et al.<sup>13</sup> proposed that the maintenance of synteny between *E. grandis* and *P. trichocarpa* was possibly due to either: (1) favored preservation of perennial woody-habit genes, or (2) introduced internal centromeres and telomeres (3 Mb and 74 Mb within *E. grandis*) that have repressed subsequent recombination and gene expression<sup>13</sup>. To investigate this possibility, differences in gene expression, recombination, and synteny on CCV chromosome 3 were considered. We observed no significant differences (ANOVA;  $P > 0.05$ ) in the average gene expression within chromosome 3 and other chromosomes (Supplementary Fig. 7), nor higher repetitive content. Looking broadly across all chromosomes, chromosome 3 had lower recombination than average (2.67 cM/Mb; mean: 2.85 cM/Mb), but not the lowest rate (chromosome 5: 2.39 cM/Mb; Supplementary Table 2). The mean recombination rate of CCV is consistent with *E. grandis*, *E. globulus*, and *E. urophylla*<sup>49</sup>, but it is worth noting that within *E. globulus*, recombination rates for chromosomes 3 and 5 were significantly less than other chromosomes, likely due to their large



**Fig. 4 Patterns of paralog retention in *Corymbia* resulting from whole-genome duplications (WGD).** **a** Intra-species retained paralogous sequences from the shared eudicot and Myrtales specific WGD events. Ks- synonymous mutation rate. Green distribution- Ks values among paralogous sequences within *E. grandis*. Yellow distribution- Ks values among paralogous sequences within *C. variegata*. Purple distribution- Ks values among paralogous sequences within *Vitis vinifera*. Gray box- paralogous sequences derived from the Myrtales WGD event. Pink Box- paralogous sequences derived from the Eudicot WGD event. CCV *Corymbia citriodora* subsp. *variegata*, EG *Eucalyptus grandis*, VV *Vitis vinifera*. **b** *Corymbia* chromosomal dispersal pattern of paralogs from the eudicot and Myrtales WGD event. Lines between chromosomes represent intra-specific paralogous sequences that arose from each duplication.

size. CCV chromosomes are more uniform in length than *E. grandis* (CCV average chromosome length = 37.5 Mb; standard deviation = 10 Mb) with similar differences in recombination rate and number of crossover events (Supplementary Tables 2 and 3).

To investigate the possibility that conserved woody-habit genes may require retention of syntenic order, all CCV chromosome 3 orthologs that are maintained within syntenic blocks in both *E. grandis* and *P. trichocarpa* (chromosomes XVIII and VI [which share synteny due to the Salicoid WGD event]) were extracted (1:1:2;  $n = 173$  genes) (Supplementary Fig. 8; Supplementary Data File 7) and compared for enriched gene functions. The top KEGG enrichment pathways for this gene set included phenylalanine and tyrosine metabolism ( $P = 0.001$ ;  $P = 0.003$ ; respectively),

alkaloid biosynthesis (tropane and isoquinoline) ( $P = 0.02$ ), and glycerolipid metabolism ( $P = 0.02$ ), each of which play critical roles in plant primary and specialized metabolism (e.g., monolignol biosynthesis), cell wall formation, defense and stress signaling<sup>60–63</sup>. Regarding plant defense, quantitative trait loci (QTLs) for fungal disease resistance have been repeatedly reported on chromosome 3 in *E. grandis* from several independent studies using different methods of genetic evaluation in breeding populations, clearly demonstrating a major involvement of this chromosome in the pathogen resistance response<sup>64</sup>. While these results favor the syntenic maintenance of critical gene functions on chromosome 3, gene order conservation needs to be examined further across a larger cohort of woody angiosperm genomes.

**Corymbia gene family analysis.** Comparative genomics between CCV and *E. grandis* allowed investigation of gene family expansions that had occurred since eucalypts diverged from other Rosids, as well as expansions specific to CCV itself. Within our constrained syntenic orthogroups, 124 had eucalypt-specific expansions (orthogroups with >5 genes and >70% of genes derived from both CCV and *E. grandis*) containing 1,494 genes (735-CCV; 759-*E. grandis*). Functional KEGG pathway enrichments within these expansions included the molecular processes of phenylpropanoid biosynthesis ( $P = 0.00006$ ), cyanoamino acid metabolism ( $P = 0.006$ ), pentose/glucuronate interconversions ( $P = 0.01$ ), isoflavonoid biosynthesis ( $P = 0.02$ ), monoterpene biosynthesis ( $P = 0.03$ ), glucosinolate biosynthesis ( $P = 0.03$ ), and plant-pathogen interaction ( $P = 0.04$ ) (Supplementary Table 4). These expansions are consistent with general response mechanisms for biotic/abiotic stress, where carbohydrate pathway activation enables rapid signaling and energy for terpene and secondary metabolite biosynthesis<sup>65–68</sup>.

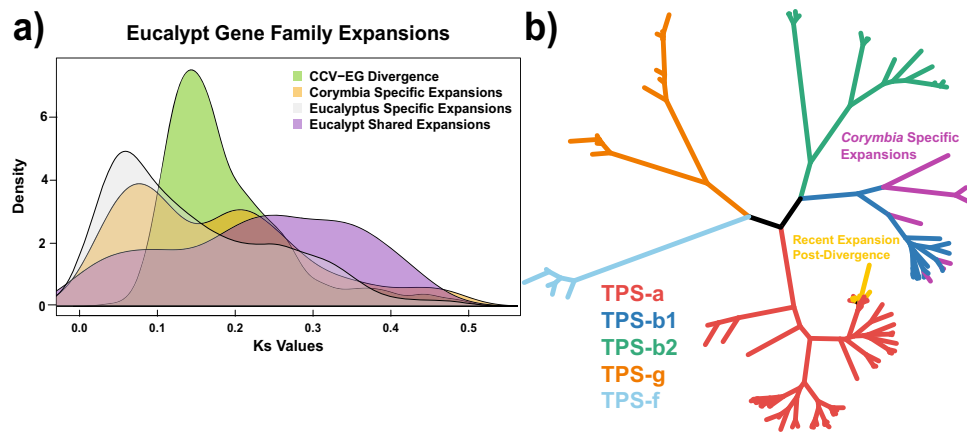
Similarly, we characterized gene families that had undergone species-specific expansions within *E. grandis* and CCV relative to the other woody angiosperm genomes. Within *E. grandis*, investigation of expanded gene families (orthogroups with at least five genes and  $\geq 50\%$  derived from *E. grandis*) found 179 expanded orthogroups containing 1,479 genes. KEGG pathway enrichment analysis within this dataset found the greatest enrichments for galactose metabolism ( $P = 0.00008$ ), phenylpropanoid biosynthesis ( $P = 0.006$ ), flavonoid biosynthesis ( $P = 0.002$ ), pentose and glucuronate interconversions ( $P = 0.002$ ), and alpha-linolenic acid metabolism ( $P = 0.003$ ) (Supplementary Table 5). Within CCV (using the same criteria for expansion noted above), there were 75 expanded gene family orthogroups containing 501 genes. Functional enrichments within these expanded gene families found significant KEGG pathway enrichments relating to plant-pathogen interaction ( $P = 0.003$ ), phenylpropanoid biosynthesis ( $P = 0.003$ ), ether lipid metabolism ( $P = 0.02$ ), sesquiterpenoid and triterpenoid biosynthesis ( $P = 0.02$ ), and cutin, suberine and wax biosynthesis ( $P = 0.03$ ) (Supplementary Table 6). Considering the substantial overlap among significant KEGG enrichments between CCV and *E. grandis* expansions, genes associated with these terms were mapped to their closest ortholog among shared KEGG pathways (Supplementary Figs. 9–12). We found that CCV and *E. grandis* both had both similar (e.g., cutin and sesquiterpenoid biosynthesis) and separate expansions (e.g., phenylpropanoid biosynthesis).

Specifically, when comparing overlaps among species-specific expanded gene families within plant-pathogen interaction KEGG pathways, CCV displayed signatures of expansion in gene families that were absent in *E. grandis* (Supplementary Fig. 12). Gene families with shared expansions in both CCV and *E. grandis* were related to disease resistance protein 2 (*rps2*) and mitogen-activated protein kinase kinase 1 (*mekk1*), which are part of separate stress response pathways within the cytoplasm. *Mekk1* is a mitogen-activated protein kinase (MAPK) signal cascading gene within the pathogen-associated molecular pattern triggered immunity (PTI) pathway that is tightly associated with abiotic stress response such as temperature, drought, salinity, as well as wounding<sup>69</sup>. *Rps2* is a resistance NB-LRR gene that upon recognition with bacterial effector proteins, generates an effector-triggered immune (ETI) response and can elicit a localized hypersensitive response (HR) where cells undergo programmed death to prevent pathogen spread<sup>70</sup>. Investigation of CCV-specific gene family expansion revealed separate expansions in the same pathogen interaction pathways: *rpm1*, *fls2*, and *bak1/bkk1*. Similar to *rps2*, *rpm1* is an R gene and elicits an ETI/HR response upon detection of bacterial effector proteins<sup>71,72</sup>. *Fls2* and *bak1* however, are part of the PTI immune response pathway, both being plasma-membrane bound receptor kinases

that form a signaling complex, that upon activation, cascade signals to cytoplasmic kinases as part of PTI responses. *Fls2* recognizes a specific peptide sequence of bacterial flagellin, and while *bak1* is part of the same plasma membrane complex, it can initiate signaling independently of *fls2*, recognizing the EF-Tu bacterial receptors, lipopolysaccharides, peptidoglycans and whose function is critical as part of the plant innate immune response<sup>73,74</sup>.

Based on the prevalence of tandem gene arrays in *E. grandis*, we investigated tandem gene duplication in CCV as a mechanism for gene family expansion whose functions are enriched for climate niche adaptation for hot season rainfall and semi-arid environments<sup>75</sup>. CCV has a large number of tandemly duplicated genes, with a similar number of arrays in extended syntenic blocks as *E. grandis* (8,366 vs 8,679; 23% vs 24% of all genes, respectively). This number is lower than previously reported for *E. grandis* ( $n = 12,570$ )<sup>13</sup>, as we used MCScanX<sup>76</sup> as our standardized, more conservative, methodology for identifying tandem repeats within each genome, as these can often be difficult to define. As hypothesized, there was an 81% overlap ( $n = 408$ ) between CCV-specific gene family expansions and tandem duplicates which were significantly enriched for the same KEGG pathways, thus tandem gene duplication appears to be a major mechanism of gene family expansion in eucalypts.

To estimate the relative ages of these gene family expansions, comparisons within eucalypt-specific, CCV-specific and *E. grandis* specific expansions were investigated. First, CCV and *E. grandis* genes derived from 1:1:1 orthogroups among *Corymbia: Eucalyptus: Vitis* (outgroup) were used to visualize the Ks peak when *Corymbia-Eucalyptus* diverged. Then, eucalypt-specific (both CCV and *E. grandis*; *V. vitis* outgroup) expansions, CCV-specific expansions, and *E. grandis*-specific expansions were compared to find whether those expansions were relatively younger or older than the *Corymbia-Eucalyptus* split. The divergence peak among *Corymbia* and *Eucalyptus* genes occurs at Ks  $\sim 0.15$  (Fig. 5a; Supplementary Data File 8). The eucalypt gene family expansions pre-date this divergence, with its main peak spread between Ks  $\sim 0.2$ – $0.4$  (total = 550; CCV = 265; *E. grandis* = 285). CCV-specific expansions displayed a bimodal peak, with some expansions occurring prior to divergence (Ks  $\sim 0.21$ ;  $n = 101$ ) and others undergoing a relatively recent expansion (Ks  $\sim 0.08$ ;  $n = 132$ ). These recent expansions suggest a dynamic mechanism for increasing gene numbers where function is enriched for sesquiterpenoid and triterpenoid biosynthesis ( $P = 0.02$ ) as well as cutin, suberine, and wax biosynthesis ( $P = 0.03$ ) (Supplementary Figs. 10–11; Supplementary Table 7). Plant cuticular wax compounds perform functions essential for the survival of terrestrial plants, including limiting non-stomatal water loss and gas exchange, protecting from ultraviolet radiation<sup>77</sup>, and forming physical barriers to herbivores and pathogens<sup>78</sup>. Eucalypts in particular require strong defense mechanisms to protect leaves during development, as young trees present large amounts of juvenile foliage making them a target for insect and fungal pests<sup>79</sup>. *Eucalyptus grandis* and CCV differ somewhat in wax morphology, with *E. grandis* exhibiting sparser wax coverage and irregular structure compared to *Corymbia*<sup>80</sup>. Concentrations of cuticular wax compounds in eucalypts have been linked to water loss response, for instance, a 10-fold increase in concentrations of n-alkane (wax compounds) was observed along an aridity gradient in both *Eucalyptus* and *Corymbia*<sup>81</sup>. Leaf wax and water repellency have also been linked to frost tolerance along an altitudinal cline in *Eucalyptus*<sup>82</sup>. The expansion in CCV genes linked to wax biosynthesis and the subsequent increase in concentration is likely selected for by the highly seasonal and variable rainfall environments occupied by CCV. However, the expansion of this gene family may be species-



**Fig. 5 Relative timing of eucalypt gene family expansions.** **a** Synonymous substitution rates (Ks) among orthologs derived from shared eucalypt expansion and those that are *E. grandis* and *C. c. variegata* specific. Green distribution- *Corymbia-Eucalyptus* divergence peak among orthologs. Gray distribution- *E. grandis* expanded orthogroups. Orange distribution- *C. c. variegata* expanded orthogroups. Purple distribution- *E. grandis* and *C. c. variegata* shared expanded orthogroups. **b** *C. c. variegata* terpene gene families. Red branches- Terpene synthase family-a genes. Blue branches- Terpene synthase family-b1 genes. Green branches- Terpene synthase family-b2 genes. Orange branches- Terpene synthase family-g genes. Light-blue branches- Terpene synthase family-f genes. Purple branches- Terpene synthases that are generally expanded in *C. c. variegata*. Yellow branches- Terpene synthases that expanded after *C. c. variegata* diverged from *E. grandis*.

rather than genus-specific given the variation wax concentrations exhibited across species of *Eucalyptus* and *Corymbia*<sup>81</sup>.

Terpenes, in addition to regulating growth and developmental processes<sup>83</sup>, contribute chemical barriers to herbivory<sup>84,85</sup>, pollinator attraction<sup>86</sup>, and thermotolerance<sup>87</sup>. The largest terpene synthase subfamily in CCV is TPS-a (responsible for sesquiterpene biosynthesis), which has undergone a recent expansion since diverging from *Eucalyptus* (Fig. 5b; Supplementary Data File 8). The TPS-b1 subfamily (responsible for monoterpene synthesis), was also generally expanded in *Corymbia*, based on the CCV-expanded orthogroups. The majority of terpenes synthesized by the two subfamilies are volatile and responsible for attracting pollinators and for plant defense via tritrophic interactions<sup>88</sup>. However, some sesquiterpenes are non-volatile phytoalexins that directly protect against fungal and bacterial pathogens<sup>89</sup>. Similarly, *E. grandis* specific expansions occurred post-divergence (Ks ~0.06;  $n = 296$ ), with gene function enrichments related to sesquiterpenoid/triterpenoid biosynthesis (TPS-a) ( $P = 0.02$ ) and glycan degradation ( $P = 0.03$ ) (Supplementary Table 8). These CCV and *E. grandis* lineage-specific gene family expansions are likely due to selective pressures in environmental niches occupied by these two genera and appear to provide evidence of concerted evolution in eucalypts, but requires investigation of more species to lend support. In particular, genes involved in terpene biosynthesis have undergone separate but parallel expansions via tandem gene duplication<sup>32</sup>.

## Discussion

The generation of a high-quality *de novo* sequenced genome for *Corymbia citriodora* subsp. *variegata* has provided the opportunity to understand how evolutionary history has contributed to genome evolution within the Myrtaceae, an important and diverse group of angiosperms that have radiated across the Southern Hemisphere. After the gamma paleohexaploidy WGD (140 MYA) and divergence from other Rosids, Myrtales (the taxonomic order to which Myrtaceae belongs) underwent another lineage-specific WGD tetraploidy event (109 MYA). Paralogous sequence retention in the CCV genome underpins the importance of this event, finding functional enrichments for genes involved in pathogen inhibition, heat tolerance, and desiccation resistance, as

well as pollinator attraction via unsaturated fatty acid metabolism and wax and terpene biosynthesis. This ancestral state of the eucalypt progenitor has been maintained even after *Corymbia* and *Eucalyptus* diverged, where large gene families responsible for mono- and sesqui-terpenes synthesis, leaf cuticle wax synthesis, and stress response pathways have expanded further. After the ancient tetraploidy event, most Myrtaceae underwent diploidization, and with few exceptions maintained a haploid chromosome number of eleven whilst exhibiting large differences in genome size (234–1110 Mb)<sup>15</sup>.

Within the eucalypts, *Corymbia* and *Eucalyptus* diverged ~60 MYA but nonetheless maintained synteny among their chromosomes despite also undergoing large chromosomal inversions (chromosomes 2,4,6,8,9,10 and 11) and translocations (chromosomes 2 and 6). Seventy-one percent of genes were captured in large syntenic blocks between the two genomes, with chromosome 3 being the most conserved. This conservation of chromosome 3 also extends to Northern Hemisphere *P. trichocarpa* (chromosomes XVIII and VI) and may be a result of the required synteny of conserved genes with related critical functions, which warrants further investigation. Throughout their evolutionary histories, both CCV and *Eucalyptus* have undergone gene family expansions whose function mainly relates to biotic and abiotic stress response. These gene family expansions have occurred in both separate (phenylpropanoid biosynthesis) and shared enzymatic pathways (cutin and terpenoid biosynthesis), while CCV also shows unique signatures of expanded key signaling components within the PTI pathway. While a number of these gene family expansions are shared, there is evidence of concerted and parallel evolution within CCV and *E. grandis* where gene families related to terpene biosynthesis (TPS-a) have expanded via tandem duplication in both species since they diverged.

The sequencing and description of the CCV genome will help inform future conservation efforts, molecular breeding, and global deployment of this taxa. In addition to Australia, *Corymbia* has been deployed to plantations in China, India, Sri Lanka, South Africa, Congo, and Kenya<sup>90</sup>. More recently, it has been successfully established in Brazilian plantations for hardwood, charcoal, and essential oil production and can outperform *Eucalyptus* in areas negatively impacted by climate-driven abiotic/biotic stresses<sup>91</sup>. *Eucalyptus* breeding programs and resources are



well established, and knowledge regarding the stability of the *Corymbia* genome in terms of synteny, recombination, and tandem duplication will accelerate molecular breeding for both taxa and allow genomic resources already established for *Eucalyptus* (e.g. 60 K SNP chip, SSR markers)<sup>92</sup> to be more easily transferable to other eucalypts. The release of additional Myrtaceae reference genomes will hopefully enable more extensive insights into evolutionary history based on comparative genomics across this important and diverse lineage of plants.

## Methods

**Illumina DNA library construction and sequencing.** Genomic DNA was extracted<sup>93</sup> from leaf tissue of *Corymbia citriodora* subsp. *variegata* genotype CCV2-018. Approximately 100 nanograms of DNA was sheared to 500 and 800 bp using the Covaris LE220 (Covaris), then size selected with SPRI beads (Beckman Coulter). DNA fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). The prepared libraries (insert sizes 400 and 800 bp) were quantified using the next-generation sequencing KAPA Biosystem library qPCR kit, run on a Roche LightCycler 480 qPCR instrument. The two PCR-free Illumina libraries were multiplexed into pools then prepared for sequencing with a TruSeq paired-end cluster kit (v3) and Illumina cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2500 platform using a TruSeq SBS sequencing kit (v3) following a 2×250 indexed run recipe.

**PacBio library preparation and sequencing.** Genomic DNA was sheared using the Covaris g-tube 20 Kb centrifugation protocol and purified using a 0.45X Ampure PB purification step. Single-stranded DNA fragments were removed using an exonuclease treatment, followed by DNA damage repair, end repair, and SMRTbell adapter ligation. In addition, a second exonuclease step removed failed ligation products (SMRTbell template prep kit 1.0). Ligated fragments were then size selected for those >7 Kb in length (as short fragments are preferentially sequenced) then sequenced on the RSII instrument using P6-C4 chemistry and 4-h movie lengths. Reads were then processed using SMRTPortal (version 2.3.0) RS subreads protocol with default filtering settings (min subread length: 50; min polymerase read quality: 75; min polymerase read length: 50).

**RNA collection and sequencing.** RNA was isolated from five tissue types: expanding and fully expanded leaves, flower buds and initials, and the outer chlorophyllous layer of bark cortex (Supplementary Fig. 3)<sup>94</sup>. Tissues were obtained from the CCV2-054 genotype (genetic map parent)<sup>37</sup> and immediately preserved in a cryogenic shipping unit in the field for transport and storage prior to extraction. Total RNA was prepared 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. A total of 75 Gb of RNA-seq was generated across all five libraries, 25 Gb of 100 bp single-end reads, and 50 Gb of 100 bp of paired end reads.

**Genome assembly.** Contig assembly and initial scaffolding were conducted using Illumina paired-end reads. A total of 462,039,870 reads (representing ~163× genome coverage) were assembled Arachne (v.20071016)<sup>35</sup>, modified to handle larger datasets (data IO/sort functions/increased number of reads/alignments in memory). Arachne assembly parameters: macliq=800, remove\_duplicate\_r\_reads=True, correct1\_passes=1, BINGE\_AND\_PURGE\_2HAP=True, max\_bad\_look=1000. Contig assembly and initial scaffolding steps produced 37,263 contigs into 32,740 scaffolds, totaling 563.0 Mb. Scaffold N50 length of the assembly was 132.6 Kb, with 1,430 scaffolds larger than 100 Kb. The resulting scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. In addition, scaffolds were removed if they were (a) repetitive, defined as scaffolds less than 50 Kb consisting of >95% 24 mers that occur four or more times in scaffolds >50 Kb, (b) contained only unanchored RNA, (c) <1 Kb in length, or (d) alternative haplotypes, defined as scaffolds <10 kb that align to scaffolds >10 Kb scaffolds with at least 95% identity and 95% coverage.

**PacBio patching.** Gaps in the assembly were patched using ~25× sequencing coverage of PACBIO filtered subreads. Gaps were patched by first breaking scaffolds into contigs. Contigs <1 Kb were excluded from the gap patching process. Subsequently, 1 Kb of the sequence was trimmed off the contig ends and the trimmed portion was broken into 100 mers. The 100 mers were aligned to the PACBIO reads using the short read aligner bwa<sup>95</sup>, and individual PACBIO reads were mapped to scaffolds indicated by the 100mer alignments. QUIVER was used to assemble gap crossing reads for gaps with more than 5 filtered subreads crossing them. The resultant assembled sequence was used to patch the gap. A total of 3,149

gaps were patched, with a total loss of 55,511 bases from the raw assembly due to the presence of negative gaps in the assembly. Mis-assemblies were assessed by identifying gaps where 5 or more PACBIO reads have >1 Kb regions of the read aligning to two different scaffolds. A total of 166 mis-joins were identified and the breaks made, with the associated join being made using the reads that indicated a break. A total of 485 additional joins were made using the PACBIO reads. Additional scaffolding of the genome was performed using SSPACE-Standard (Version 2.0) with Nextera long mate pair prepared libraries (Insert size 4 and 8 Kb). SSPACE scaffolding was performed using default parameters and no extension ( $x = 0$ ).

**Anchoring scaffolds to linkage maps.** The retained assembly from the gap patched assembly and SSPACE scaffolding were anchored into pseudomolecules using the ALLMAPs pipeline<sup>39</sup>. Three individual genetic maps were generated representing each parent of the two pedigrees (male map- CCV2-054, female maps- CT2-050 and CT2-018). The use of a ‘marker binning’ procedure, and stringent criteria for the inclusion of markers, resulted in robust marker orders in the linkage maps evidenced by high rank-order correlations among shared markers<sup>37</sup>. The sequence of each marker was used to anchor (and orient, if a second marker was available) contigs with matching sequence onto specific linkage groups, with a greater weighting given to the order of markers from CCV.

ALLMAPs incorporates a methodology for computing a scaffold order that maximizes collinearity across a collection of maps and generates outputs of ordered and orientated scaffolds. The pipeline was run using the default settings, except that filtering was applied so that linkage groups with <20 markers were removed from the analysis, joins between scaffolds were padded with 100 N, and a weighting of 1 (i.e., highest confidence of marker order) was applied to the CCV2-054 map, and a weighting of 2 was applied to both of the *C. torelliana* maps. A lower weighting applied to the *C. torelliana* was used to allow for a higher likelihood of the possibility of marker reordering in the nonfocal species. Similarity searches aimed at matching DArT-seq markers sequence tags from the genetic maps with scaffolds were determined by using the blastn program from BLAST<sup>96</sup> where a threshold  $e$ -value of  $1 \times 10^{-10}$  was used as a cutoff and only the best match was taken. Chromosomes and subsequent tracks in Fig. 1 were created using Circa (<http://omgenomics.com/circa>).

**Single-copy gene analysis.** *Eucalyptus grandis* protein sequences were re-aligned to the *E. grandis* genome sequence using BLAT<sup>97</sup> (-noHead -extendThroughN t=dnax q=prot) to find single-copy genes sequences. 20,256 single-copy genes were identified within *E. grandis* (no tandem duplications, no gene splice variants, 90% gene coverage, 85% gene identity, >300 bp). Aligned to CCV (using the same BLAT parameters), 16,245 proteins were found similar (>75% ID; 90% coverage), of which 14,911 were also present in single copy (92%) in the CCV assembly and of those 90% were present on pseudomolecules.

**Protein-coding gene classification and annotation.** Adequate RNA mapping for transcript assembly and protein prediction was verified (88% average mapping across tissues), as well as global sequence identity between genotypes (97% global; 98% CDS). Transcript assemblies were made from ~260 M pairs of 2 × 100 stranded paired-end Illumina mRNA-seq reads using PERTRAN as used in other plant genome annotations<sup>98</sup>. 99,336 transcript assemblies were constructed using PASA<sup>99</sup> from mRNA-seq transcript assemblies above. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from arabi (*Arabidopsis thaliana*), soybean, poplar, tomato, Kitaake rice, brachy, aquilegia, eucalyptus, grape, and Swiss-Prot proteomes to the repeat-soft-masked genome assembly using RepeatMasker<sup>100</sup> with up to 2 Kb extension on both ends unless extending into another locus on the same strand. Repeat library consists of *de novo* repeats by RepeatModeler<sup>101</sup> on the CCV genome. Gene models were predicted by homology-based predictors, FGENESH+<sup>102</sup>, FGENESH\_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and GenomeScan<sup>103</sup>, PASA assembly ORFs (in-house homology constrained ORF finder) and from AUGUSTUS via BRAKER1<sup>104</sup>. The best-scored predictions for each locus were selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. The improvement includes adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to above-mentioned proteomes to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is the highest percentage of protein aligned to the best of homologs. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with repeats. The transcripts were selected if its Cscore was larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or it had EST coverage, but its CDS overlapped with repeats less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cscore was at least 0.9 and homology coverage was at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose proteins were more than 30% in Pfam. TE domains were removed along with weak gene models.

Incomplete gene models, models where there was low homology support without fully transcriptome support, short single exons (<300 BP CDS) without a protein domain, or a lack of good expression gene model, were manually filtered out.

**Syntenic blocks.** All pairwise BLAST hits were calculated with Diamond<sup>105</sup> either separately, or within the Orthofinder<sup>106</sup> program. Hits were then culled to the top two hits within each haplotype (so, if a diploid is mapped to a diploid, four hits would be retained for each gene; if it were mapped to a tetraploid, eight hits would be retained). All hits with a bit score <50 were dropped.

**Initial orthogroup inference:** A separate run of Orthofinder was made for each pair of genomes using the culled BLAST results. Only those hits within 'orthogroups' were retained.

**Initial block construction:** Blocks were formed using MCSanX from the culled and orthogroup-constrained BLAST results, allowing 5× as many gaps in the alignment as 50% of the minimum block size (MBS, default = 10). Block were then pruned with DBSCAN to blocks with MBS hits within a fixed gene-rank radius 5× MBS. All orthogroups are then 'completed', where igraph expanded the orthogroups to include all possible combinations among genomes. These blast hits were then pruned with dbscan with identical parameters as above. **Block cleaning and extension:** To fill potential gaps in blocks left by the stochastic nature of varying orthogroup connectivity, we pull all blast results that passed the score threshold (agnostic to orthogroup identity) that were within a 100-gene radius of any syntenic block and re-form blocks with MCSanX with the same parameters as above.

**Syntenic orthology inference:** All BLAST results were culled to those within a 50-gene rank radius of any syntenic block for all genomes. Orthofinder was run on this entire set, and BLAST hits were parsed into orthologs, paralogs, or unclustered homologs. By default, hits that were not in an orthogroup (neither orthologs or paralogs) with a score <50 or <50% of the best bit score for that gene-by-unique genome combination were dropped from this dataset.

**Gene family analysis and mutation rate.** Single-copy gene orthologs and gene family expansions were characterized using Orthofinder (v2.2.7)<sup>106</sup>. All on all protein sequences from *C. citriodora*, *E. grandis*, *S. purpurea*, *P. trichocarpa*, and *V. vinifera* were performed using Diamond<sup>105</sup>, then clustered into orthogroups using Orthofinder. Single-copy orthologs were determined as clusters containing one protein sequence from each of the five species. Protein alignments among species were performed using MAFFT (v7.464)<sup>107</sup> and the coding sequence was extracted using a pal2nal<sup>108</sup> perl script. Pairwise synonymous mutation rates were calculated from coding sequences using PAML<sup>109</sup> codeml. Ks mutation rate/site/year (R) was calculated as:  $R = Ks / (2 * \text{divergence age})$ . Estimates of population mutation rate ( $4N_e\mu$ ) was obtained from the CCV parental library and four unrelated genotypes (CCV2-019; -025; -045; -046) following maximum likelihood estimators based on the alignment of shotgun sequence data<sup>110</sup> to the CCV genome sequence, using bwa mem (version 0.7.17-r1188)<sup>95</sup>. Upper and lower confidence bounds of estimates per chromosome were extracted per diploid genotype. At each position in the chromosome we count the four different nucleotides,  $n = (n_A, n_C, n_G, n_T)$ , and such a quartet of counts was called a profile, while the sum of counts,  $n = n_A + n_C + n_G + n_T$ , was the coverage of the profile's position. Heterozygous profile position were taken as those passing SNP calling using the GATK (version 3.8) program<sup>111</sup> following the best practices pipeline<sup>112</sup>, while homozygous profile positions were taken using the minimum depth of 4. The neutral mutation rate ( $\mu$ ) per base pair per generation (year) was re-estimated based on these estimates of population mutation rate from shotgun data, assuming a generation time of 15 years in CCV and an ancestral population size equals to 112,421, which was suggested previously to be consistent with the demographic past of the related species of *E. grandis*<sup>46</sup>.

**Paralogs and whole-genome duplication.** Paralogs among *C. citriodora*, *E. grandis*, and *V. vinifera* (outgroup) were extracted from orthogroup and ortholog information generated from Orthofinder. Pairwise Ks substitution rates among paralogs for each species were calculated using codeml. *Corymbia* paralog gene pairs underlying the eucalypt specific peak (0.33–0.45) were extracted and investigated for GO and KEGG pathway enrichment.

**Differential tissue expression.** Tissue-specific RNA libraries were aligned to the indexed *Corymbia* genome using STAR<sup>113</sup> (v2.5.3a; parameters: -outFilterMultimapNmax 7 -outFilterMismatchNmax 4). Gene count tables were exported and analyzed using the edgeR package<sup>114</sup> to obtain reads per kilobase million (rpkm) expression values for each tissue. Normalized gene counts were verified using a heatmap to ensure expression among related tissues was consistent.

**Gene family expansions.** Eucalypt, *Eucalyptus*, and *Corymbia* specific gene family expansions (as defined in "Results") were isolated from Orthofinder orthogroups. Pairwise Ks values among sequences were calculated using codeml within each of the three classes of gene family expansion to investigate which families had accumulated relatively more or less synonymous mutations than *Corymbia-Eucalyptus* single-copy orthologs. Terpene genes were defined using best BLAT<sup>97</sup> hits among *Corymbia* proteins to sequences from Kulheim et al.<sup>33</sup>. The unrooted

terpene gene family tree was generated using FastTree<sup>115</sup> (v2.1.10) and visualized using iTOL<sup>116</sup>, from MAFFT<sup>107</sup> aligned *Corymbia* terpene transcript sequences.

**GO and KEGG pathway enrichment analysis.** Gene ontology (GO) enrichment analysis was carried out using topGO, an R Bioconductor package<sup>117</sup> with Fisher's exact test; only GO terms with a  $P < 0.05$  were considered significant. To identify redundant GO terms, semantic similarity among GO terms were measured using Wang's method implemented in the GOSemSim, an R package<sup>118</sup>. KEGG<sup>119</sup> pathway enrichment analysis was performed based on hypergeometric distribution test and pathways with  $P < 0.05$  were considered enriched.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Additional work to support the findings of this manuscript can be found in the supplementary data section. Raw Illumina (paired end and mate pair) and PacBio reads are available from the National Center for Biotechnology Information Short Read Archive (SRA) under accession: PRJNA234431. Additional CCV resequencing genotypes are available from SRA under accessions: PRJNA333377, PRJNA333376, PRJNA333375, PRJNA333374. Illumina RNASeq data are available at NCBI under BioProject: PRJNA629009. The genome assembly and annotation are freely available at Phytozome ([https://phytozome-next.jgi.doe.gov/info/Ccitriodora\\_v2\\_1](https://phytozome-next.jgi.doe.gov/info/Ccitriodora_v2_1)). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JABURB000000000. The version described in this paper is version JABURB010000000. All relevant data are available upon request from the corresponding author (Adam Healey).

Received: 27 July 2020; Accepted: 5 March 2021;

Published online: 10 May 2021

## References

- Koonin, E. V. Darwinian evolution in the light of genomics. *Nucleic Acids Res.* **37**, 1011–1034 (2009).
- Flood, P. J. & Hancock, A. M. The genomic basis of adaptation in plants. *Curr. Opin. Plant Biol.* **36**, 88–94 (2017).
- Isabel, N., Holliday, J. A. & Aitken, S. N. Forest genomics: advancing climate adaptation, forest health, productivity, and conservation. *Evol. Appl.* **13**, 3–10 (2020).
- Tuskan, G. A. et al. Hardwood tree genomics: unlocking woody plant biology. *Front. Plant Sci.* **8**, 1–9 (2018).
- Daccord, N. et al. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
- Wu, J. et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* **23**, 396–408 (2013).
- Verde, I. et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013).
- Tuskan, G. A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
- Staton, M. et al. Substantial genome synteny preservation among woody angiosperm species: Comparative genomics of Chinese chestnut (*Castanea mollissima*) and plant reference genomes. *BMC Genomics* **16**, 1–13 (2015).
- Plomion, C. et al. Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452 (2018).
- Sollars, E. S. A. et al. Genome sequence and genetic diversity of European ash trees. *Nature* **541**, 212–216 (2017).
- Cruz, F. et al. Genome sequence of the olive tree, *Olea europaea*. *Gigascience* **5**, 29 (2016).
- Myburg, A. A. et al. The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).
- Govaerts, R. et al. *World Checklist of Myrtaceae* (Royal Botanic Gardens, 2008).
- Grattapaglia, D. et al. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genet. Genomes* **8**, 463–508 (2012).
- Thornhill, A. H. et al. A dated molecular perspective of eucalypt taxonomy, evolution and diversification. *Aust. Syst. Bot.* **32**, 29–48 (2019).
- Nicolle, D. Classification of the eucalypts (*Angophora*, *Corymbia* and *Eucalyptus*). *Version 4*, 1–56 (2019).
- Thornhill, A. H., Ho, S. Y. W., K ulheim, C. & Crisp, M. D. Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. *Mol. Phylogenet. Evol.* **93**, 29–43 (2015).

19. González-Orozco, C. E., Thornhill, A. H., Knerr, N., Laffan, S. & Miller, J. T. Biogeographical regions and phytogeography of the *Eucalypts*. *Divers. Distrib.* **20**, 46–58 (2014).
20. Bui, E. N. et al. Climate and geochemistry as drivers of eucalypt diversification in Australia. *Geobiology* **15**, 427–440 (2017).
21. Crisp, M. D., Burrows, G. E., Cook, L. G., Thornhill, A. H. & Bowman, D. M. J. S. Flammable biomes dominated by eucalypts originated at the Cretaceous–Palaeogene boundary. *Nat. Commun.* **2**, 193 (2011).
22. Williams, S. E., Whittaker, J. M., Halpin, J. A. & Müller, R. D. Australian–Antarctic breakup and seafloor spreading: balancing geological and geophysical constraints. *Earth-Sci. Rev.* **188**, 41–58 (2019).
23. Gill, A. M., Belbin, L. & Chippendale, G. *Phytogeography Of Eucalyptus in Australia* (Bureau of Flora and fauna, 1985).
24. Schuster, T. M. et al. Chloroplast variation is incongruent with classification of the Australian bloodwood eucalypts (genus *Corymbia*, family Myrtaceae). *PLoS ONE* **13**, 1–28 (2018).
25. Lee, D. J. Achievements in forest tree genetic improvement in Australia and New Zealand 2: development of *Corymbia* species and hybrids for plantations in eastern Australia. *Aust. For.* **70**, 11–16 (2007).
26. Brawner, J. T., Lee, D. J., Meder, R., Almeida, A. C. & Dieters, M. J. Classifying genotype by environment interactions for targeted germplasm deployment with a focus on *Eucalyptus*. *Euphytica* **191**, 403–414 (2013).
27. Dickinson, G. R., Wallace, H. M. & Lee, D. J. Reciprocal and advanced generation hybrids between *Corymbia citriodora* and *C. torelliana*: forestry breeding and the risk of gene flow. *Ann. Sci.* **70**, 1–10 (2013).
28. Grant, J. C. et al. Depth distribution of roots of *Eucalyptus dunnii* and *Corymbia citriodora* subsp. *variegata* in different soil conditions. *For. Ecol. Manag.* **269**, 249–258 (2012).
29. Butler, J. B. et al. Independent QTL underlie resistance to the native pathogen *Quambalaria pitereka* and the exotic pathogen *Austropuccinia psidii* in *Corymbia*. *Tree Genet. Genomes* **15**, 72 (2019).
30. Brawner, J. T., Lee, D. J., Hardner, C. M. & Dieters, M. J. Relationships between early growth and *Quambalaria* shoot blight tolerance in *Corymbia citriodora* progeny trials established in Queensland, Australia. *Tree Genet. Genomes* **7**, 759–772 (2011).
31. Padovan, A., Keszei, A., Külheim, C. & Foley, W. J. The evolution of foliar terpene diversity in Myrtaceae. *Phytochem. Rev.* **13**, 695–716 (2014).
32. Butler, J. B. et al. Annotation of the *Corymbia* terpene synthase gene family shows broad conservation but dynamic evolution of physical clusters relative to *Eucalyptus*. *Heredity* **121**, 87–104 (2018).
33. Külheim, C. et al. The *Eucalyptus* terpene synthase gene family. *BMC Genomics* **16**, 450 (2015).
34. Healey, A. L. Genomic and phenotypic characterization of commercial *Corymbia* hybrids for lignocellulosic biofuel production. PhD Thesis, Queensland Alliance for Agriculture and Food Innovation (The University of Queensland, 2017).
35. Jaffe, D. B. et al. Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
36. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
37. Butler, J. B. et al. Comparative genomics of *Eucalyptus* and *Corymbia* reveals low rates of genome structural rearrangement. *BMC Genomics* **18**, 1–13 (2017).
38. Sansaloni, C. et al. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proc.* **5**, P54 (2011).
39. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 1–15 (2015).
40. Grattapaglia, D. & Bradshaw, H. J. Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Can. J. For. Res.* **24**, 1074–1078 (1994).
41. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
42. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **D1**, 1178–1186 (2012).
43. Zhou, R. et al. Characterization of a large sex determination region in *Salix purpurea* L. (Salicaceae). *Mol. Genet. Genomics* **293**, 1437–1452 (2018).
44. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
45. Dai, X. et al. The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res.* **24**, 1274–1277 (2014).
46. Silva-Junior, O. B. & Grattapaglia, D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. *N. Phytol.* **208**, 830–845 (2015).
47. Bacles, C. F. E. et al. Reproductive biology of *Corymbia citriodora* subsp. *variegata* and effective pollination across its native range in Queensland, Australia. *South. For.* **71**, 125–132 (2009).
48. Bortoloto, T. M. et al. Identification of a molecular marker associated with lignotuber in *Eucalyptus* ssp. *Sci. Rep.* **10**, 1–8 (2020).
49. Gion, J. M. et al. Genome-wide variation in recombination rate in *Eucalyptus*. *BMC Genomics* **17**, 1–12 (2016).
50. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433 (2016).
51. Da Costa, I. R., Dornelas, M. C. & Forni-Martins, E. R. Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae. *Plant Syst. Evol.* **276**, 209–217 (2008).
52. Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann. Bot.* **95**, 45–90 (2005).
53. Wang, W. et al. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. *Gigascience* **9**, 1–12 (2020).
54. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, 1–14 (2012).
55. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-visited. *Am. J. Bot.* **97**, 1296–1303 (2010).
56. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
57. Ermolaeva, M. D., Wu, M., Eisen, J. A. & Salzberg, S. L. The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol. Biol.* **51**, 859–866 (2003).
58. Dos Santos, I. B. et al. The *Eucalyptus* cuticular waxes contribute in preformed defense against *Austropuccinia psidii*. *Front. Plant Sci.* **9**, 1–13 (2019).
59. Berthon, K., Esperon-Rodriguez, M., Beaumont, L. J., Carnegie, A. J. & Leishman, M. R. Assessment and prioritisation of plant species at risk from myrtle rust (*Austropuccinia psidii*) under current and future climates in Australia. *Biol. Conserv.* **218**, 154–162 (2018).
60. Adams, Z. P., Ehrling, J. & Edwards, R. The regulatory role of shikimate in plant phenylalanine metabolism. *J. Theor. Biol.* **462**, 158–170 (2019).
61. Lavell, A. A. & Benning, C. Cellular organization and regulation of plant glycerolipid metabolism. *Plant Cell Physiol.* **60**, 1176–1183 (2019).
62. Facchini, P. J. Alkaloid biosynthesis in plants: Biochemistry, cell biology, molecular regulation, and metabolic engineering applications. *Annu. Rev. Plant Biol.* **52**, 29–66 (2001).
63. Ziegler, J. & Facchini, P. J. Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* **59**, 735–769 (2008).
64. Resende, R. T. et al. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *N. Phytol.* **213**, 1287–1300 (2017).
65. Zhang, M. et al. Comparative transcriptome profiling of the maize primary, crown and seminal root in response to salinity stress. *PLoS ONE* **10**, 1–16 (2015).
66. Winkel-Shirley, B. Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* **5**, 218–223 (2002).
67. Adrian, M. et al. Metabolic fingerprint of ps3-induced resistance of grapevine leaves against *Plasmopara viticola* revealed differences in elicitor-triggered defenses. *Front. Plant Sci.* **8**, 1–14 (2017).
68. Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C. & Flügge, U. I. The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*. *Plant J.* **51**, 247–261 (2007).
69. Danquah, A., de Zelicourt, A., Colcombet, J. & Hirt, H. The role of ABA and MAPK signaling pathways in plant abiotic stress responses. *Biotechnol. Adv.* **32**, 40–52 (2014).
70. Qi, Y. et al. Physical association of *Arabidopsis* Hypersensitive Induced Reaction proteins (HIRs) with the immune receptor RPS2. *J. Biol. Chem.* **286**, 31297–31307 (2011).
71. Gao, Z., Chung, E.-H., Eitas, T. & Dangel, J. L. Plant intracellular innate immune receptor resistance to *Pseudomonas syringae* pv. *maculicola* 1 (PRM1) is activated at, and functions on, the plasma membrane. *Proc. Natl Acad. Sci. USA* **108**, 8914 (2011).
72. Kawasaki, T. et al. A duplicated pair of *Arabidopsis* RING-finger E3 ligases contribute to the RPM1- and RPS2-mediated hypersensitive response. *Plant J.* **44**, 258–270 (2005).
73. Schwessinger, B. et al. Phosphorylation-dependent differential regulation of plant growth, cell death, and innate immunity by the regulatory receptor-like kinase BAK1. *PLoS Genet.* **7**, e1002046 (2011).
74. Robatzek, S. & Wirthmueller, L. Mapping FLS2 function to structure: LRRs, kinase and its working bits. *Protoplasma* **250**, 671–681 (2013).
75. Ladiges, P. et al. Historical biogeographical patterns in continental Australia: congruence among areas of endemism of two major clades of eucalypts. *Cladistics* **27**, 29–41 (2011).
76. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, 1–14 (2012).
77. Yeats, T. H. & Rose, J. K. C. The formation and function of plant cuticles. *Plant Physiol.* **163**, 5–20 (2013).
78. Tucker, D. J. et al. A metabolomic approach to identifying chemical mediators of mammal-plant interactions. *J. Chem. Ecol.* **36**, 727–735 (2010).

79. Bonora, F. S., Hayes, R. A., Nahrung, H. F. & Lee, D. J. Spotted gums and hybrids: impact of pests and diseases, ontogeny and climate on tree performance. *For. Ecol. Manag.* **472**, 118235 (2020).
80. Hallam, N. D. & Chambers, T. C. The leaf waxes of the genus *Eucalyptus* L'Héritier. *Aust. J. Bot.* **18**, 335–386 (1970).
81. Hoffmann, B., Kahmen, A., Cernusak, L. A., Arndt, S. K. & Sachse, D. Abundance and distribution of leaf wax n-alkanes in leaves of acacia and eucalyptus trees along a strong humidity gradient in Northern Australia. *Org. Geochem.* **62**, 62–67 (2013).
82. Thomas, D. & Barber, H. Studies on leaf characteristics of a cline of *Eucalyptus urnigera* from Mount Wellington, Tasmania. I. Water repellency and the freezing of leaves. *Aust. J. Bot.* **22**, 501–512 (1974).
83. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. 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 (2011).
84. Lawler, I. R., Stapley, J., Foley, W. J. & Eschler, B. M. 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 (1999).
85. O'Reilly-Wapstra, J. M., McArthur, C. & Potts, B. M. Linking plant genotype, plant defensive chemistry and mammal browsing in a *Eucalyptus* species. *Funct. Ecol.* **18**, 677–684 (2004).
86. Pichersky, E. & Gershenzon, J. The formation and function of plant volatiles: perfumes for pollinator attraction and defense. *Curr. Opin. Plant Biol.* **5**, 237–243 (2002).
87. Peñuelas, J., Llusià, J., Asensio, D. & Munné-Bosch, S. Linking isoprene with plant thermotolerance, antioxidants and monoterpene emissions. *Plant, Cell Environ.* **28**, 278–286 (2005).
88. Cheng, A. X. et al. Plant terpenoids: biosynthesis and ecological functions. *J. Integr. Plant Biol.* **49**, 179–186 (2007).
89. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl Acad. Sci. USA* **95**, 4126–4133 (1998).
90. Dillon, S. K., Brawner, J. T., Meder, R., Lee, D. J. & Southerton, S. G. Association genetics in *Corymbia citriodora* subsp. *variegata* identifies single nucleotide polymorphisms affecting wood growth and cellulose pulp yield. *N. Phytol.* **195**, 596–608 (2012).
91. Tambarussi, E. V., Pereira, F. B., da Silva, P. H. M., Lee, D. & Bush, D. Are tree breeders properly predicting genetic gain? A case study involving *Corymbia* species. *Euphytica* **214**, 150 (2018).
92. Silva-Junior, O. B., Faria, D. A. & Grattapaglia, D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *N. Phytol.* **206**, 1527–1540 (2015).
93. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 1–8 (2014).
94. Tibbits, J., McManus, L. J., Spokevicius, A. & Bossinger, G. A rapid method for tissue collection and high throughput genomic DNA isolation from mature trees. *Plant Mol. Biol. Rep.* **24**, 1–11 (2006).
95. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
96. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
97. Kent, W. J. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
98. Lovell, J. T. et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).
99. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
100. Smit, A. F., Hubley, R., Green, P. & Smit, H. RepeatMasker Open-3.0. (1996).
101. Smit, A. F. RepeatModeler.
102. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
103. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
104. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
105. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
106. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
107. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
108. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, 609–612 (2006).
109. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
110. Haubold, B., Pfaffelhuber, P. & Lynch, M. MRho - A program for estimating the population mutation and recombination rates from shotgun sequenced diploid genomes. *Mol. Ecol.* **19**, 277–284 (2010).
111. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010).
112. Depristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).
113. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
114. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
115. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
116. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
117. Alexa, A. & Rahnenfurer, J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.38.1. (2016).
118. Yu, G. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
119. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

## Acknowledgements

The authors acknowledge nuclear sequence data were produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>) in collaboration with the user community and the Joint BioEnergy Institute (<http://www.jbei.org/>). Sequencing of RNA transcripts was carried out by AGRF funded by Southern Cross Plant Sciences internal funding. The work conducted by the Joint Genome Institute and the Joint BioEnergy Institute is supported by the Office of Science in the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. Germplasm used in this study was provided by the Queensland Department of Agriculture and Fisheries. The Future Biofuels project was funded by the Queensland Government Smart Futures Research Partnerships Program. The *Corymbia* genetic map construction was supported by the Australian Research Council (grant numbers DP140102552, DP110101621) and an Australian Government Research Training Program Scholarship. Travel for this research was supported by the University of Queensland Joan Allsop Scholarship. The authors also thank Joe Carlson of the Joint Genome Institute for coordinating the submission of the genome to GenBank, Ramil Mauleon of Southern Cross University for submitting RNASeq data into the SRA, and the anonymous reviewers whose comments and suggestions greatly improved the manuscript.

## Author contributions

A.L.H., M.S., G.J.K., D.J.L., A.B., J.S.F., B.M.P., J.B.B., D.G., O.S.J., A.F., R.E.V., and R.J.H. constitute members of the *Corymbia* genome consortium. M.S., G.J.K., D.J.L., B.M.P., B. A.S., D.P., J.S., R.E.V., and R.J.H. initiated project and coordinated research and resources. J.G., H.H., K.B., and J.S. generated sequencing data. J.J., J.B.B., J.S.F., S.S., and A.B. generated the initial genome assembly, genetic maps, chromosome build, and annotation. A.L.H., A.S., J.L., and O.S.J. performed comparative genome analysis, syntenic block construction, and enrichment analysis. A.L.H. wrote the paper with significant contributions and input from M.S., G.J.K., D.J.L., B.M.P., J.B.B., J.S.F., and O.S.J. A.F., D.G., B.A.S., and R.J.H. assisted with manuscript editing. All authors have read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02009-0>.

**Correspondence** and requests for materials should be addressed to A.L.H.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021