5-7-2019

High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in Eucalyptus

David Kainer
Oak Ridge National Laboratories

Amanda Padovan
The Australian National University

Joerg Degenhardt
Martin-Luther Universität Halle-Wittenberg

Sandra Krause
Martin-Luther Universität Halle-Wittenberg

Produyut Mondal
Martin-Luther Universität Halle-Wittenberg

See next page for additional authors

DR CARSTEN KULHEIM (Orcid ID : 0000-0002-0798-3324)

Article type : Regular Manuscript

High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*

David Kainer[1,2*], Amanda Padovan[2,3], Joerg Degenhardt[4], Sandra Krause[4], Prodyut Mondal[4], William J. Foley[2], Carsten Külheim[2,5]

[1] Center for BioEnergy Innovation, Bioscience Division, Oak Ridge National Laboratories, Oak Ridge, TN 37831 USA; [2] Research School of Biology, The Australian National University, Acton 2601, Australia; [3] CSIRO, Clunies Ross Street, Canberra 2601, Australia; [4] Institut für Pharmazie, Martin-Luther Universität Halle-Wittenberg, 06120 Halle (Saale), Germany; [5] School of Forest Resources and Environmental Science, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931 USA

Corresponding Author
Carsten Külheim
ph: +19064871615
ckulheim@mtu.edu

Received: *16 October 2018*
Accepted: *26 April 2019*

ORCID IDs: DK: 0000-0001-7271-4676; AP: 0000-0002-8118-9137; JD: 0000-0003-0510-1006, SK: 0000-0001-8415-1970; PM: 0000-0002-8356-5634; WJF: 0000-0001-8587-1814; CK: 0000-0002-0798-3324

Summary

- Terpenoid based essential oils are economically important commodities, yet beyond their biosynthetic pathways, little is known about the genetic architecture of terpene oil yield from plants. Transport, storage, evaporative loss, transcriptional regulation and precursor competition may be important contributors to this complex trait.
- Here, we associate 2.39 M single nucleotide polymorphisms derived from shallow whole genome sequencing of 468 *Eucalyptus polybractea* individuals with 12 traits related to the overall terpene yield, eight direct measures of terpene concentration and four biomass-related traits.
- Our results show that in addition to terpene biosynthesis, development of secretory cavities where terpenes are both synthesised and stored, and transport of terpenes were important components of terpene yield. For sesquiterpene concentrations, the availability of precursors in the cytosol was important. Candidate terpene synthase genes for the production of 1,8-cineole and α-pinene, and β-pinene, (which made up more than 80% of the total terpenes) were functionally characterised as a 1,8-cineole synthase and a β / α-pinene synthase.
- Our results provide novel insights of the genomic architecture of terpene yield and we provide candidate genes for breeding or engineering of crops for biofuels or the production of industrially valuable terpenes.

Keywords: 1,8-cineole, essential oil, eucalypt, genome-wide association study (GWAS), pinene, terpene oil yield, terpenes, terpene synthase

Introduction

Plant derived essential oils such as lavender oil, tea tree oil and eucalyptus oil are valued for their use in pharmaceuticals, cosmetics, cleaning products and industrial solvents and are gaining in commercial importance (Bohlmann & Keeling, 2008; Vickers *et al.*, 2014). Although they typically contain dozens to hundreds of chemical compounds, they are often dominated by one or a few terpenoids. For example, lavender oil is dominated by the monoterpene alcohol linalool and the monoterpene ester linalyl acetate (Renaud *et al.*, 2001), tea tree oil is dominated by the monoterpene alcohol terpinen-4-ol (Butcher *et al.*, 1994), and eucalyptus oil by the monoterpene 1,8-cineole (Barton, 1999). Often these dominant terpenes give the essential oil its commercial value, and certain terpenes are increasingly being sought as feedstock for new products. For example, for the production of advanced renewable biofuels, the monoterpenes α- and β-pinene, camphene and limonene are particularly useful (Mewalal *et al.*, 2017), and may be sourced from many plant species.

Essential oil producing plant species are largely undomesticated and huge variation exists within species both quantitatively (up to 20-fold intraspecific variation; (King *et al.*, 2006; Webb *et al.*, 2014; Kainer *et al.*, 2017)) and qualitatively (Keefover-Ring *et al.*, 2009;

Padovan *et al.*, 2014), which affects the consistency of harvested oil yield, or more specifically, the yield of valuable terpenes. Some efforts have been made to improve oil yield and consistency through traditional breeding. For example, tea tree breeding programs have succeeded in doubling terpene yield over two decades (Baker *et al.*, 2014). However, knowledge of the genetic basis of complex phenotypic variation, here terpene yield, may lead to greater improvement in a shorter time. In order to do so, we need to investigate the genetic components of the complex trait of terpene yield, which is a function of the concentration of terpenes in the leaf and the total amount of leaf mass that can be harvested (Kainer *et al.*, 2017).

The most studied aspect of terpene yield is the biosynthesis of terpenes (reviewed by (Lichtenthaler, 1999; Lange *et al.*, 2000; Keszei *et al.*, 2008; Vranova *et al.*, 2013); Figure 1). All terpenes are derived from the C5 precursor isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP), which are synthesised in two spatially separate pathways in plants (Lange *et al.*, 2000). The methylerythritol phosphate (MEP) pathway, of the chloroplast, produces IPP and DMAPP from pyruvate and glyceraldehyde-3-phosphate. The mevalonate (MVA) pathway, in the cytosol, produces IPP from acetyl-CoA, some of which is converted to DMAPP by the isopentenyl diphosphate isomerase (IPPI) enzyme. Downstream, IPP and DMAPP are condensed into prenyl-diphosphates (GPP, GGPP, FPP) which are the immediate substrates for terpene synthesis. In the chloroplast, terpene synthase (TPS) enzymes produce monoterpenes, diterpenes and tetraterpenes, while sesquiterpenes and triterpenes are produced in the cytosol. Further modification of terpenes often takes place through the actions of cytochrome P450 monooxygenases (Hamberger & Bohlmann, 2006) or glucosyl transferases (Rivas *et al.*, 2013).

Studies of the genetic basis of terpene variation have focused on candidate genes found in the MEP and MVA pathways. Within the Myrtaceae, Külheim *et al.* (2011) found allelic variants in the *hds* and *hdr* genes at the end of the MEP pathway that were significantly associated with the foliar concentration of the monoterpene 1,8-cineole in a population of *E. globulus*. Similarly, in *E. loxophleba hds* and *mcs* were associated with variation in concentration of 1,8-cineole and *mcs, hdr* and *mct* were associated with total terpene concentration (Padovan *et al.*, 2017). Such candidate gene studies, however, are limited by their inherent bias. That is, they can only reveal details in what is already known from biochemical studies (Cappa *et al.*, 2013), and typically explain only a small portion of the additive genetic heritability of a trait.

Other aspects of the genetic basis of variation in foliar terpene yield such as biomass accumulation, loss of volatile terpenes, transport of terpenes and their storage, are less studied. Terpenes can either be continuously released to the atmosphere or stored in specialized structures such as trichomes or schizogenic secretory cavities. In the case of tea tree and *Eucalyptus*, terpenes are stored in secretory cavities within the leaf, and their capacity may set a limit to the accumulation of oil in a leaf (King *et al.*, 2004). Goodger and Woodrow (2012) showed that foliar terpene concentration in *E. polybractea* is tightly correlated with the total volume of secretory cavities, yet the genetic basis of the formation of these cavities is poorly understood. Thus, there are potentially hundreds of other quantitative trait loci (QTL) outside of the MEP or MVA pathways and *tps* genes affecting terpene yield.

Genome-wide association studies (GWAS) can identify the genomic architecture of a trait, through identification of candidate genes, followed by functional characterization. To date, several GWAS have been performed in forest tree species (Cappa *et al.*, 2013; Uchiyama *et al.*, 2013; Evans *et al.*, 2014; McKown *et al.*, 2014; Resende *et al.*, 2017; Müller *et al.*, 2019), but none for terpene traits. One of the reasons for this is that the undomesticated, outcrossing nature of forest tree species causes very rapid decay of LD (Külheim *et al.*, 2009; Thavamanikumar *et al.*, 2013), which is simultaneously a curse and a blessing from the perspective of GWAS. Since the power to detect significant QTLs is a function of the extent of LD (Long & Langley, 1999), populations with very short LD require very high marker densities to ensure that most QTL are in linkage with at least one marker, making this approach cost prohibitive in reasonable large populations until recently. On the other hand, with short range LD and dense markers obtained from NGS sequencing, GWAS becomes an ideal tool to locate significant loci at the gene and sub-gene level. This is not possible using sparse markers in *Eucalyptus*, such as those obtainable by GBS or the EuCHIP60K SNP chip (Silva-Junior *et al.*, 2015).

In this study, we perform GWAS in *Eucalyptus polybractea* using a dense sequence-based marker set. We obtained millions of biallelic SNPs via low-depth whole genome re-sequencing of a population of 480 *E. polybractea* from 40 open pollinated half-sib families. However, the cost of sequencing hundreds of whole genomes means that each individual was sequenced to low (1-8 x) depth, which presents challenges for accurate genotyping and association testing. We applied an LD-aware genotyping approach to reduce genotyping error rates due to the low-depth samples. We performed single-SNP association tests with 12 traits being eight related to terpene yield and four to biomass traits. We also performed multivariate and regional GWAS analysis to increase power in this moderately sized experiment and find SNPs and regions that affect multiple terpene-related traits. The outcomes of this study provide a greater understanding of the genetic architecture of terpene yield and will help inform programs to selectively breed or genetically engineer Myrtaceae species for improved terpene production.

Materials and Methods

Phenotypes

We analysed phenotypes and genotypes of trees from a progeny experiment of *Eucalyptus polybractea* (R.T. Baker) located on the property of GR Davis Pty Ltd, near West Wyalong, NSW Australia (Lat. 33°58’S Long. 147°03’E). The experiment contains half-siblings from 40 open-pollinated families, where all 40 maternal trees originated from six sites around the immediate West Wyalong region (Supporting Information Fig. S1). We selected 12 trees from each of the 40 families with visually clear phenotypic diversity from large to small individuals, for a total of n = 480. Details of the phenotyping protocol and analysis of terpene and biomass traits can be found in Kainer *et al.* (2017).

Twelve phenotypes were selected for downstream analysis: total terpene concentration (TTC), total monoterpene concentration (MONO), total sesquiterpene concentration (SESQ), monoterpene:sesquiterpene ratio (MSratio), 1,8-cineole concentration (CIN), 1,8-cineole proportion (PCIN), α-pinene concentration (APIN), β-pinene concentration (BPIN), leaf area (LA), height (HT), 1 year change in height (ΔHT between 1, and 2 years post coppice), 1 year change in crown area (ΔCA between 1, and 2 years post coppice). Some traits were transformed in R when they were not normally distributed using log transform (MSratio, APIN and BPIN) and square root transform (SESQ).

DNA extraction, library preparation, sequencing, QC and alignment

DNA from all 480 individuals was extracted and sequenced as described in Kainer *et al.* (2018). Briefly, after DNA extraction with lysis stage extended to 20 min (DNeasy Plant kit, Qiagen, Valencia, CA), 500 – 1000 ng DNA / sample was fragmented with a Diagenode Bioruptor NGS (Diagenode, Denville NJ), with 13 cycles of 30 s on and 30 s off at maximum intensity. After blunt-end repair, a universal Illumina P7 adapter and one of 96 barcoded P5 adapters were ligated to the ends of genomic DNA fragments from each sample using a protocol as described by Rohland and Reich (2012). The resulting libraries were pooled by barcode distance and DNA concentration into 22 pools. We first sequenced one individual at the Biomolecular Resource Facility (Australian National University;MiSeq, 2 x 150 bp) to test our custom made library preparation method. This was followed by testing a pool of 13 samples at our local facility to verify that multiplexing worked on the HiSeq2500 platform (2 x 150 bp). The remainder of pools were sequenced at Macrogen (Republic of Korea; 1 pool of 14, 1 pool of 20 and 18 pools of 24 samples, 2 x 125 bp). Five lanes produced relatively poor total output and were sequenced a second time, creating a set of 120 low-depth replicates for the samples in those pools.

Sequence read quality was assessed with FastQC (Andrews 2010) and we used the BBtools suite v36 (Bushnell, 2016) to remove adapters and trim low quality bases from the pooled fastq reads, followed by demultiplexing with Flexbar v2.5 (Dodt *et al.*, 2012) allowing for 1 mismatch per barcode. *Eucalyptus grandis* is one of a few eucalypt species planted around tropical and subtropical regions worldwide and was the first in its genus with an annotated genome sequence (Myburg *et al.,* 2014). Due to the estimated divergence of *E. grandis* and *E. polybractea* ~21 mya (Thornhill *et al.*, 2015; González-Orozco *et al.*, 2016), we anticipated that there would be significant sequence diversity. To reduce the mismatch rate we used a two stage alignment strategy. First, we aligned the cleaned reads of each sample to the *E. grandis* v2.0 reference genome (Myburg *et al.*, 2014; Bartholomé *et al.*, 2015) using BWA MEM v0.7.12 (Li & Durbin, 2009) on default settings. Next, we called variants in 40 individuals using both Freebayes v0.9.21 (Garrison & Marth, 2010) and Samtools v1.3.1 (Li *et al.*, 2009). After variant filtering (100% call rate and Minor Allele Frequency (MAF) > 0.80) we derived the intersection of the two sets. This represented an *E. polybractea* "species" set of fixed or nearly fixed variants in *E. polybractea* relative to *E. grandis*. We

then replaced sites in the *E. grandis* reference genome with the alternate alleles from *E. polybractea* using BCFtools (Li *et al.*, 2009). Each of the 480 samples plus 120 replicate samples were then realigned to the new reference genome, resulting in 600 BAM files. We merged replicate BAMs for those 120 samples that were sequenced twice. In preparation for genotyping we marked PCR and optical duplicates, followed by left alignment around INDELs using GATK 3.6 (DePristo *et al.*, 2011). Ten samples were deemed to have inadequate coverage for variant calling.

Genotyping

Due to the low sequencing coverage, we elected to make use of LD-aware imputation and genotype refinement techniques provided by Thunder (Li *et al.*, 2011) as part of the GotCloud pipeline (Jun *et al.*, 2015). This pipeline calculates genotype likelihoods, phases haplotypes, imputes missing genotypes, and finally refines error-prone genotype calls based on LD and haplotype frequencies using samples with higher sequencing depth as guidance. The population-scaled mutation rate prior was set to 0.02 as calculated from genotype likelihoods by Angsd 0.913 (Korneliussen *et al.*, 2014). Thunder was run with 20 iterations on the National Computer Infrastructure HPC (nci.org.au). The variants output from Thunder excluded those with low (<0.95) average genotype posterior probability and low (< 0.90) imputation confidence. The performance of the genotyping approach was assessed by including two low depth replicates (labelled A and B) for 12 of the samples that were sequenced twice, and then measuring the percentage of genotypes that were discordant between A and B, and between each individual replicate (A or B) and its corresponding merged sample (AB).

Variant Filtering

Variants output from Thunder were filtered to remove those with low average genotype quality (GQ < 20), low or very high population-level depth (DP<800; DP>3300), or low MAF (< 0.02). Variants were also removed if they fell within regions of low complexity (entropy), which were calculated with the BBtools suite *bbmask* feature. In order to test whether SNPs departed from Hardy-Weinberg Equilibrium (HWE) without family structure, we randomly selected one individual per family and recorded the *p*-value from the HWE test of each SNP in that unrelated group using the SNPRelate v1.8.0 R package (Zheng *et al.*, 2012). We repeated this process with 100 different sets of random unrelated individuals. SNPs with a median *p*-value across the 100 tests of less than 1.0×10^{-2} were deemed to be consistently out of HWE and were thus removed. Finally, any SNP where more than one of the 12 duplicated samples had discordant genotypes within its replicates was excluded from further analysis. The remaining SNPs are described as the Thunder SNP set. SNP data is available at https://figshare.com/s/be57a3a4d49742dd6fdf.

Population Stratification and kinship

For population stratification analysis we performed a linkage disequilibrium (LD)-pruning, with the Thunder SNP set, using the --*indep* function in PLINK v1.9 (Chang *et al.*, 2015), which produces a set of SNPs that have reduced collinearity between them. We used default parameters as described in the Plink documentation: a sliding window of 50 bases shifting by 5 bases, a variance inflation factor (VIF) of 1.5 and MAF > 0.05. Principal components were calculated using the GENESIS v2.4.0 R package (Conomos & Thornton, 2016). We then used these in the estimation of pairwise kinship with GENESIS, refining pairwise kinship coefficients (θ_{IBD}) by removing the signal of ancestral population structure, resulting in only recent familial and/or cryptic relatedness. We estimated linkage disequilibrium (LD) decay for 2.39 M SNPs and all 468 individuals, as well as three families (7, 35 and 37) and from a random selection of 12 individuals using the method described in Marroni *et al.* (2011).

Single-SNP GWAS, regional heritability mapping, and multivariate GWAS

Detailed methods for single-SNP GWAS, regional heritability mapping and multivariate GWAS are described in the Supplemental Information Methods S1; in brief, for single-SNP association testing we used a slightly pruned Thunder SNP set (VIF=50) with the GENESIS R package using the 'assocTestMM' function. Bonferroni correction for multiple testing was considered too stringent because the GWAS SNP set contains many SNPs in relatively strong LD, which results in non-independence of each association test. We applied the less stringent false discovery rate (FDR) with the Benjamini and Hochberg method on a per chromosome basis and selected SNPs that had a FDR < 0.1 as significant. Furthermore, we used experimental validation of several GWAS hits to verify some of the results of the GWAS (see methods below). For regional heritability mapping, the full set of 2.39 M SNPs was divided into regions of 100 contiguous SNPs without overlap between regions using a custom script in R. We then performed regional heritability mapping using the Genome-wide complex trait analysis tool (GCTA) for each region and trait (Yang *et al.*, 2011). We applied false discovery rate (FDR) with the Benjamini and Hochberg method on a per chromosome basis and selected regions that had a FDR < 0.1 as significant. For the multivariate GWAS, we used the multivariate version of BIMBAM (Stephens, 2013) to perform genetic association of multiple related phenotypes. We used five terpene traits: CIN, APIN, SESQ, MONOREST (where MONOREST = MONO – CIN – APIN) and PCIN.

Terpene synthase functional characterisation

We selected several *tps* genes that were declared significantly associated with 1,8-cineole (CIN), α- or β-pinene (APIN or BPIN) to test whether we can infer function from association. Detailed methods are described in the Supplemental Information Methods S2, in short total RNA was extracted from samples that had a high proportion of either 1,8-cineole or α-pinene followed by cDNA synthesis. Genes were amplified with primers (Supporting Information Table S1) based on the *E. polybractea* pseudo reference. PCR products were cloned into the pCR4-TOPO or pJET vector and sequenced with Sanger sequencing technique. Three full-

length ORFs were isolated which showed high similarity to *tps* genes and were called *Eptps1*, *Eptps2* and *Eptps3* (GenBank submission: #2219851).

The ORFs were cloned into the expression vector pASK-IBA37+ and amplified with nested primers (Supporting Information Table S1). Vectors containing the three *TPS* genes were transformed into *E. coli* TOP10 cells and grown on selective medium. Gene expression was induced for 20 h at 18°C. Cells were disruption by sonication and the crude protein extract was transferred into assay buffer.

Terpene synthase activity assays were performed with 30 µl crude enzyme extract and 70 µl reaction mix. Enzyme products were collected by a solid phase microextraction fiber (SPME), exposed to the headspace of the assay mixture for 45 min at 35°C in a water bath. Terpene products were identified with a gas chromatograph mass spectrometer.

Results

Sequencing and Alignment

Illumina sequencing produced 1.52 terabases of data. We removed low quality bases, barcodes and adapters, which reduced the amount of data to 1.37 terabases (Supporting Information Table S2). During the first stage of the alignment strategy, we generated a pseudo *E. polybractea* reference, by replacing 3.35 M (0.547% of 620 Mbp) sites in the *E. grandis* reference with high frequency homozygous alternate variants from our population. When we re-aligned all samples to the new pseudo reference. The average mapping quality (MQ) improved by 0.6 and slightly increased the number of successfully mapped reads and percentage of proper pairs. This was a consequence of average mismatches per read dropping from 5.5 to 4.9 (Supporting Information Table S3). Across the *E. grandis* genome, 56% was covered by at least one read in each sample, with one sample covering up to 73% (Supporting Information Table S3, Fig. S2). On average 22% of the genome was covered by a depth of 4x or greater across all samples.

Genotyping

The GotCloud genotyping pipeline produced 6.28 M SNPs in the 11 chromosomes of the reference genome. Applying filters reduced the SNP number to 2.39 M. The MAF distribution of the 2.39 M SNPs showed that the majority of SNPs had a MAF < 0.1 (Supporting Information Fig. S3). Of the 2.39 M SNPs, 1.03 M were located within the 34,110 annotated genes from the 11 primary chromosome scaffolds of *E. grandis* (Supporting Information Table S4), and another 0.82 M SNPs were within 5 kb of a gene. We defined a candidate gene set for terpene traits that contained 123 genes that are directly involved in terpenoid biosynthesis (MEP and MVA pathway genes plus *tps* genes) and located in the 11 *E. grandis* chromosomes (Supporting Information Table S5). Of these, 89 genes contained at least one SNP in the Thunder SNP set.

Population Stratification

We used a pruned SNP set of 307,136 SNPs with MAF > 0.05 for the population stratification analyses. The principal component analysis showed a small degree of ancestral population structure although the majority of samples clustered into one general population (Supporting Information Fig. S4). The first three PCs, respectively captured 2.92%, 2.37% and 2.16% of the variation in the genotypic data. Pairwise kinship analysis revealed that some siblings were closer to a full-sib relationship than the expected half-sib relationship of 0.25 (see Figure 1 in Kainer *et al.,* 2018). Family 37 showed clear signatures of inbreeding, as shown by a much slower rate of LD decay compared to the other families (Supporting Information Fig. S5). Two sequentially labelled samples appeared to be unintentional sequencing duplicates (Coeff of Ancestry = 1.03), and were removed from further analysis leaving 468 individuals for the GWAS.

Single-SNP GWAS

We used the Thunder SNP set of 2.39 M SNPs for association analyses using a linear mixed model approach. Trait heritability estimates based on the null model from the analysis ranged from h^2 = 0.15 for change in crown area (ΔCA) to 0.76 for concentration of sesquiterpenes (SESQ; Supporting Information Table S6). We found 2,623 SNPs associating (Benjamini Hochberg FDR α = 0.1) with one or several of the traits investigated here (Supportive Information Table S7). Composite terpene traits total terpene concentration (TTC) and total monoterpene concentration (MONO) yielded relatively few SNPs with strong significance and a noticeable paucity of significant SNPs in the candidate gene set (see Table 1, Supporting Information Table S7). There was a slight enrichment of the effect of SNPs in these candidate genes relative to all other genes (Table 1). For example, for concentration of 1,8-cineole (CIN), the mean absolute SNP-effect of SNPs in the candidate set ($|EFF| = 1.26$) was significantly higher than the mean absolute SNP-effect of SNPs in all genes ($|EFF| = 1.23$; $p = 0.046$). In contrast, for biomass-related traits such as height (HT) and leaf area (LA), the mean SNP-effect in terpene candidate genes was not significantly different to the mean SNP-effect in all genes, as expected. This trend was also present when comparing the SNP effect of 100 random sets of 100 SNPs from a similar number of genes as the candidate gene set (Supporting Information Fig. S6).

The strongest genetic associations were identified for β-pinene (BPIN), forming a QTL on chromosome 1 from 25.45 to 25.50 Mbp, with a peak SNP at 25.455 Mbp (FDR 1.6 x 10^{-41}; Figure 2, Figure 3a). *Eucalyptus grandis* 2.0 annotations from Phytozome show no annotated genes in this immediate region, nor is there any recorded gene expression, the closest annotated terpene synthase is EgTPS060 at 24.5 Mbp (annotation: (Külheim *et al.*, 2015)). However, the BlastX track in Phytozome JBrowse (Figure 3b) revealed high similarity to putative monoterpene synthase genes found in *Arabidopsis thaliana*, such as AT4G16730.1 (https://phytozome.jgi.doe.gov/jbrowse/index.html?data=genomes%2FEgrandis&loc=Chr01

%3A25450024..25500024). Based on sequence similarity, the two closest *tps* of *E. grandis* are EgTPS055 and EgTPS058 (data not shown).

Figure 4 shows Manhattan plots for terpene traits, with details of high-ranking SNP associations shown in Table 2 and Supporting Information Table S7. Aside from BPIN, the strongest associations for any trait were two SNPs (Chr10_27384836, FDR 1.56 x 10^{-9} & Chr10_27384848, FDR 2.73 x 10^{-9}) associated with SESQ and located in *phosphoenolpyruvate (pep)/phosphate translocator 2* (Eucgr.J02222). These two SNPs were also highly significantly associated with the derived trait monoterpene:sesquiterpene ratio (MSratio) but not with any other trait. A SNP on chromosome 1 (Chr01_31758048), located within a cluster of monoterpene synthases, was the most significant association with α-pinene concentration (APIN; FDR 2.54 x 10^{-5}) and was also one of the most significant associations with CIN (FDR 0.042). Several SNPs at Chr10 19.02 Mbp were associated with both TTC and SESQ and are located within a gene encoding an S-adenosyl-L-methionine-dependent methyltransferase (SAM-Mtase) gene.

Growth in tree height from 1-year post-coppice to 2-years post-coppice (ΔHT) presented no associations with notably strong statistical significance. Nevertheless, several of the top-ranking associations were located within or very near to genes that have been implicated in structural development or growth regulation such as Eucgr.J01613 (FDR 0.079; *SPT16*, involved in regulation of vegetative development in *Arabidopsis thaliana*) and Eucgr.J00015 (FDR 0.079; *actin 4*).

Regional heritability mapping

Testing for significant associations between non-overlapping regions of 100 SNPs and phenotypic traits provided results, which overlapped with the single SNP GWAS for terpene related traits (Table 2, Supporting Information Fig. S7, Table S8). We found 80 associated regions with nine traits, with several regions associating to multiple traits for a total of 93 associations (e.g. region 1391 associated with APIN, CIN, and the proportion of 1,8-cineole (PCIN)). Several regions on chromosome 5 associated with biomass related traits (HT, ΔHT). We observed no clear peaks as could be expected for regional QTL; instead, many of the SNPs across this chromosome were significant and should therefore be interpreted with caution.

Multivariate GWAS

Multivariate association of five terpene traits with the GWAS SNP set produced Bayes factors (BFs) for each SNP, representing how much that SNP improved a null model for any combination of the five terpene traits. A total of 197 SNPs had a BF > 4, 48 SNPs had a BF > 5 and 17 SNPs had a BF > 6. By examining the 48 SNPs with BF > 5.0 we can see how regions of significant association affect various terpene traits differently (Figure 5). For

example, three SNPs on chromosome 6 (e.g. Chr06_21496016, unknown function) appear to affect PCIN directly plus several other terpene traits indirectly, but with little effect on CIN. This indicates that the region may affect the production of non-cineole terpenes, thereby altering the proportion of 1,8-cineole in stored terpenes, but not affecting the amount of 1,8-cineole.

The SNPs on chromosome 1 around 31.75 Mbp are located within a cluster of 8 monoterpene synthases and show a strong probability of directly affecting CIN and APIN, and lesser probability of affecting other monoterpenes (MONOREST), PCIN and SESQ. This provides evidence for the presence of 1,8-cineole and/or α-pinene synthases in this region leading us to investigate this region in more detail (Figure 6). Interestingly, SNP Chr01_25455245, which is located within the previously described β-pinene QTL on chromosome 1, appears to have a strong direct effect on APIN and an indirect effect on MONOREST.

Figure 6 reveals a lack of SNPs within most of the terpene synthases in this cluster. The strong LD block around the prominently associated SNPs near TPS058 (31.75 Mbp) also forms a moderately strong LD block with downstream SNPs around TPS061 (31.95 Mbp) and upstream SNPs near TPS054 (31.52 Mbp). It is worth noting that the small region of dense SNPs just after 31.9 Mbp is probably a rearrangement in *E. polybractea* relative to *E. grandis*, so TPS061 could be physically closer to the rest of the TPS group than it appears. It is apparent from the top panel that CIN and APIN are the traits most likely to be directly affected by SNPs throughout this region. APIN is also least likely to be affected indirectly (Supporting Information Fig. S8).

Functional Gene Validation

Terpene synthase genes putatively responsible for 1,8-cineole and α-pinene production were localised to chromosome 1. We amplified two complete terpene synthases, which showed high similarity to one of the genes, EgTPS055, from two individual plants. *Eptps3* was isolated from an individual with a high proportion of 1,8-cineole (DK238; 92% 1,8-cineole) and *Eptps2* was isolated from an individual with a low proportion of 1,8-cineole (DK264; 48% 1,8-cineole). The two gene products have an amino acid sequence identity of 95.1%. Both EpTPS2 and EpTPS3 were heterologously expressed and formed 1,8-cineole as the major product with minor amounts of other terpenes (Figure 7B and 7C). We cloned a third terpene synthase with sequence similarity to EgTPS060 from individual DK264. The corresponding enzyme, EpTPS1 produced β- and α-pinene (Figure 7A).

Discussion

We performed the first genome-wide association for terpene traits using dense markers obtained by whole genome re-sequencing. Candidate gene studies have previously revealed significant associations of terpene concentration with key genes in the MEP and MVA pathways and downstream terpene synthases. Here, we used 2.39 million SNPs to explore the genomic architecture in these candidate genes and all other regions of the genome. The use of

very high density SNPs enabled us to perform GWAS on a fine scale, though the moderately small population size meant that power to detect QTLs of smaller effect was limited (Visscher *et al.*, 2017). Nevertheless, through single-, regional- and multivariate association we pinpointed the location of several commercially important terpene synthases and then functionally validated them. Furthermore, we found evidence for allelic influence on important factors beyond the synthesis of terpenoids, such as precursor availability, terpene transport, cavity formation and storage.

One quarter (20/80) of the significant region associations also contained SNPs that were significantly associated in the single-SNP GWAS (Supporting Information Tables S7, S8). Strongly significant SNPs were one of the main drivers for this overlap (e.g. SNP Chr01_31758048 associated with CIN, PCIN and APIN, was included in region 1392, which associated with the same three traits). Another driver were regions that contained multiple significant SNPs (e.g. region 19021 (TTC and SESQ) contained seven SNPs significantly associated with TTC, SESQ, MONO and MSRATIO). Finally, many significant regions that did not contain individual significant SNPs may provide valuable information about small QTL of small effect that we were unable to detect in our single-SNP GWAS.

Identification of terpene synthases and implications for industry

The significant GWAS associations found in a cluster of *TPS* genes on chromosome 1 led to the successful functional validation of two of those genes, which showed that they produced 1,8-cineole and, to a lesser extent, α-pinene. This was in agreement with the dual effect of these genes shown by multivariate GWAS (Figure 6), as well as previous studies that have shown that cineole synthases in *A. thaliana* and *Salvia officinalis* also produce lesser amounts of pinenes (Wise *et al.*, 1998; Chen *et al.*, 2004). Overall, the successful validation of these *tps* genes provides support for performing GWAS in highly related populations using low depth sequencing, as well as the strength of multivariate approaches.

In the *Eucalyptus* essential oil industry 1,8-cineole is the most important terpene. Growers seek genotypes with high terpene yield that is qualitatively almost entirely 1,8-cineole. Therefore, identifying which genes, out of dozens of putative monoterpene synthases, actually produce 1,8-cineole is beneficial for molecular breeding. Given the identity of cineole synthases, it may be possible to evaluate prospective parental genotypes for copy-number variation in these genes, or engineer individuals with more copies of 1,8-cineole synthases than would be found naturally. On the other hand, if biofuel production via dimerization of non-oxygenated monoterpenes is the goal (Meylemans *et al.*, 2012; Mewalal *et al.*, 2017), it may be possible to knockout 1,8-cineole synthases in high yielding genotypes, while simultaneously boosting the copy-number of pinene synthases to convert the excess of GPP substrate to α- or β-pinene. This process can be guided by the functionally validated pinene synthase (Figure 7A), as well as the strong association between β-pinene and two un-annotated monoterpene synthases at another locus on chromosome 1 (Figure 3). The β-/α-pinene synthase EpTPS1 shows high sequence similarity to EgTPS060, which is 1 Mbp distant from the strong QTL for β-pinene. It is possible that EgTPS060/EpTPS1 is closer to the β-pinene QTL due to structural variation or deletions in *E. polybractea* relative to *E. grandis*.

Terpene transport and storage

The strongest signal of association with TTC and MONO was a cluster of seven UDP-Glycosyltransferase (UDP-GT) genes on chromosome 10 between 10.55 and 10.65 Mbp. Glycosylation of terpenes enables transport and storage of hydrophobic terpenes (Rivas *et al.*, 2013; Schwab *et al.*, 2015). Terpene glycosylation has been shown to occur naturally in many plants, including cultured *Eucalyptus perriniana* cells (Shimoda *et al.*, 2006) and *A. thaliana* (Caputi *et al.*, 2008). AtUGT85A2, which showed strong activity with citronellol and geraniol is the closest *Arabidopsis thaliana* homolog for three of the GTs within the associated region on chromosome 10 (Eucgr.J00971, Eucgr.J00972 and Eucgr.J00973), providing evidence that allelic variants at this locus may affect the glycosylation of monoterpenes, which in turn may have an effect on monoterpene and total terpene concentration. Although eucalypts have secretory cavities for terpene storage, it is possible that a significant proportion of the terpenoids produced by the MVA and MEP pathways are subsequently glycosylated and stored in that form, perhaps as a protective barrier for cells around the perimeter of the secretory cavities (Goodger *et al.*, 2016).

Terpenoid biosynthesis is highly polygenic

The GWAS showed that there were relatively few loci strongly associated with TTC or MONO. This reflects the highly polygenic architecture of this trait and supports findings in Myrtaceae that gene expression and transcript abundance are regulators of variation in terpene concentration (Webb *et al.*, 2013). None of the significant SNPs were located in or near the Candidate Gene set, which is in contrast with previous QTL and SNP association studies in candidate genes which show that a sizeable proportion of quantitative variation in TTC is due to allelic variation within the MEP and/or MVA pathway (Henery *et al.*, 2007; Külheim *et al.*, 2011; O'Reilly-Wapstra *et al.*, 2011; Padovan *et al.*, 2017).

There are several possible explanations for this result. Firstly, associations detected by candidate gene studies are often subject to the 'winners curse' or Beavis effect (Beavis, 1994), which inflates effect sizes. Indeed, simulating a SNP set derived only from the candidate gene set, the genes containing the most significant associations are those previously shown to have an effect on terpene concentration, such as *dxs*, *dxr, hds, gpps, ggpps* as well as several terpene synthases and *hmgr* and *pmd1* from the MVA pathway. In the limited context of just the candidate gene set it is easy to interpret this subset of results, genome-wide they may explain only a small proportion of the heritability.

Secondly, there is a large body of evidence showing that much of the flux through the MVA and MEP pathways is due to a complex network of regulatory mechanisms (Hemmerlin, 2013; Webb *et al.*, 2013; Vickers *et al.*, 2014; Davies *et al.*, 2015). Allelic variants within genes encoding MEP and MVA enzymes could affect the efficiency of these enzymes, but that may be a relatively small signal of variation compared to the effects of transcriptional, post-transcriptional, translational, and post-translational regulation, precursor availability and transport. For example, an individual with a well-optimized terpene biosynthesis pathway

may produce monoterpenes at lower capacity if the availability of G3P or pyruvate in the chloroplast is sub-optimal.

Thirdly, allelic variation within the candidate gene set may indeed cause variation in the total production of terpenes, but this may be masked by the relative ability of an individual plant to store the terpenes in foliar secretory cavities. TTC (and MONO to a lesser extent) is as much a measure of an individual's terpene storage capacity as its capacity to actually synthesise terpenes, and since TTC and total cavity volume are highly correlated (Goodger & Woodrow, 2012). Association for TTC may also reveal genes involved in the formation and maintenance of cavity space in the leaves, which will confound any signal from the candidate gene set. Furthermore, eucalypts emit large amounts of volatile terpenes (Winters *et al.*, 2009; Kanagendran *et al.*, 2018), which may indicate excess production capacity beyond the storage capacity of the secretory cavities, or perhaps synthesis of terpenes in the mesophyll and not exclusively in the secretory cavities (Winters, 2010). A recent investigation of the molecular mechanism of formation of schizogenous intercellular spaces in *Marchantia polymorpha*, found that two proteins were essential for its initiation: a U-box E3 ubiquitin ligase which interacts with a Leucine-rich repeat kinase (Ishizaki, 2015). After the initiation, genes for cell wall remodelling including expansins, cellulases and xyloglucan endotransglucolases/hydrolases are induced (Ishizaki, 2015). Here, we identified a U-box E3 ubiquitin-protein ligase (Eucgr.F00384, Chr06_4956541, FDR = 0.05), a leucine-rich repeat kinase (Eucgr.G02301, Chr07_43651762, FDR = 0.04) as well as an expansin-like gene (Eucgr.E00317, Chr05_2988627, uncorrected *p*-value 1.45 x 10^{-6}, FDR = *ns*) that are all associated with the foliar concentration of terpenes (Table 2). While the evolutionary distance between a liverworth and eucalypts is large, the mechanism of initiation of secretory cavity formation may be convergent (Ishizaki, 2015).

Precursor availability

The second strongest single-SNP and multivariate associations were between SESQ and SNPs in the *phosphoenolpyruvate/phosphate translocator* (PPT2) gene (Eucgr.J02222, Chr10_27384836, FDR = 1.6 x 10^{-9}). This translocator resides in the plastid inner-envelope membrane and is responsible for importing phosphoenolpyruvate (PEP) from the cytosol into the chloroplast (Knappe *et al.*, 2003; Linka & Weber, 2010). PEP is the direct precursor to the shikimate pathway and may also be converted to pyruvate via pyruvate kinase, though this mechanism has not been demonstrated yet in plants (Banerjee & Sharkey, 2014). As both G3P and pyruvate are required by the MEP pathway for IPP production, the import of PEP may provide a secondary source of pyruvate that contributes to a high rate of terpene production in mature leaf when volatile isoprene emissions are greatest (Banerjee & Sharkey, 2014). This is supported by studies of *Arabidopsis thaliana cue1* mutants with defective PPT2, which produce only one third the carotenoids, one quarter the chlorophylls, have vastly reduced flux through the shikimate pathway, yet show no reduction in fatty acid biosynthesis (Streatfield *et al.*, 1999).

While there is evidence that PPT2 plays a role in terpenoid biosynthesis in the chloroplast, it is difficult to explain why Eucgr.J02222 is directly associated with SESQ, but not with terpene traits such as MONO and CIN that are perhaps more dependent on MEP precursor availability. We hypothesise that while a loss of function in PPT2 (as in the *cue1* mutants) results in reduced availability of precursors to both the MEP and shikimate pathways, over-expression of PPT2 may not result in greater production of monoterpenes if flux through the MEP pathway is already at maximum. Loreto *et al.* (2007) showed evidence that cytosolic PEP is partitioned between respiration, anabolic metabolism and chloroplast import. Increased transport of PEP from the cytosol to the chloroplast may reduce production of cytosolic acetyl-CoA, depriving the MVA pathway of its precursor and therefore reducing the production of sesquiterpenes.

Conclusion

In this study, we used over 2 million markers obtained from whole genome re-sequencing of 480 *E. polybractea* individuals to explore the genomic architecture of terpene and biomass related traits. We showed that variation in monoterpene, sesquiterpene and total terpene concentration may be influenced by many loci of small effect beyond the genes of their biosynthetic pathways, with few loci of large effect. We observed significant effects of the chloroplast PEP translocator (PPT2), and evidence from associations with multiple UDP-glycosyltransferase genes that glycosylation of terpenes may influence the accumulation of extractable terpenes. Finally, terpene synthases that are responsible for 1,8-cineole and α-pinene production are localised in a cluster of *tps* genes on chromosome 1. These results provide a new list of candidate genes including several for formation of secretory cavities that warrant further investigation to provide additional avenues beyond the MEP and MVA pathways for improvement of terpene production in Myrtaceae.

Acknowledgements

This work was supported by the Australian Research Council (DP140101755) and funding by Rural Industries Research Development Corporation, Australia. The Center for BioEnergy Innovation is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE office of Science. We wish to acknowledge Richard Davis and GR Davis Pty Ltd for granting access to the *E. polybractea* progeny trial. This project received assistance from the National Computational Infrastructure, which is supported by the Australian Government.

Author contributions

CK and WJF planned and designed the research. CK and DK performed field sampling and laboratory work with the exception of tps functional characterisation. DK developed the bioinformatics pipeline and analysed the data. AP, CK and DK evaluated gene function. SK, PM and JD cloned tps genes and functionally characterised them. DK and CK wrote the

manuscript and all authors provided comments on drafts of the manuscript and approved the final manuscript.

References

Baker GR, Doran JC, Williams ER, Morris G 2014. Highly improved tea tree varieties to maximise profit. Barton, ACT Australia: Rural Industries Research and Development Corporation. 91.

Banerjee A, Sharkey TD. 2014. Methylerythritol 4-phosphate (MEP) pathway metabolic regulation. *Natural Product Reports* **31**(8): 1043-1055.

Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion J-M. 2015. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* **206**(4): 1283-1296.

Barton AFM. 1999. The oil mallee project. *Journal of Industrial Ecology* **3**(2-3): 161-176.

Beavis WD 1994. The power and deceit of QTL experiments: lessons from comparative QTL studies. *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference*. 250-266.

Bohlmann J, Keeling CI. 2008. Terpenoid biomaterials. *The Plant journal : for cell and molecular biology* **54**(4): 656-669.

Bushnell B 2016. BBTools. sourceforge.net/projects/bbmap/ accessed May 2016

Butcher PA, Doran JC, Slee MU. 1994. Intraspecific variation in leaf oils of *Melaleuca alternifolia* (Myrtaceae). *Biochemical Systematics and Ecology* **22**(4): 419-430.

Cappa EP, El-Kassaby YA, Garcia MN, Acuña C, Borralho NMG, Grattapaglia D, Marcucci Poltri SN. 2013. Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. *PLOS ONE* **8**(11): e81267.

Caputi L, Lim EK, Bowles DJ. 2008. Discovery of new biocatalysts for the glycosylation of terpenoid scaffolds. *Chemistry - A European Journal* **14**(22): 6656-6662.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**(1): 7.

Chen F, Ro DK, Petri J, Gershenzon J, Bohlmann J, Pichersky E, Tholl D. 2004. Characterization of a root-specific *Arabidopsis* terpene synthase responsible for the formation of the volatile monoterpene 1,8-cineole. *Plant Physiol* **135**: 1956 - 1966.

Conomos MP, Thornton T 2016. GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package 2.4.0.

Davies FK, Jinkerson RE, Posewitz MC. 2015. Toward a photosynthetic microbial platform for terpenoid engineering. *Photosynthesis research* **123**(3): 265-284.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**(5): 491-498.

Dodt M, Roehr JT, Ahmed R, Dieterich C. 2012. FLEXBAR—Flexible Barcode and Adapter Processing for next-generation sequencing platforms. *Biology* **1**(3): 895-905.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics* **46**(10): 1089-1096.

Garrison E, Marth G. 2010. Haplotype-based variant detection from short-read sequencing. *arXiv* **1207.3907**.

González-Orozco CE, Pollock Laura J, Thornhill Andrew H, Mishler Brent D, Knerr N, Laffan Shawn W, Miller Joseph T, Rosauer Dan F, Faith Daniel P, Nipperess

David A, et al. 2016. Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**: 1110.

Goodger JQD, Seneratne SL, Nicolle D, Woodrow IE. 2016. Foliar Essential Oil Glands of *Eucalyptus* Subgenus Eucalyptus (Myrtaceae) Are a Rich Source of Flavonoids and Related Non-Volatile Constituents. *PLOS ONE* **11**(3): e0151432.

Goodger JQD, Woodrow IE. 2012. Genetic determinants of oil yield in *Eucalyptus polybractea* R.T. Baker. *Trees* **26**(6): 1951-1956.

Hamberger B, Bohlmann J. 2006. Cytochrome P450 mono-oxygenases in conifer genomes: discovery of members of the terpenoid oxygenase superfamily in spruce and pine. *Biochem Soc Trans* **34**(Pt 6): 1209-1214.

Hemmerlin A. 2013. Post-translational events and modifications regulating plant enzymes involved in isoprenoid precursor biosynthesis. *Plant science : an international journal of experimental plant biology* **203-204**: 41-54.

Henery ML, Moran GF, Wallis IR, Foley WJ. 2007. Identification of quantitative trait loci influencing foliar concentrations of terpenes and formylated phloroglucinol compounds in *Eucalyptus nitens*. *New Phytologist* **176**(1): 82-95.

Ishizaki K. 2015. Development of schizogenous intercellular spaces in plants. *Frontiers in Plant Science* **6**: 497.

Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research* **25**(6): 918-925.

Kainer D, Bush D, Foley WJ, Külheim C. 2017. Assessment of a non-destructive method to predict oil yield in *Eucalyptus polybractea* (blue mallee). *Industrial Crops and Products* **102**: 32-44.

Kainer D, Stone EA, Padovan A, Foley WJ, Külheim C. 2018. Accuracy of Genomic Prediction for Foliar Terpene Traits in *Eucalyptus polybractea*. *G3: Genes|Genomes|Genetics* **8**(8): 2573-2583.

Kanagendran A, Pazouki L, Bichele R, Külheim C, Niinemets Ü. 2018. Temporal regulation of terpene synthase gene expression in *Eucalyptus globulus* leaves upon ozone and wounding stresses: relationships with stomatal ozone uptake and emission responses. *Environmental and Experimental Botany* **155**: 552-565.

Keefover-Ring K, Thompson JD, Linhart YB. 2009. Beyond six scents: defining a seventh *Thymus vulgaris* chemotype new to southern France by ethanol extraction. *Flavour and Fragrance Journal* **24**(3): 117-122.

Keszei A, Brubaker CL, Foley WJ. 2008. A molecular perspective on terpene variation in Australian Myrtaceae. *Australian Journal of Botany* **56**(3): 197-213.

King DJ, Gleadow RM, Woodrow IE. 2004. Terpene deployment in *Eucalyptus polybractea*; relationships with leaf structure, environmental stresses, and growth. *Functional Plant Biology* **31**(5): 451-460.

King DJ, Gleadow RM, Woodrow IE. 2006. The accumulation of terpenoid oils does not incur a growth cost in *Eucalyptus polybractea* seedlings. *Functional Plant Biology* **33**(5): 497-505.

Knappe S, Lottgert T, Schneider A, Voll L, Flügge U-I, Fischer K. 2003. Characterization of two functional phosphoenolpyruvate/phosphate translocator (PPT) genes in *Arabidopsis*-AtPPT1 may be involved in the provision of signals for correct mesophyll development. *The Plant journal : for cell and molecular biology* **36**(3): 411-420.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Külheim C, Padovan A, Hefer CA, Krause ST, Koellner T, Myburg AA, Degenhardt J, Foley WJ. 2015. The *Eucalyptus* terpene synthase gene family. *BMC genomics* **16**: 450.

Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF. 2009. Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC genomics* **10**: 452.

Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF, Foley WJ. 2011. The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* **191**: 1041-1053.

Lange BM, Rujan T, Martin W, Croteau RB. 2000. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences* **97**(24): 13172-13177.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research* **21**(6): 940-951.

Lichtenthaler HK. 1999. The 1-deoxy-d-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annual review of plant physiology and plant molecular biology* **50**: 47-65.

Linka N, Weber APM. 2010. Intracellular metabolite transporters in plants. *Molecular Plant* **3**(1): 21-53.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**: 720-731.

Loreto F, Centritto M, Barta C, Calfapietra C, Fares S, Monson RK. 2007. The relationship between isoprene emission rate and dark respiration rate in white poplar (*Populus alba* L.) leaves. *Plant, Cell & Environment* **30**(5): 662-669.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. 2011. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics & Genomes* **7**(5): 1011-1023.

McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J, et al. 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist* **203**(2): 535-553.

Mewalal R, Rai DK, Kainer D, Chen F, Külheim C, Peter GF, Tuskan GA. 2017. Plant-Derived Terpenes: A Feedstock for Specialty Biofuels. *Trends in biotechnology* **35**(3): 227-240.

Meylemans HA, Quintana RL, Harvey BG. 2012. Efficient conversion of pure and mixed terpene feedstocks to high density fuels. *Fuel* **97**: 560-568.

Müller BSF, de Almeida Filho JE, Lima BM, Garcia CC, Missiaggia A, Aguiar AM, Takahashi E, Kirst M, Gezan SA, Silva-Junior OB, et al. 2019. Independent and Joint-GWAS for growth traits in *Eucalyptus* by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytologist* **221**(2): 818-833.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. 2014. The genome of *Eucalyptus grandis*. *Nature* **510**(7505): 356-362.

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 Genetics & Genomes* **7**(3): 485-498.

Padovan A, Keszei A, Külheim C, Foley W. 2014. The evolution of foliar terpene diversity in Myrtaceae. *Phytochemistry Reviews* **13**(3): 695-716.

Padovan A, Webb H, Mazanec R, Grayling P, Bartle J, Foley WJ, Külheim C. 2017. Association genetics of essential oil traits in *Eucalyptus loxophleba*: explaining variation in oil yield. *Molecular Breeding* **37**(6): 73.

Renaud ENC, Charles DJ, Simon JE. 2001. Essential Oil Quantity and Composition from 10 Cultivars of Organically Grown Lavender and Lavandin. *Journal of Essential Oil Research* **13**(4): 269-273.

Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D. 2017. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytologist* **213**: 1287-1300.

Rivas F, Parra A, Martinez A, Garcia-Granados A. 2013. Enzymatic glycosylation of terpenoids. *Phytochemistry Reviews* **12**(2): 327-339.

Schwab W, Fischer T, Wüst M. 2015. Terpene glucoside production: Improved biocatalytic processes using glycosyltransferases. *Engineering in Life Sciences* **15**(4): 376-386.

Shimoda K, Kondo Y, Nishida T, Hamada H, Nakajima N, Hamada H. 2006. Biotransformation of thymol, carvacrol, and eugenol by cultured cells of *Eucalyptus perriniana*. *Phytochemistry* **67**(20): 2256-2261.

Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytologist* **206**(4): 1527-1540.

Stephens M. 2013. A unified framework for association analysis with multiple related phenotypes. *PLOS ONE* **8**(7): e65245.

Streatfield SJ, Weber APM, Kinsman EA, Häusler RE, Li J, Post-Beittenmiller D, Kaiser WM, Pyke KA, Flügge U-I, Chory J. 1999. The phosphoenolpyruvate/phosphate translocator is required for phenolic metabolism, palisade cell development, and plastid-dependent nuclear gene expression. *The Plant Cell* **11**(9): 1609-1621.

Thavamanikumar S, Southerton SG, Bossinger G, Thumma BR. 2013. Dissection of complex traits in forest trees — opportunities for marker-assisted selection. *Tree Genetics & Genomes* **9**(3): 627-639.

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

Uchiyama K, Iwata H, Moriguchi Y, Ujino-Ihara T, Ueno S, Taguchi Y, Tsubomura M, Mishima K, Iki T, Watanabe A, et al. 2013. Demonstration of genome-wide association studies for identifying markers for wood property and male strobili traits in *Cryptomeria japonica*. *PLOS ONE* **8**(11): e79866.

Vickers CE, Bongers M, Liu Q, Delatte T, Bouwmeester H. 2014. Metabolic engineering of volatile isoprenoids in plants and microbes. *Plant, Cell & Environment* **37**(8): 1753-1775.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**(1): 5-22.

Vranova E, Coman D, Gruissem W. 2013. Network Analysis of the MVA and MEP Pathways for Isoprenoid Synthesis. *Annual Review of Plant Biology* **64**: 665-700.

Webb H, Foley WJ, Külheim C. 2014. The genetic basis of foliar terpene yield: Implications for breeding and profitability of Australian essential oil crops. *Plant Biotechnology* **31**(5): 363-376.

Webb H, Lanfear R, Hamill J, Foley WJ, Külheim C. 2013. The Yield of Essential Oils in *Melaleuca alternifolia* (Myrtaceae) Is Regulated through Transcript Abundance of Genes in the MEP Pathway. *PLOS ONE* **8**(3): e60631.

Winters AJ. 2010. *The composition and emission of volatile organic compounds in* Eucalyptus *spp. leaves.*Thesis M4 - Citavi, Faculty of Science Sydney.

Winters AJ, Adams MA, Bleby TM, Rennenberg H, Steigner D, Steinbrecher R, Kreuzwieser J. 2009. Emissions of isoprene, monoterpene and short-chained carbonyl compounds from *Eucalyptus* spp. in southern Australia. *Atmospheric Environment* **43**: 3035-3043

Wise ML, Savage TJ, Katahira E, Croteau R. 1998. Monoterpene Synthases from Common Sage (Salvia officinalis) : cDNA isolation, characterization, and functional expression of (+)-sabinene synthase, 1,8-cineole synthase, and (+)-bornyl diphosphate synthase. *Journal of Biological Chemistry* **273**(24): 14891-14899.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**(1): 76-82.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie CC, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**(24): 3326-3328.

Figure legend

Figure 1: Schematic overview of mono- and sequiterpene biosynthesis. Genes are shown with gene copy number based on the genome of *Eucalyptus grandis*. Examples of possible terpene modifications are shown. Terpenes are in bold type.

Figure 2: Manhattan plot of β-pinene (BPIN) concentration. Only the single nucleotide polymorphisms within the highly significant QTL on chromosome 1 of *Eucalyptus polybractea* passed the genome-wide significance threshold denoted by the red horizontal line. The blue horizontal line denotes an arbitrary significance threshold of $p = 10^{-5}$.

Figure 3: Regional Manhattan plot and BLASTX track of quantitative trait loci for β-pinene concentration (BPIN). **a)** Manhattan plot of single nucleotide polymorphism associations (SNPs) with BPIN on chromosome 1 (25,440,000 – 25,520,000 bp) of *Eucalyptus polybractea*. The horizontal dotted line indicates the Bonferroni threshold of FDR = 0.1. Black circled dots are SNPs at FDR < 0.1 with Benjamini-Hochberg method on a per chromosome basis; **b)** The BLASTX track against *Arabidopsis thaliana* from Phytozome. Annotation of BLASTX hits: 1 and 2 mono terpene synthase (At2g24210, At3g25810, At3g25820, At4g16730 and At4g16740).

Figure 4: **Manhattan plots of genome-wide associations in six terpene traits in *Eucalyptus polybractea***. The x-axis represents the genomic position of each SNP, the y-axes shows the strength of association measured as the $-\log_{10}(p\text{-value})$ for each single nucleotide polymorphism (SNP) and trait. Panels are APIN: α-pinene concentration, CIN: 1,8-cineole concentration, MONO: monoterpene concentration, PCIN: proportion of 1,8-cineole, SESQ: sesquiterpene concentration, TTC: total terpene concentration. The dashed horizontal line shows the highly conservative genome-wide Bonferroni threshold. SNPs with a Benjamini Hochberg corrected FDR < 0.10 at the genome-wide level are coloured in orange, while SNPs with a FDR < 0.10 at the less stringent chromosome level are shown in blue. Vertical orange lines are drawn at SNPs Chr01_31758048, which is located within a cluster of terpene synthase genes, and Chr10_10603390, which is located in a cluster of UDP-glycosyltransferases.

Figure 5: Heatmap of most significant multivariate associated single nucleotide polymorphisms (SNPs) in *Eucalyptus polybractea* (BF > 5). Color indicates the SNP having a direct effect on the trait (PrD), the posterior probability of the SNP having any effect on the trait (PrA), or an indirect effect on the trait (PrI). Red indicates higher probability, blue indicates lower. SNPs (rows) are ordered by genomic position. The green shaded annotations on the left indicate the multivariate BF for each SNP.

Figure 6: Multivariate association within cluster of seven terpene synthase genes on chromosome 1 in *Eucalyptus polybractea*. a) posterior probability that a given single nucleotide polymorphism (SNP) has a direct effect on one of five terpene traits, where brighter yellow means greatest probability and darker purple means least. **b)** Manhattan plot of Bayes factors for multiple terpene traits. Points are sized according to their Bayes factor and coloured according to their pairwise linkage disequilibrium (LD) with the most significant association at 31.75 Mbp (labelled). GWAS, genome-wide association study. **c)** Pairwise LD between each SNP in the 500 kb region, where lighter colour signifies stronger LD.

Figure 7: Terpene products of the heterologously expressed enzymes from *Eucalyptus polybractea*. a) EpTPS1, b) EpTPS2 and c) EpTPS3. The compounds were collected with solid phase micro extraction and analysed by gas chromatography-mass spectroscopy. They were identified as 1. α-pinene, 2. sabinene, 3. β-pinene, 4. myrcene, 5. limonene, 6. 1,8-cineole, and 7. α-terpineol.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Primers used in this study.

Table S2 Next generation sequencing quality statistics.

Table S3 Reference alignment statistics per sample.

Table S4 Single nucleotide polymorphism filtering statistics.

Table S5 123 Terpene biosynthesis related genes with their location in the *Eucalyptus grandis* genome, gene annotation and gene model.

Table S6 Heritability estimates for each trait based on the Genesis null models for single SNP GWAS.

Table S7 Complete table of all SNPs that passed the Benjamini Hochberg FDR alpha = 0.1 for each trait, including up to three of the closest *E. grandis* gene models.

Table S8 Regions that were significantly associated with phenotypic traits with an FDR significance threshold of 0.1 on a per chromosome basis.

Figure S1 Maternal seed sources for *Eucalyptus polybractea* progeny trial.

Figure S2 Sequencing and genotype depth for shallow whole genome sequencing of *Eucalyptus polybractea*.

Figure S3 Minor allele frequency distribution of 2.39 million SNPs across the population of 468 *Eucalyptus polybractea* individuals.

Figure S4 Principal Component Analysis of 468 individuals of *Eucalyptus polybractea* labelled by family.

Figure S5 Linkage disequilibrium decay from 468 individuals of *Eucalyptus polybractea*.

Figure S6 Tests for trait-effect size between single nucleotide polymorphisms from random genes compared to the candidate gene set.

Figure S7 Genome-wide regional heritability mapping manhattan plots for eight oil yield related traits.

Figure S8 Probability of single nucleotide polymorphisms indirect effect on oil traits from multivariate associations of a terpene synthase rich region on chromosome 1.

Method S1 Detailed information about the methods and models used for single-SNP GWAS, regional heritability mapping and multivariate GWAS.

Method S2 Detailed information about the methods for functional characterisation of terpene synthase genes.

Tables

Table 1 – *p*-values of single nucleotide polymorphisms (SNPs) located within genes and mean absolute SNP effect size |EFF| in *Eucalyptus polybractea*.

	WG		**CAND**		**GENES**	**CAND**	
threshold	FDR 0.1	FDR 0.2	FDR 0.1	FDR 0.2	|EFF|	|EFF|	t-test
Oil							
TTC	2	9	0	0	1.703	1.707	0.429
MONO	0	0	0	0	1.554	1.556	0.460
CIN	0	2	0	0	1.230	1.260	0.046
PCIN	1	2	0	0	0.014	0.014	0.910
APIN	2	2	0	0	0.097	0.099	0.120
BPIN	0	5	0	0	0.153	0.156	0.057
SESQ	9	16	0	0	0.969	0.987	0.079
MSratio*	905	8425	2	24	0.150	0.151	0.210
Biomass							
LA	0	2	0	0	26.357	25.059	1.000
HT	0	0	0	0	6.779	6.622	0.967
ΔHT	0	0	0	0	3.224	3.165	0.923
ΔCA	0	0	0	0	0.032	0.032	0.373

The number of SNPs with *p*-values below the given thresholds within the whole genome (WG) or within oil candidate genes (CAND) are shown (left side of the table). Mean SNP effects size is shown in all genes (GENES) and candidate genes (CAND; right side of the table). T-test reports the *p*-value from a 1-sided t-test to see if estimated SNP effects are greater in candidate genes than in all genes. *MSratio showed signs of *p*-value inflation.

Table 2 – Selected single nucleotide polymorphisms (SNPs) associations by trait in *Eucalyptus polybractea*.

Traits	SNP ID	FDR	MAF	R^2	Nearest gene	Annotation	Possible function
BPIN	Chr01_25455245 *	7.74E-43	0.309	0.338	NA	Putative monoterpene synthases	TERP
SESQ	Chr10_27384836 *	1.56E-9	0.473	0.114	Eucgr.J02222	phosphoenolpyruvate (pep)/phosphate translocator 2	PRE
APIN, CIN, PCIN	Chr01_31758048 *	2.54E-5	0.258	0.081	Eucgr.TPS058	Monoterpene synthase cluster	TERP
SESQ, TTC, MONO	Chr10_19020945 *	0.0007	0.056	0.064	Eucgr.J01534	SAM-Mtase superfamily protein	STRESS
TTC, MONO	Chr02_38232403 *	0.008	0.056	0.058	Eucgr.B01785	Disease resistance protein (CC-NBS-LRR class) family	STRESS
TTC, MONO	Chr10_10603390*	0.0124	0.079	0.059	Eucgr.J00973	UDP-Glycosyltransferase superfamily protein	TRAN, STOR
SESQ	Chr06_45031390	0.019	0.029	0.058	Eucgr.F03370	S-locus lectin protein kinase family protein	STRESS
TTC, MONO	Chr02_52005554	0.032	0.045	0.050	Eucgr.B03157	plasmodesmata callose-binding protein 3	TRAN
SESQ	Chr05_13544393	0.038	0.054	0.055	Eucgr.E01295	NB-ARC domain-containing disease resistance protein	STRESS
TTC	Chr07_43651762	0.039	0.121	0.054	Eucgr.G02301	leucine-rich repeat transmembrane protein kinase family protein	CAV
TTC, APIN	Chr07_13026919 *	0.039	0.051	0.052	Eucgr.G00843	HXXXD-type acyl-transferase family protein	STOR
MONO	Chr06_4956541	0.052	0.120	0.054	Eucgr.F00384	RING/U-box E3 ubiquitin-protein ligase	CAV
MONO	Chr10_19980843	0.054	0.067	0.046	Eucgr.J01589	Calcium-dependent lipid-binding (CaLB domain) family protein	TRAN
APIN	Chr07_37694713	0.054	0.050	0.048	Eucgr.G01944	translocase inner membrane subunit 44-2	TRAN

ΔHT	Chr10_20285569 *	0.079	0.041	0.050	Eucgr.J01613	GLOBAL TRANSCRIPTION FACTOR C, SPT16	GROW
ΔHT	Chr10_236988	0.080	0.052	0.049	Eucgr.J00015	actin-4	GROW
SESQ	Chr05_27914866 *	0.083	0.037	0.049	Eucgr.E02010	Vacuolar sorting protein 39	TRAN
CIN	Chr02_55962524	0.099	0.036	0.049	Eucgr.B03623	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	TRAN

SNP IDs with * indicate that there are multiple SNPs at similar significance levels near this locus. Nearest gene shown from *E. grandis*. Annotation: Annotation for *Eucalyptus grandis* genes were taken from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) with the exception of Chr01_25455245, which was annotated based on the BLASTX track of phytozome. Abbreviations: TRAN (transport), STOR (storage), STRESS (stress response), CAV (cavity formation or cavity structure), PRE (precursor availability), TERP (terpenoid biosynthesis), GROW (growth regulation).