Genomic patterns of species diversity and divergence in *Eucalyptus*

Corey J Hudson[1], Jules S Freeman[1,2,3], Alexander A Myburg[4], Brad M Potts[1,2], and René E Vaillancourt[1,2]

[1]School of Biological Sciences, University of Tasmania, Private Bag 55 Hobart, Tasmania 7001, Australia.
[2]National Centre for Future Forest Industries, University of Tasmania, Private Bag 55, Hobart, Tas., 7001, Australia
[3]Faculty of Science, Health, Education and Engineering, and Collaborative Research Network, University of the Sunshine Coast, Locked Bag 4, Maroochydore, Qld., 4558, Australia
[4]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

Corresponding author: Jules S Freeman
Email: Jules.Freeman@utas.edu.au
Phone: +61 3 6226 1828; Fax: +61 3 6226 2698

Word counts
Main body: 6498
Introduction: 876
Materials and Methods: 1931
Results: 2214
Discussion: 1355
Acknowledgements: 122
Number of Figures: 5
Number of Tables: 3
Supporting Information: Tables: 7, Figures: 6, and Notes: 1

Summary

- We examined genome-wide patterns of DNA sequence diversity and divergence among six species of the important tree genus *Eucalyptus* and investigate their relationship with genomic architecture.
- Using ~90 range-wide individuals of each species (*E. grandis, E. urophylla, E. globulus, E. nitens, E. dunnii* and *E. camaldulensis*), genetic diversity and divergence were estimated from 2,840 polymorphic Diversity Arrays Technology markers covering the 11 chromosomes. Species differentiating markers (SDMs) identified in each of 15 pair-wise species comparisons, along with species diversity (H_{HW}) and divergence (F_{ST}) were projected onto the *E. grandis* reference genome.
- Across all species comparisons, SDMs totalled 1.1-5.3% of markers and were widely distributed throughout the genome. Marker divergence (F_{ST} and SDMs) and diversity differed among and within chromosomes. Patterns of diversity and divergence were broadly conserved across species and significantly associated with genomic features including the proximity of markers to genes, the relative number of clusters of tandem duplications, and gene density within or among chromosomes.
- These results suggest genomic architecture influences patterns of species diversity and divergence in the genus. This influence is evident across the six species, encompassing diverse phylogenetic lineages, geography and ecology.

Key words:

Eucalyptus, genomic architecture, population genomics, genome-scan, outlier, gene density, tandem gene duplication, speciation

Introduction

Elucidating the genetic basis of population and species divergence represents a central goal of evolutionary biology (Coyne & Orr, 2004; Storz, 2005; Seehausen *et al.*, 2014; Cruickshank & Hahn 2014). Genomic analysis of large, representative sets of individuals from multiple populations can help unravel the genetic architecture of adaptive divergence by identifying genomic regions that are under selection (Storz, 2005; Oleksyk *et al.*, 2010; Pritchard *et al.*, 2010; Ellegren, 2014). Such analyses rely on the concept that selection distorts patterns of neutral variation throughout the genome in predictable ways and these patterns can be detected through genome-wide analysis (Pritchard *et al.*, 2010). For example, a standard neutral model predicts that mutation, genetic drift and migration affect all loci equally, whereas natural selection affects specific loci and will thus result in detectable signatures of selection (Achere *et al.*, 2005). Loci subject to contrasting selection pressures in different populations and/or species (i.e. divergent selection) can be identified using outlier tests, which identify loci having higher levels of divergence (e.g. marker F_{ST}) relative to neutral, background levels (Scotti-Saintagne *et al.*, 2004; Holliday *et al.*, 2012; Steane *et al.*, 2014). These outlier markers can then be positioned on a reference genome sequence or linkage map to identify genomic regions affected by selection and candidate genes potentially involved in adaptation (Pannell & Fields, 2014). To provide insights into the functional nature of genes influenced by selection, enrichment analyses of gene ontology (GO) classes at the genome level is increasingly used to complement population genomic analyses (Harr 2006; Turner *et al*., 2008; Eckert *et al*., 2010).

Advances in DNA sequencing and array technology coupled with declining genotyping costs now make it feasible to generate the large number of molecular markers distributed across the genome required for population genomics studies in many organisms (Luikart *et al.*, 2003; Pannell & Fields, 2014; Seehausen *et al.*, 2014). We here apply a population genomic approach to study species diversity and divergence in *Eucalyptus*, a southern hemisphere tree genus with a unique evolutionary history (Grattapaglia *et al.*, 2012; Myburg *et al.*, 2014) compared to most species previously studied.

The genus *Eucalyptus* contains over 700 species, including some of the most widely planted hardwood species in the world (Doughty, 2000). The worldwide eucalypt plantation estate has been estimated at 20 million hectares (Iglesias-Trabado & Wilstermann, 2008)

comprising mostly species of the subgenus *Symphyomyrtus* (Grattapaglia & Kirst, 2008). Here we study six economically important *Symphyomyrtus* species from sections (after Brooker 2000) *Latoangulatae* (*E. grandis* Hill ex Maiden and *E. urophylla* S. T. Blake), *Maidenaria* (*E. globulus* Labill., *E. nitens* Deane and Maiden, Maiden and *E. dunnii* Maiden) and *Exsertaria* (*E. camaldulensis* Dehnh.). In their natural environments, these species extend across an extraordinary range of habitats, from temperate regions of south-eastern Australia to tropical Indonesia (Figure 1) and from near sea level up to almost 3,000 m elevation (Euclid, 2006). Of the six species examined, only *E. dunnii* and *E. grandis* have some overlap in their native range. Despite this, no hybrids have been reported for this species pair or any of the six studied species in the wild (Griffin *et al.*, 1988); suggesting the presence of strong reproductive barriers despite natural and artificial hybridisation being commonly reported in eucalypts. Intraspecific genetic population structure has also been well studied in *Eucalyptus*. In general, widely dispersed species with few disjunctions tend to have low levels of structure, while those with small and/or disjunct populations tend to exhibit greater structure (Grattapaglia *et al.*, 2012).

Population genomics provides a framework to examine the molecular basis of adaptation and speciation in these ecologically diverse and economically important eucalypt species. Such studies are especially relevant to the understanding of speciation which occurs in the absence of polyploidy or major chromosomal re-arrangements (Coyne & Orr, 2004). *Eucalyptus* is a good model for studying such speciation as all species have the same haploid chromosome number (n=11; Grattapaglia *et al.*, 2012) and, while there is some variation in genome size, evidence to date suggests high genome synteny and colinearity within subgenus *Symphyomyrtus* (Myburg *et al.*, 2004; Grattapaglia *et al.*, 2012; Hudson *et al.*, 2012). We utilise (i) the reference *E. grandis* genome sequence (Myburg *et al.*, 2014) and (ii) a DNA marker array for *Eucalyptus* which provides species-transferrable markers with wide genome coverage, the majority of which are annotated on the genome sequence (Sansaloni *et al.*, 2010; Steane *et al.*, 2011; Petroli *et al.*, 2012).

Using range-wide samples, genome-wide scans were conducted to examine the genetic diversity within and among the six eucalypt species and to identify genomic regions which differentiated species. The projection of these parameters onto the 11 chromosome assemblies of *E. grandis* provides the first insights into the genomic patterns of species

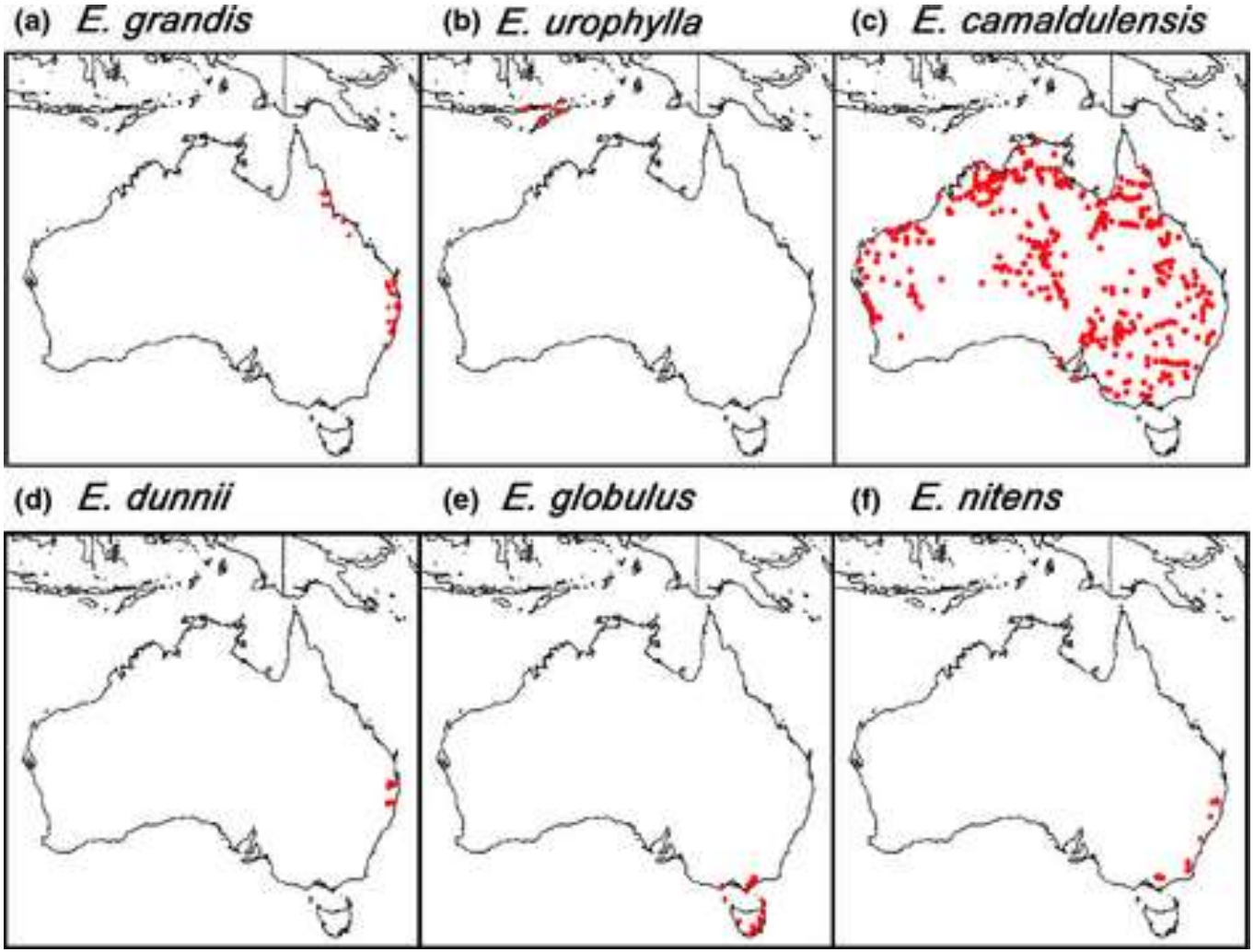

Figure 1. Native distribution of the six Eucalyptus species included in this study: (a) E. grandis, (b) E. urophylla, (c) E. camaldulensis, (d) E. dunnii, (e) E. globulus, and (f) E. nitens. Samples were derived from throughout the geographic range of all species except E. nitens (where populations in the far north were not sampled) and E. urophylla (where populations in the east of East Timor were not sampled; see Payn et al., [1]).

diversity and differentiation in the genus. While the importance of biogeographical and/or population histories in interpreting genomic patterns of diversity and species divergence is often emphasised (Nosil & Feder, 2012; Strasburg *et al.*, 2012), the influence of genomic architecture (such as the distribution of genes and other functional elements; variation in recombination rate; chromosome structure, number and length) is increasingly being recognised (e.g. Flowers *et al.*, 2012; Renaut *et al.*, 2013; Slotte 2014). Here we examine the relationship between genomic architecture and patterns of diversity and divergence amongst these six eucalypt species.

Materials and Methods

Samples

A total of 84-93 samples for each of the six eucalypt species were included in this study (Table 1). These range-wide samples were selected based on the species natural distributions, in addition to consideration of subspecies (*E. camaldulensis*) and/or racial classification (*E. globulus*) to encompass the range of genetic diversity within species. Leaf samples from individuals were collected from either natural populations (*E. globulus*and *E. camaldulensis*) or commercial field trials established directly from individual-tree open-pollinated seedlots collected from natural stands (*E. nitens, E. urophylla, E. grandis* and *E. dunnii* [except seven samples of *E. dunnii*]). Sample locality and source information is provided in Supporting Information Table S1. *E. camaldulensis* genomic DNA was isolated by Butcher *et al*. (2009). *Eucalyptus urophylla, E. grandis* and *E. dunnii* samples were isolated using Qiagen Plant DNeasy kits (Payn *et al.,* 2008). A modified CTAB extraction protocol (McKinnon *et al.*, 2004) was used for the remaining samples.

Genotyping

Samples were genotyped with a Diversity Arrays Technology (DArT) microarray containing 7,680 dominant markers, developed originally from 64 phylogenetically diverse eucalypt species (Sansaloni *et al.*, 2010; Steane *et al.*, 2011). Individuals were assayed by DArT Pty Ltd (Canberra, ACT). The eucalypt DArT array contains a substantial proportion of redundant markers (34-44% based on sequence similarity analyses using different assembly parameters; see Petroli *et al.*, 2012). Two methods were used to identify, and remove, redundant markers from our dataset. First, results from a sequence-based redundancy analysis of eucalypt DArT markers (Petroli *et al.*, 2012; equivalent to A3 stringency) were used to assign markers to unique- or multi-marker bins. From each bin, only the highest quality marker (e.g. with least missing data and highest reproducibility score) was retained. Following this, non-redundant markers were combined with an additional 424 markers not included in the redundancy analysis of Petroli *et al*. (2012). Hamming distances were then calculated in GenAlEx (Peakall & Smouse, 2006) using the marker genotype scores of individuals to identify additional redundant markers. Marker-pairs having identical (or near-identical) genotype scores were inspected with the best quality marker being retained.

Table 1 Number of samples included in this study for each of six *Eucalyptus* species. The number of individual localities and regions / races / subspecies sampled for each species and the state distribution of individuals is also shown.

Species	*N*	Localities	Regions / Races / Subspecies	State distribution of samples[f]
E. camaldulensis	92	88	7[a]	VIC, SA, WA, NT, NSW, QLD.
E. dunnii	89	21	N/A	NSW, QLD.
E. globulus	84	48	13[b]	VIC, TAS.
E. grandis	90	27	4[c]	NSW, QLD.
E. nitens	85	40	7[d]	VIC, NSW.
E. urophylla	93	45	7[e]	N/A

[a]Number of subspecies. [b]Number of races following Dutkowski & Potts (1999).[c-d]Number of geographical regions. [e]Number of Lesser Sunda islands. [f]State: TAS = Tasmania, VIC = Victoria, SA = South Australia, WA = Western Australia, NT = Northern Territory, QLD = Queensland, NSW = New South Wales, N/A = not applicable.

The six species dataset received from DArT Pty Ltd contained 4,876 markers. Redundancy analyses identified 1,847 redundant markers (38% of total). Also excluded from the final dataset were markers with >40% missing data in any one species (143 markers) and 46 low quality markers with reproducibility values (a percent score of how reproducible a marker is within an individual hybridisation) <95%. The final dataset contained 2,840 high-confidence, non-redundant markers, of which 2,408 (85%) had unique placements on the 11 chromosomes of the *E. grandis* reference genome (Myburg *et al.*, 2014) (termed 'anchored markers'). The remaining DArT markers were either annotated to unassembled *E. grandis* scaffolds (104 markers) or not annotated (328 markers). The average interval between annotated markers was 249kbp and only 12 intervals exceeded 2Mbp; the greatest being 3.2Mbp. Overall, the annotated DArT markers spanned between 96.4 to 99.9% of each chromosome, including 597.3Mbp of the total 605.8Mbp (98.6%) sequence assembled in the 11 *E. grandis* chromosomes.

Data analyses

To check that all individuals were assigned to their correct species, a principal coordinate analysis (PCA) based on DArT genotypes was performed in GenAlEx (Peakall & Smouse, 2006) following the generation of a pair-wise genetic distance matrix (Figure 2). Individual marker and species level genetic diversity estimates were then calculated for each species. As *F* estimates (inbreeding coefficient) could not be estimated directly from the dominant DArT markers we estimated genetic diversity within species for each locus using H_{HW} (Kremer *et al.*, 2005; see Supporting Information Table S2). The non-parametric Kruskal-Wallis test in SAS ™ (PROC NPAR1WAY WILCOXON statement; Version 9.2, SAS Institute, Cary, USA) was used to test for differences in H_{HW} estimates between species. As significant differences were detected between species, marker H_{HW} values were standardised for comparing genome-wide patterns of diversity (Standardised H_{HW}=H_{HW} value–species mean). A 'global' H_{HW} for each marker was obtained from the average of the individual species H_{HW} estimates. The overall (termed 'Global') and pair-wise divergence amongst the six species was estimated for each marker using F_{ST}, calculated using BayeScan V2.01 (Foll & Gaggiotti, 2008; Foll, 2010).

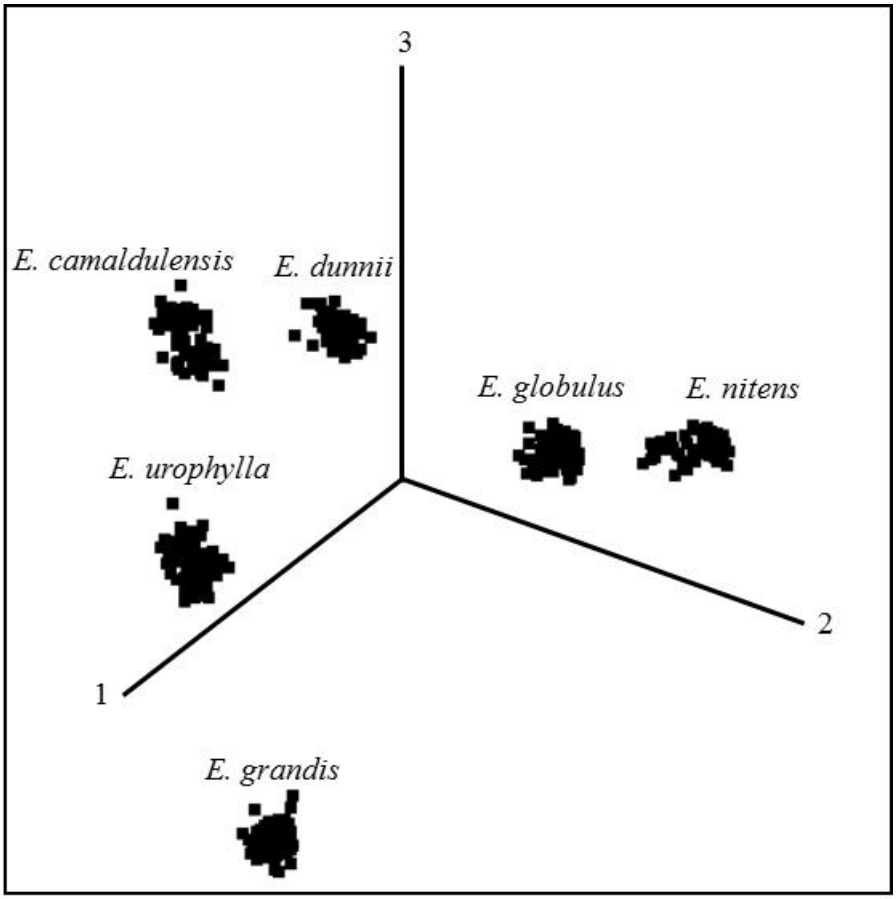

Figure . Principal coordinates analysis of Eucalyptus individuals using diversity arrays technology (rT) mar er genotypes (mar ers).

BayeScan was also used to identify outlier loci in each of 15 pair-wise species comparisons. This program uses a Bayesian method to identify markers with F_{ST} coefficients significantly different from the distribution of values expected under neutral theory (Foll & Gaggiotti, 2008). To avoid biasing the distribution of F_{ST} towards zero, only DArT markers scored as "1" in ≥2% and ≤98% of individuals (e.g. across species pairs or all species in the global analysis) were included in each comparison; a more conservative approach for detecting outliers (Savolainen *et al.*, 2006). Posterior probabilities were calculated for each locus using default BayeScan settings and a prior odds value of 10. Posterior probabilities are not directly comparable to standard *P* values, instead we followed 'Jeffrey's interpretation' (see Foll, 2010), and defined markers with a posterior probability threshold of ≥0.5 as outliers potentially under selection, and those with a posterior probability of ≥0.76 as being under 'substantial' selection'. In each pair-wise species comparison, markers detected as being significant outliers by BayeScan and those not detected in these analyses but which were fixed for different alleles in each species (fixed marker differences, FMDs), were defined as species differentiating markers (SDMs). The failure to detect FMDs as outliers was more common in more divergent comparisons, consistent with increasing background, or putative neutral, differentiation between species decreasing the statistical power to detect outlier markers (Butlin, 2010; Perez-Figueroa *et al.*, 2010). Markers which consistently differentiated the species from different taxonomic sections were also identified. As reported in previous studies (Steane *et al.*, 2011), the species from which the DArT markers originated did affect estimates of species and global H_{HW}, but global F_{ST} and the proportion of (markers which were) SDMs were not significantly affected (Supporting Information Note S1). Similarly while the use of relative (e.g. F_{ST}) rather than absolute measures of species divergence have been questioned (Cruickshank & Hahn, 2014), an average of 82% of SDMs were FMDs.

Genomic analysis

The genome coordinates of DArT markers annotated (at http://www.eucgenie.org/) on the 11 chromosomes of the *E. grandis* genome (V1.0; Myburg *et al.*, 2014) were used to plot standardised H_{HW} estimates (using an 11 marker sliding window average), as well as species global and individual pair-wise F_{ST} estimates for the DArT markers. The differences in global F_{ST}, global H_{HW} and the proportion of SDMs within and between chromosomes were tested with non-parametric Kruskal-Wallis tests using individual markers as the replicates. The within chromosome tests were performed by comparing the differences amongst markers

in 5Mbp bins. This bin size was chosen to ensure good marker representation (average of 19.7 markers per bin).

To test whether variation in H_{HW}, F_{ST} and proportion of SDMs was correlated across species (H_{HW}) and across species pair-wise comparisons (F_{ST} and proportion of SDMs), Spearman rank correlations were calculated. These correlations were computed at the individual marker, 5 Mbp bin and chromosome levels. The tests between species comparisons involved calculation of the pairwise correlations amongst the F_{ST} values derived from various species contrasts. For example, the F_{ST} for each marker from the comparison between *E. grandis* and *E. urophylla* was correlated with that obtained from comparing *E. globulus* and *E. nitens*. There were 15 pairwise comparisons of the six species, resulting in 105 possible correlations (15*14/2).

To test whether the patterns of variation in H_{HW}, F_{ST} and proportion of SDMs could be explained by genomic attributes, Spearman rank correlations were calculated between these genetic parameters and various chromosome features (chromosome length, recombination rate, gene density, density of transposable elements and duplication features). Gene density (expressed as number of genes per Mbp) was calculated based on the gene models annotated in V1.0 of the *E. grandis* genome, for each chromosome as well as for each 5Mbp bin. The relationships amongst gene density, individual species and global H_{HW}, global F_{ST} and the proportion of SDMs was tested at the chromosome and 5Mbp bin levels using Spearman's rank correlation coefficient (r_s) calculated with PROC CORR of SAS™. In order to remove differences between chromosomes and examine the correlations which exist within chromosomes, the 5Mbp bin data was standardised to a mean of zero for each chromosome using PROC STANDARD of SAS™. Regressions of chromosomal variation in global H_{HW}, global F_{ST} and proportion of SDMs on gene density were undertaken with PROC SGSCATTER of SAS™.

Association of global H_{HW}, global F_{ST} and the proportion of SDMs with other chromosomal features (reported in Myburg *et al*. [2014], and Petroli *et al*. [2011]) was tested using Spearman rank correlations at the chromosome level. The duplication features used for these analyses included: the number of clusters of tandem duplicate genes; and the numbers of pairs of duplicated genes linked to segmental duplication; lineage-specific genome-wide duplication and ancient hexaploidization – each expressed relative to the number of genes on

the chromosome (Myburg *et al*., 2014, Supplementary data 19). We tested for the relationship with the density of three classes of transposable elements (DNA transposons, retrotransposons and uncategorised transposons; Myburg *et al*., 2014, Supplementary data 13) relative to chromosome length. We also examined the relationship with recombination rate (cM/Mbp) based on chromosome-level values reported for an *E. grandis* × *urophylla* hybrid pedigree (Petroli *et al.* 2012).

To test whether the proximity of DArT markers to genes affected the patterns of diversity, we classified markers into three classes based on predicted gene models in V1.0 of the *E. grandis* genome – those (i) within genes, (ii) within 5kbp of a gene, and (iii) more than 5kbp from a gene. Consistent with the observations of Petroli *et al.* (2012), most DArT markers were in genes. The 5kbp interval was chosen to identify markers proximal to genes, which may be influenced by the combined effects of linkage disequilibrium and selection on adjacent genes or regulatory elements (Pannell & Fields, 2014). Non-parametric Kruskal-Wallis tests were used to test for the differences in diversity and divergence parameters between these three marker classes. We further partitioned the DArT markers within genes into those within and outside of tandem duplicate gene clusters (as identified in Myburg *et al.*, 2014), and tested for differences using non-parametric Kruskal-Wallis tests. As we found significant differences between marker classes (see results), we re-ran the Spearman correlation analyses detailed in the previous paragraph using the DArT marker subsets. This was undertaken to remove the influence of variation in the relative frequency of DArT markers within each of the marker classes on the correlations with chromosomal features.

To provide insights into the functional nature of the genes putatively involved in species differentiation, genes located within 5kb up- and down-stream of DArT markers (termed 10kbp regions) were used in GO enrichment analyses using Blast2GO (Conesa *et al.*, 2005). For this analysis of functional category over-representation in the identified genes, Gene Pfam annotations (available from http://www.phytozome.net/eucalyptus.php) were firstly converted to GO terms using a conversion file obtained from http://www.geneontology.org (file version 14/07/2012). A functional classification of genes within the 10kbp regions was conducted in SDM and non-SDM groups. A two-tailed Fisher's exact test was used to test for enriched GO terms within the SDM gene group compared with the non-SDM gene group. In addition we specifically checked whether SDMs were in, or within 5kbp of, predicted gene models from (economically and evolutionary) important gene classes identified in Myburg *et*

al. (2014). These comprised gene models associated with lignocellulosic biomass production, secondary metabolites and oils, the MADS Box gene family, and the S-domain-Receptor-Like Kinase (SDRLK) gene family (Myburg *et al.*, 2014, Supplementary notes S5 to S8).

Results

Species differentiating markers

The six eucalypt species sampled were all well differentiated by the DArT markers, with no evidence of admixture (Figure 2). The 881 pair-wise SDMs listed in Table 2 were due to 365 differentiated markers, of which 62% were involved in more than one pair-wise comparison (Table 2; Supporting Information Table S3). Nineteen of these 365 SDMs were detected as fixed marker differences (FMDs) only, the remaining 346 were detected as outliers in BayeScan analyses and most of these were also FMDs. All outliers had F_{ST} estimates greater than the pair-wise species mean values, consistent with divergent selection (Foll, 2010). A greater proportion of the SDMs were detected only as BayeScan outliers, as opposed to fixed marker differences, in the comparisons involving the *Latoangulatae* and *Exertaria* species compared with those involving the more differentiated species from section *Maidenaria* (Table 2). This is consistent with decreased power to detect outliers in more differentiated species. Consequently, not all loci potentially subject to selection may have been detected.

The markers associated with species differentiation tended to be similar regardless of the species pairs being compared. There was a positive correlation of the marker F_{ST} values (n=2,840) involving the fifteen pair-wise comparisons of the six species. The Spearman r_s's averaged 0.50 and were highly significant (P<0.001) for all 105 correlations possible amongst the 15 species pairs (range 0.31-0.70). The presence of SDMs were less consistent (mean r_s=0.15), with 30% of the 105 comparisons not significant (P≥0.05). Of the unique SDMs, 138 (38%) were only detected in a single pair-wise comparison. Of the remaining SDMs, 28 (12%) consistently differentiated species from different sections (section differentiating markers; Figure 3; Supporting Information Fig. S1), and are thus candidates for more ancient phylogenetic divergence. For example, two section differentiating SDMs (on chromosomes 6 [ePt-573990] and 11[ePt-636783]) were found to differentiate the *Maidenaria* species from those in sections *Latoangulatae* and *Exsertaria*.

Genomic structuring of genetic diversity and divergence

The annotation of DArT markers on the *E. grandis* genome sequence allowed the genome-wide patterns of diversity and divergence to be investigated. Of the 365 unique SDMs identified, the physical positions of 297 (81%) were identified. The genome-wide patterns of

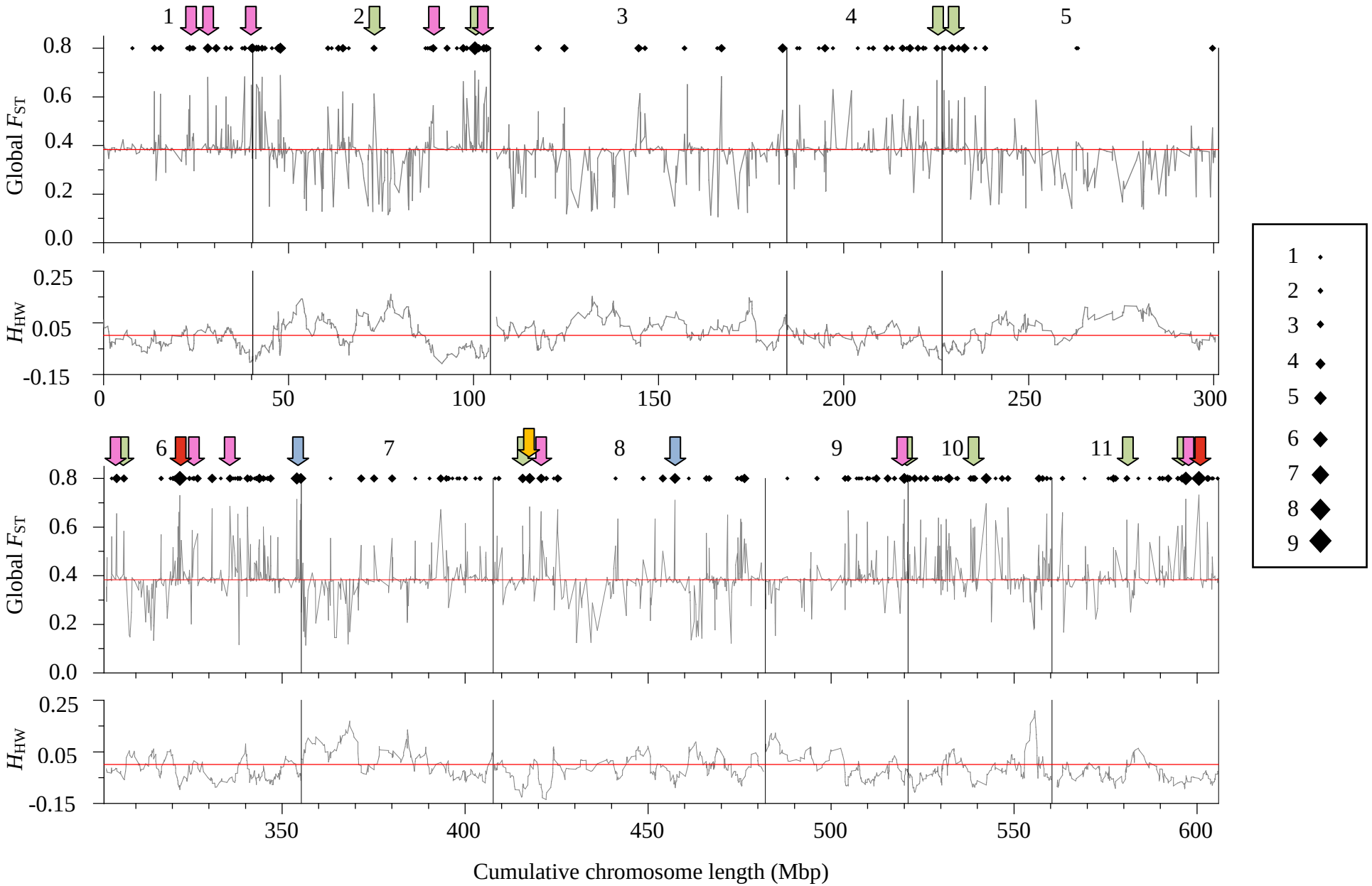

Figure . hromosome wide distributions of global FST mar er estimates and genetic diversity (standardi ed ; average mar er values for all six Eucalyptus species plotted using an 11 mar er sliding window average). ertical bars indicate chromosome boundaries, with chromosome numbers given above figures. ed lines indicate average FST (.) and (in this standardi ed data). The positions of species differentiating mar ers (S s) detected in pairwise species comparisons are indicated by diamond symbols; symbol si e is proportional to the number of times each mar er was detected as an S across all 1 comparisons (see scale on right). The positions of section differentiating mar ers are indicated by coloured arrows above S s: red, section aidenaria vs atoangulatae Exsertaria; orange, Exsertaria vs aidenaria atoangulatae; blue, aidenaria vs atoangulatae; green, aidenaria vs Exsertaria; pin , atoangulatae vs Exsertaria

Table 2 Number of species differentiating markers (SDMs) detected from BayeScan analyses in each *Eucalyptus* species comparison.

Section/species comparison	FMDs	BayeScan outlier results[1]					Total SDMs[2]
		n	Mean F_{ST}	Number and significance level of outlier markers			
				≥ 0.50	≥ 0.76	Total	
Exertaria* vs. *Maidenaria							
E. camaldulensis vs. *E. dunnii*	83	2517	0.398	78	4	82	83
E. camaldulensis vs. *E. globulus*	39	2354	0.349	7	39	46	46
E. camaldulensis vs. *E. nitens*	48	2345	0.381	39	8	47	48
Exertaria* vs. *Latoangulatae							
E. camaldulensis vs. *E. urophylla*	2	2387	0.243	7	23	30	30
E. camaldulensis vs. *E. grandis*	14	2370	0.303	13	33	46	46
Latoangulatae* vs. *Maidenaria							
E. grandis vs. *E. dunnii*	116	2488	0.440			0	116
E. grandis vs. *E. globulus*	41	2271	0.380		41	41	41
E. grandis vs. *E. nitens*	57	2237	0.413			0	57
E. urophylla vs. *E. dunnii*	62	2548	0.409	35		35	62
E. urophylla vs. *E. globulus*	25	2389	0.357	4	25	29	29
E. urophylla vs. *E. nitens*	34	2349	0.386	33		33	34
Intra-section comparisons							
E. grandis vs. *E. urophylla*	2	2296	0.267	8	21	29	29
E. globulus vs. *E. dunnii*	76	2131	0.317	15	98	113	113
E. globulus vs. *E. nitens*	22	1923	0.309	9	32	41	41
E. nitens vs. *E. dunnii*	106	2149	0.401	94		94	106
Average	48.5	2317	0.357	-	-	44	59
Total		-	-	342	324	666	881

[1]*n*=number of markers included in each comparison (applying 2-98% criteria; see Methods). The number of outlier markers having posterior probability values of ≥0.50 to <0.76, and ≥0.76 are indicated.

[2]Total SDMs include outlier markers and/or fixed marker differences (FMDs) between species.

variation in global F_{ST} and global H_{HW} and the distribution of the 297 SDMs are shown in Figure 3 (see also Supporting Information Fig. S1). The patterns of variation in H_{HW} of each component species, as well as pair-wise F_{ST} estimates and SDMs are given in Supporting Information Figs. S1 and S2. For the species and pair-wise comparisons many of the same trends can be observed as summarised by the variation in global F_{ST} and global H_{HW} shown in Figure 3. For example, at the 5 Mbp bin level, the Spearman correlation of the H_{HW} of individual species with the global H_{HW} shown in Figure 3 was positive (P<0.001) and ranged from 0.68 (*E. urophylla*) to 0.79 (*E. globulus*) (Supporting Information Table S2).

Variation between chromosomes

The distribution of marker diversity and divergence was highly structured across the genome, with variation both between and within chromosomes. Among chromosomes, significant differences were detected for species global H_{HW} (Kruskal-Wallis χ^2 54.5, df=10, P<0.001), global F_{ST} (Kruskal-Wallis χ^2 113.5, df=10, P<0.001) and proportion SDMs (Kruskal-Wallis χ^2 39.5, df=10, P<0.001) (Figure 3).

The pattern of structural variation in these diversity and divergence parameters was not independent. At the chromosome level, mean global F_{ST} and the proportion of SDMs were highly positively related (n=11; r_s=0.83, P=0.002), and both were significantly (P<0.001) inversely correlated with global H_{HW} (Figure 3; mean global F_{ST} r_s=-0.93 and proportions of SDMs r_s=-0.83). The differences amongst chromosomes in marker diversity and divergence were relatively stable across species. All pair-wise species comparisons were positively correlated for the chromosome average H_{HW}, with 9 of the 15 significant at P<0.05 (Table 3). Chromosome average F_{ST} was positively correlated across the 105 comparisons involving the 15 pair-wise F_{ST} estimates amongst the six species, with 52% significant at P <0.05 (mean r_s=0.60, range 0.08-0.94). In contrast, only 18% of chromosome level correlations for the proportion of SDMs in the pair-wise comparisons were significant (n=105; P<0.05), these were all positive and most involved a shared species in the pair-wise comparisons (e.g. *E. grandis* vs. *E. dunnii* and *E. camaldulensis* vs. *E. dunnii*, r_s=0.95, P<0.001). These results suggest that while the differences amongst chromosomes in average H_{HW} and F_{ST} were relatively stable amongst species, this was less so for the location of pair-wise SDMs.

Table 3 Pair-wise Spearman correlations (r_s) of genetic diversity (H_{HW}) within six eucalypt species at the a) marker , b) chromosome, and c) within chromosome (amongst 5 Mbp bins) levels, and between H_{HW} and gene density at the between and within chromosome levels.

Correlation[1]	Section:	*Latoangulatae*		*Exsertaria*	*Maidenaria*		
	Species:	*E. grandis*	*E. urophylla*	*E. camaldulensis*	*E. globulus*	*E. nitens*	*E. dunnii*
	Species Mean $H_{HW:}$	0.205	0.238	0.226	0.183	0.154	0.197
(a) Amongst markers (n=2840)	*E. urophylla*	0.34***					
	E. camaldulensis	0.31***	0.39***				
	E. globulus	0.25***	0.2***	0.30***			
	E. nitens	0.23***	0.17***	0.25***	0.43***		
	E. dunnii	0.18***	0.14***	0.23***	0.46***	0.34***	
(b) Amongst chromosomes (n=11)	*E. urophylla*	0.60 ns					
	E. camaldulensis	0.39 ns	0.67*				
	E. globulus	0.68*	0.58 ns	0.59 ns			
	E. nitens	0.72*	0.65*	0.54 ns	0.74**		
	E. dunnii	0.69*	0.91***	0.68*	0.53 ns	0.79**	
	Gene density	-0.76**	-0.58 ns	-0.71*	-0.67*	-0.69*	-0.70*
(c) Within chromosomes (amongst 5 Mbp bins pooled over chromosomes [n=122])	*E. urophylla*	0.38***					
	E. camaldulensis	0.35***	0.50***				
	E. globulus	0.44***	0.38***	0.51***			
	E. nitens	0.40***	0.31***	0.38***	0.53***		
	E. dunnii	0.38***	0.34***	0.36***	0.54***	0.42*	
	Gene density	-0.48***	-0.28**	-0.29**	-0.42***	-0.32***	-0.50***

[1] ns = $P \geq 0.05$; * = $P<0.05$; ** = $P<0.01$; *** = $P<0.001$.

Variation within chromosomes

Based on the variation amongst 5Mbp bins, significant (Kruskal-Wallis test; $P<0.05$) non-random variation in species global H_{HW}, global F_{ST} and the proportion of SDMs was detected within 7-8 of the 11 chromosomes, depending upon the population parameter being tested (notably chromosomes 2, 3, 5, 6, 7, 8, and 9). Within chromosomes, regions with above-average global F_{ST} values coincided with regions of reduced global H_{HW} and *vice versa* (pooled 5Mbp bins within chromosomes r_s=-0.6, $P<0.001$) (Figure 3). Expectedly, SDMs were commonly found to reside within regions of elevated global F_{ST} values (pooled 5Mbp bins within chromosome r_s=-0.70, $P<0.001$) (Figure 3; Supporting Information Fig. S1).

Associations with gene density and other chromosome features

Of the ten chromosomal features studied, gene density and the relative number of clusters of tandem duplicate genes were the most significantly correlated with global population parameters based on the 2,408 anchored DArT markers (Supplementary Information Table S4). The density of two classes of transposable elements were also correlated with divergence between species (global F_{ST}), (DNA transposons [r_s=-0.81, Bonferroni P=0.026] and uncategorised transposons [r_s=-0.85, Bonferroni P=0.010]). These explanatory variables were significantly inter-correlated, with a decrease in the relative number of clusters of tandem duplicated genes (n=11; r_s=-0.69, P= 0.019), DNA transposon density (n=11; r_s=-0.81, P= 0.003) and uncategorised transposon density (n=11; r_s=-0.80, P= 0.003) as gene density increased. Further, the density of DNA transposons and uncategorised transposons were no longer significantly correlated with any of the global population parameters in partial correlations after removing the effect of gene density. Therefore, we interpreted the significant correlations of these two classes of DNA transposable elements as largely reflecting their strong correlation with gene density.

Gene density varied markedly amongst the *E. grandis* chromosomes, varying from 43 [chromosome 3] to 71 [chromosome 6] genes per Mbp (Myburg *et al*., 2014; average [±s.d.] length of gene models was 3.18±2.90 kp; Supporting Information Fig. S3). DArT markers on the more gene-dense chromosomes were less diverse (global H_{HW} r_s=-0.88, Bonferroni P=0.003) and exhibited higher divergence between species (n=11; global F_{ST} r_s=0.92, Bonferroni $P<0.001$) and higher proportion of SDMs (r_s=0.75, Bonferroni P=0.085) (Figure 4). Spearman correlations amongst 5Mbp bins pooled within chromosomes showed similar significant ($P<0.001$) correlations (n=122; global H_{HW} r_s=-0.55; global F_{ST} r_s=0.55; the proportion of SDMs r_s=0.47; see also Supporting Information Fig. S4). Further, this

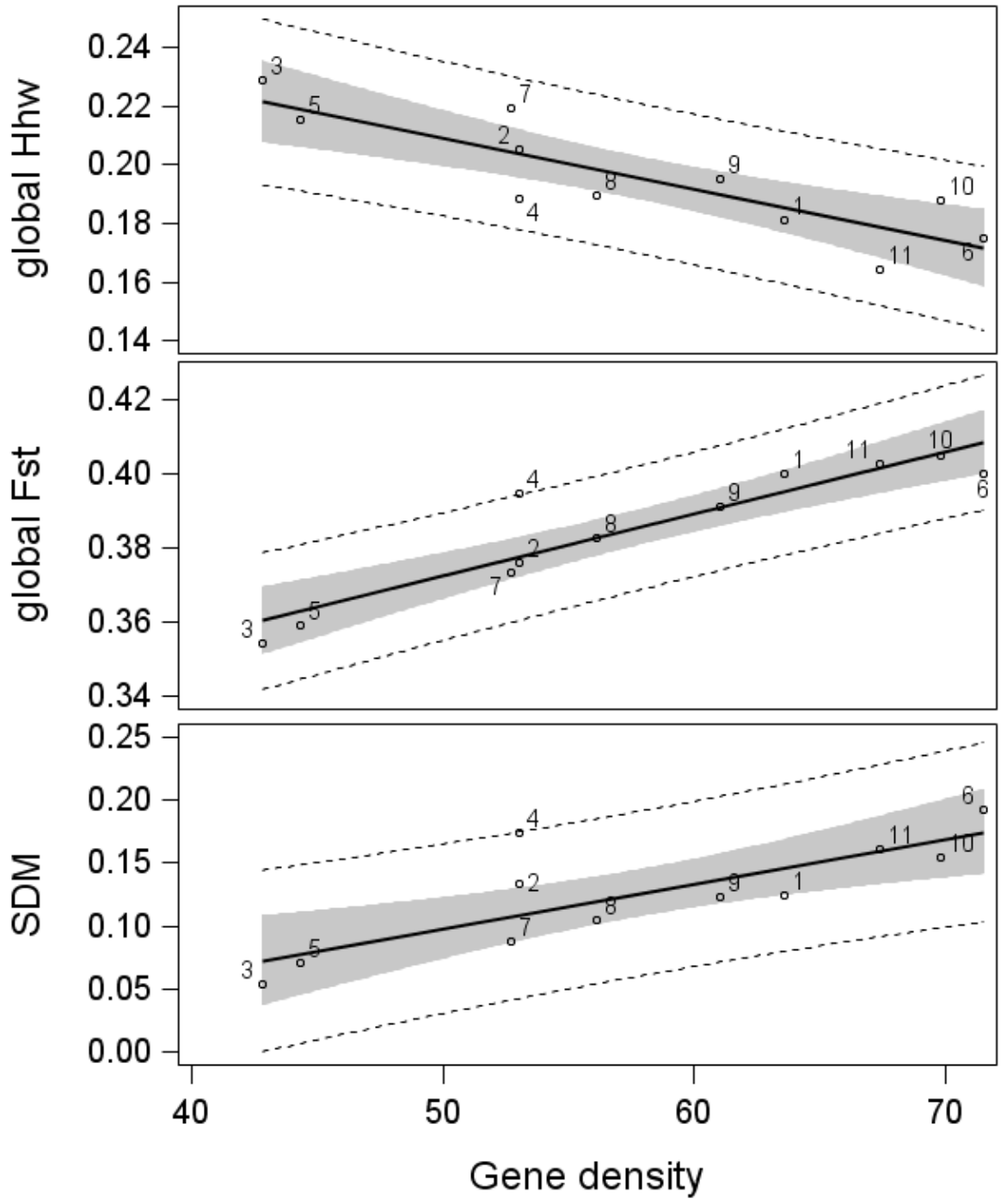

Figure . The significant relationship of global species diversity (; six Eucalyptus species, average, top), global divergence (FST, middle) and the proportion of species differentiating mar ers (S s, bottom) with chromosome mean gene density (genes bp 1). hromosome numbers are indicated and the confidence intervals for the mean (shaded) and individual (dotted line) predicted values are shown.

correlation of diversity and divergence parameters with gene density was relatively stable across species. For example, variation in H_{HW} was negatively correlated with gene density at both the chromosome and within chromosome levels in all six species (11 of the 12 correlations were significant at $P<0.05$; Table 3). The reverse trends were observed at the chromosome level for the relative number of clusters of tandem duplicate genes (n=11; global H_{HW} r_s=0.85, Bonferroni P=0.010; global F_{ST} r_s=-0.78, Bonferroni P=0.045 and proportion SDMs r_s=-0.85, Bonferroni P=0.008).

Effect of gene proximity

DArT diversity decreased and species divergence increased with increasing proximity to genes (Figure 5). The trend evident for global H_{HW} (Figure 5) was also evident for H_{HW} of the individual species (Supplementary Information Fig. S5). Nevertheless, while often not significant, the same trends with chromosome features as detailed above were evident when markers were separated into each of the three gene proximity classes (Supplementary Information Table S4). Of particular note were the DArT markers within genes. These intragenic markers showed chromosome-level relationships of global F_{ST} with gene density (n=11; r_s=0.95, Bonferroni $P<0.001$), relative number of clusters of tandem duplications (n=11; r_s=-0.85, Bonferroni P=0.01), DNA transposon density (n=11; r_s=-0.81, Bonferroni P=0.026) uncategorised transposon density (n=11; r_s=-0.84, Bonferroni P=0.013) and recombination rate (n=11; r_s=0.81, Bonferroni P=0.026), which were similar to those obtained using all DArT markers. This result argues that the relationships observed are a feature of the genes themselves rather than chromosomal variation in the proportion of markers which were within genes.

SDMs were proportionally more common in markers occurring in or near genes (Figure 5), consistent with an adaptive role of these or linked genes in species differentiation. For the anchored DArT markers, the differences in the proportion of SDMs amongst chromosomes was positively related to gene density (e.g. Figure 4; r_s=0.75, Bonferroni P=0.085). An identical association with gene density was detected with DArT markers that occurred in genes, suggesting again that this association was not simply a function of chromosomal variation in the proximity of markers to genes (Supporting Information Table S4).

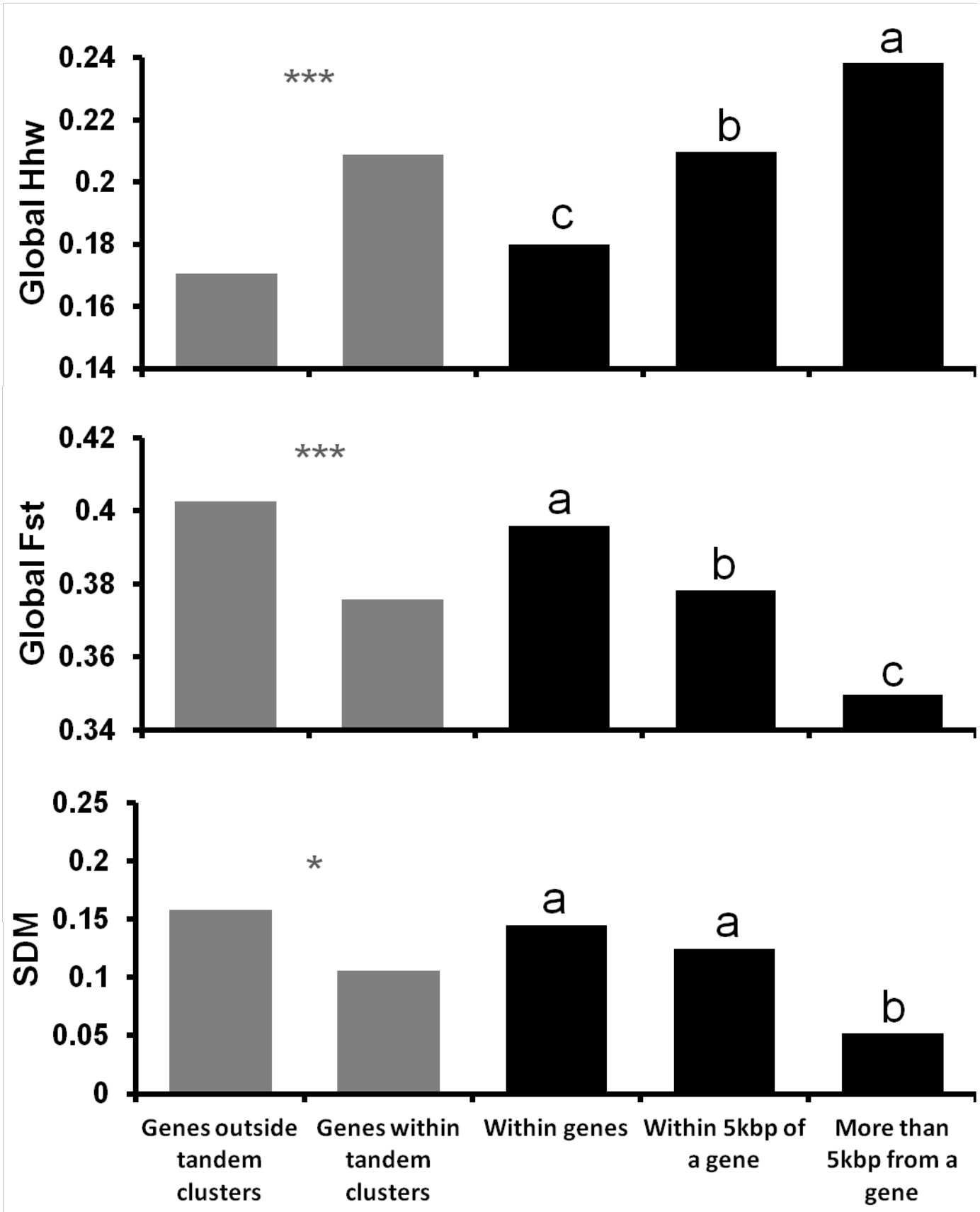

Figure . The average global species diversity (; six Eucalyptus species, average, top), global divergence (FST, middle) and the proportion of species differentiating mar ers (S s, bottom) for diversity arrays technology (rT) mar ers in genes (n 1), within b of a gene (n 1), and b from a gene (n) (blac bars) (a) and rT mar ers in genes that were outside and within clusters of tandemly duplicated genes (grey bars) (b). onsistent with other studies (Petroli et al., 1), most rT mar ers occurred in genes. eans with common letters (above blac bars) are not significantly (P .) different based on pairwise rus al allis tests. eans with different letters were significantly different at the P . 1 level. The significance of the rus al allis tests for the difference between rT mar ers within genes that were outside (n 11) and within (n) clusters of tandem duplicate genes are shown (, P . ; , P . 1).

Effect of tandem duplications

While the DArT markers in genes had higher global F_{ST} values on more gene-dense chromosomes (r_s=0.95, Bonferroni P<0.001), SDMs tended to be associated with chromosomes having relatively fewer clusters of tandem duplicate genes (r_s=-0.90, Bonferroni P=0.002) (Supporting Information Table S4). Indeed, of the ten chromosomal features examined, the only partial correlations which were significant after removing the effect of gene density were those which involved the relative number of clusters of tandem duplicate genes (e.g. for DArT markers in genes: global F_{ST} partial r_s=-0.82, Bonferroni P=0.036; SDMs partial r_s=-0.80, Bonferroni P=0.050). This suggests gene density and the relative number of clusters of tandem duplications are key explanatory variables, amongst the inter-correlated chromosomal features associated with variation in our population genetic parameters.

To explore the effect of tandem duplication on gene diversity and divergence, we partitioned DArT markers within genes into those which occurred within clusters of tandem duplications from those that did not. Consistent with the chromosomal level correlations, the markers within tandem duplications had significantly higher diversity (global H_{HW}) and lower divergence (global F_{ST} and SDMs), and were more similar to markers outside of genes (Figure 5). However, for (intragenic) DArT markersnot in clusters of tandem duplications, global species divergence (F_{ST} and SDMs) was still positively correlated with gene density (global F_{ST} n=11 r_s=0.79 Bonferroni P=0.037, SDMs n=11 r_s=0.67 Bonferroni P=0.23) and negatively correlated with the number of tandem duplicate gene clusters (global F_{ST} r_s=-0.87 Bonferroni P=0.005, SDMs n=11r_s=-0.95 Bonferroni P=0.001). No significant correlations involving these chromosomal features were detected for intragenic DArT markers within clusters of tandem duplicate genes (Supporting Information Table S4). Rather divergence in the intragenic markers within clusters was negatively correlated with chromosome length (n=11; global F_{ST} r_s=-0.89 Bonferroni P=0.002) and positively correlated with retrotransposon density (n=11; global F_{ST} r_s=0.78 Bonferroni P=0.045). This correlation with chromosome length was also evident for both classes of intergenic markers (within 5kb of a gene r_s=-0.80 Bonferroni P=0.031; more than 5kb of a gene r_s=-0.80 Bonferroni P=0.031). After Bonferroni adjustment for multiple comparisons, chromosome length and DNA transposon and retrotransposon density were the only chromosomal features correlated with any diversity and divergence parameter for (i) intergenic markers or (ii) intragenic markers within clusters of tandem duplications, and these correlations only involved global F_{ST} (Supporting Information Table S4).

Genomic characterisation of differentiated regions

Gene ontology terms could be assigned to 53.4% of the genes identified within DArT marker 10kbp regions (Supporting Information Table S5). Using SDM (containing 225 genes) and non-SDM gene groups (containing 1,272 genes), the mapping of genes to second-level GO terms revealed that the gene content of the two groups was highly similar (Supporting Information Fig. S6 and Table S6). Enrichment analysis initially detected 13 GO terms that were significantly under- or over-represented in the SDM gene group relative to the non-SDM gene group (Supporting Information Table S7). However, the significance of these terms was generally marginal and none were significant when a false discovery rate of Q=0.05 was applied.

Few genes that have been specifically studied in eucalypts were located close to SDMs. In particular, of the *E. grandis* genes for which expression data was reported in Myburg *et al.* (2014; which included 174 genes in the lignin pathway, 237 peroxidase and/or laccase protein genes, 164 cellulose and xylase related genes, 50 MYB genes, 253 S-domain-Receptor-Like Kinase [SDRLK] genes, 134 terpene synthase genes, as well as 107 MADS-box genes) only 42 (3.8%) were within 5kb of our DArT markers, 24 (2.1%) of which had DArT markers within the annotated gene. Of these 42 genes, two had embedded SDMs and another four had SDMs within 5kbp. Of note was the SDM (ePt-641068) embedded within a sucrose metabolism *SUSY* gene (*Eucgr.J01640 - ATSUS5-SUS5*), which differentiated six species pairs; and an SDM (ePt-573579) in a copper ion binding laccase gene (*Eucgr.E00461*) which differentiated three species pairs. Other genes which may be genetically linked to SDMs include a MADS-box gene (*EgrSoc5.2*), where the marker (ePt-504279) differentiated *E. dunnii* from both *E. globulus* and *E. nitens*. None of the reported genes were in close proximity to SDMs which differentiated the well-studied *E. grandis* and *E. globulus*, nor the intrasection comparisons comparing *E. nitens* with *E. globulus* (section *Maidenaria*) and *E. grandis* with *E. urophylla* (section *Latoangulatae*). Nevertheless, there were two SDMs (ePt-637106 on chromosome 3 [near *Eucgr.*C02258] and ePt-642862 on chromosome 8) which differentiated species within both intrasectional species comparisons, signalling the location of a gene(s) or genomic region that may be under selection in different phylogenetic lineages.

Discussion

This is the first major study of genomic diversity and species differentiation in *Eucalyptus* and one of the few in plants to investigate more than a single species-pair (e.g. Rymer *et al.*, 2010; Renaut *et al.*, 2013). Consistent with recent reports in plants (Flowers *et al.*, 2012; Renaut *et al.*, 2013) and animals (Ellegrin *et al.*, 2012; Cutter and Payseur 2013), our results provide evidence that genomic architecture strongly influences genomic patterns of marker diversity and divergence amongst eucalypt species. While it is difficult to disentangle the influence of individual genomic features, many of which are inter-correlated, the proximity of markers to genes, the relative number of clusters of tandem duplications and gene density emerge as key explanatory variables associated with these patterns and are discussed below.

As expected, intragenic DArT markers exhibited less diversity within species and more divergence between species than intergenic markers. This is consistent with selection reducing genetic diversity and contributing to differentiation (Roesti *et al.*, 2012; Ellegren *et al.*, 2012), although differences in mutation rates across the genome could influence the observed pattern (Meirmans & Hedrick 2011). The species differentiating DArT markers identified as likely subjects of selection (i.e. SDMs) were found more often in genomic regions with globally high F_{ST}, low average species diversity and were more frequent in genic than non-genic markers. The fact that some SDMs were detected in intergenic markers may reflect regulatory elements under selection (e.g. Zhang *et al.*, 2011) but could also arise from failure to identify a gene model or incorrect annotation of a DArT sequence. The intermediate differentiation in markers in close proximity to genes could reflect linked selection on adjacent genes (Cutter and Payseur 2013, Slotte 2014) and/or an increasing concentration of regulatory elements under selection with proximity to genes (e.g. Zhou *et al.*, 2011).

While diploid in nature, the eucalypt genome has been shaped by an early lineage-specific genome duplication (specific to the order Myrtales) and subsequent tandem gene duplication (Myburg *et al.*, 2014). Indeed, *E. grandis* has the highest proportion of tandem gene duplications yet reported in plants (Myburg *et al.*, 2014). Gene duplication has long been thought to generate evolutionary novelty (e.g. Ohno 1970). Consistent with our study, tandem duplicated genes often exhibit elevated intraspecific diversity (e.g. Clark *et al.*, 2007). This has been attributed to relaxed purifying selection on newly duplicated paralogues and

subfunctionalization of paralogues (Lynch and Conery 2000; Flagel and Wendel, 2009), although intraspecific adaptation may also contribute (Hanada *et al.*, 2008; Turner *et al.*, 2008). Further, the effects of tandem duplication appear to extend to genes beyond the clusters, such that chromosomes with proportionally more clusters exhibit increased diversity and decreased species divergence in these genes. This relationship could at least partly reflect linked selection (Cutter and Payseur 2013; Slotte 2014).

Gene density within and between chromosomes was positively correlated with interspecific divergence and negatively correlated with intraspecific diversity. These relationships were stable across the six species studied, and thus transcend phylogenetic lineages, biogeography and ecology. Such negative correlation between gene density and genetic diversity has previously been reported in humans (Payseur & Nachman, 2002), *Ceanorhabditis* (Cutter & Payseur, 2003), *Arabidopsis* (Nordborg *et al.*, 2005), *Medicago* (Branca *et al.*, 2011) and *Oryza* species (Flowers *et al.*, 2012). In each case, the correlation was attributed to more intense selection (functional constraint) in regions of greater gene density. Selection can impact on genetic diversity through positive and negative directional selection, causing mutations to be driven to fixation or extinction. The impact of such selection can spread to adjacent coding or non-coding regions of the genome through linked selection (Slotte, 2014). In these cases, both the spread of polymorphisms linked to beneficial mutations (genetic hitchhiking) and the removal of deleterious mutations (background selection) will reduce genetic diversity in the vicinity of mutations under selection (Cutter & Payseur, 2013).

In contrast to genetic diversity, reports of a correlation between interspecific divergence and gene density are rare. However, consistent with our study, gene density was negatively correlated with intraspecific diversity and positively correlated with nucleotide divergence between *Oryza* species (Flowers *et al.*, 2012). In our case, this association was strongest in genic markers, arguing the association is a property of the genes, not marker distribution, and was evident both within and between chromosomes. Greater divergence between species in gene dense regions may simply reflect enhanced selection in regions of high gene density (Flowers *et al.*, 2012), combined with linked selection as discussed above. Alternatively, it has been suggested that during genome evolution clustering of tightly linked loci involved in adaptation can occur through minor rearrangements or tandem duplication, creating clusters of co-expressed genes (Hurst *et al.*, 2004; Renaut *et al.*, 2013; Yeaman 2013). In our case this

effect does not appear to be a function of tandem duplication as these are negatively associated with species divergence at the chromosome level.

In all species comparisons, the SDMs identified were dispersed across most of the genome, consistent with previous studies investigating the genomic basis of plant speciation in both closely and distantly related populations (reviewed by Nosil *et al.*, 2009; Strasburg *et al.*, 2012). These markers were mainly in, or near, genes and may result from the action of divergent selection or endogeneous barriers to gene flow arising from intrinsic genetic incompatibilities (Bierne *et al.*, 2011; Seehausen *et al.*, 2014). Divergence between genetically isolated species is predicted to impact more loci due to the combined processes of selection pressures and genetic drift (Nosil *et al.*, 2009); as well as the build up of reproductive isolation, through extrinsic (pre- and post-zygotic) and intrinsic (post-zygotic) mechanisms (Seehausen *et al.*, 2014). The eucalypt species studied are well along the speciation continuum (Seehausen *et al.*, 2014), compared with those often studied (Cruickshank & Hahn 2014). They occupy markedly different habitats and have diverged, at least at the sectional level, c. 10-15 Mya (Crisp *et al.*, 2011). There is evidence for strong intrinsic post-zygotic barriers between sections (*Maidenaria* versus *Latoangulatae*/*Exertaria*; Myburg *et al.*, 2004; Potts & Dungey, 2004) and within section *Maidenaria* (*E. nitens* × *E. globulus*; Costa e Silva *et al.*, 2012). High F_{ST} and fixed marker differences between the parapatric *E. grandis* and *E. dunnii* attest to the strength of these isolating barriers.

Genes located within or adjacent to SDMs may represent targets of divergent selection or underlie reproductive isolation between the species in this study. Further support for such candidate genes can be provided by knowledge of their biological functions (e.g. Turner *et al.*, 2008). Specifically, the MADS-box gene (*EgrSoc5.2*), part of the *SOC* subfamily involved in flowering initiation, was located adjacent to an SDM differentiating the temperate *E. globulus* and *E. nitens* from the subtropical *E. dunnii*. Genes influencing flowering time or floral morphology are obvious candidates for reproductive isolation (Rieseberg & Blackman, 2010; Rymer *et al.*, 2010). In particular a MADS-box transcription factor (*FLC*) has been previously implicated in flowering time differences between *Arabidopsis* species (Wang *et al.*, 2006). Likewise, it has been proposed that the diversification of the *SOC* subfamily in *Eucalyptus* may have contributed to the evolutionary diversification of the genus by mediating the flowering response to a range of environmental cues, to suit a wide range of habitats (Myburg *et al.*, 2014).

Our study shows that genomic architecture influences the patterns of molecular intraspecific diversity and interspecific divergence across the eucalypt genome. Together with evidence from comparative mapping (Hudson *et al.*, 2012) and sequence analysis (Myburg *et al.*, 2014; Tibbits *et al.*, submitted), our results point to general conservation of genomic architecture across these diverse *Eucalyptus* species. This is evident despite considerable differences in genome size between species (Grattapaglia & Bradshaw, 1994; Grattapaglia *et al.*, 2012). No significantly enriched GO terms associated with species differentiation were detected in this study (after correction for the false discovery rate), arguing that the differentiation between species is multigenic and involves different classes of genes at different loci. The integration of results from this study with others (e.g. comparative transcriptomics, association genetics and QTL studies; see Faria *et al.*, 2014) combined with future analyses at higher resolution should allow the identification of genes contributing to differentiation and reproductive isolation in *Eucalyptus*.

Acknowledgements

We thank Dean Williams (Forestry Tasmania), Kelsey Joyce (Gunns Ltd) Simon Southerton and Charlie Bell (CSIRO Plant Industries) for samples; Peter Harrison and Paul Tilyard for technical assistance. We acknowledge Geoff Galloway and Andrea Louw at Sappi Forest Researchand Kitt Payn at Mondi Tree Improvement Research for *E. grandis*, *E. dunnii* and *E. urophylla* samples, as well as Sappi and Mondi (South Africa). We also thank the anonymous reviewers for their constructive suggestions. This research was supported by the Australian Research Council (DP110101621, DP0986491, DP130104220 and DP140102552), Sappi and Mondi, the Technology and Human Resources for Industry Programme (THRIP, UID 80118), the National Research Foundation (NRF, UID 71255 and 86936) and the Department of Science and Technology (DST) of South Africa.

References

Achere V, Favre JM, Besnard G, Jeandroz S. 2005. Genomic organization of molecular differentiation in Norway spruce (*Picea abies*). *Molecular Ecology* **14**: 3191-3201.

Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology* **20**: 2044-2072.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L *et al*. 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Science of the United States of America* **108**: E864–E870.

Brooker MIH. 2000. A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). *Australian Systematic Botany* **13**: 79-148.

Butcher P, McDonald M, Bell J. 2009. Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*. *Tree Genetics & Genomes* **5**: 189-210.

Butlin R. 2010. Population genomics and speciation. *Genetica* **138**: 409-418.

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA *et al*. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**: 338-342.

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

Costa e Silva J, Potts BM, Tilyard P. 2012. Epistasis causes outbreeding depression in eucalypt hybrids. *Tree Genetics & Genomes* **8**: 249-265.

Coyne JA, Orr HA. 2004. *Speciation*. Massachusetts, USA: Sinauer Associates Inc.

Crisp MD, Burrows GE, Cook LG, Thornhill AH, Bowman DMJS. 2011. Flammable biomes dominated by eucalypts originated at the Cretaceous-Palaeogene boundary. *Nature Communications* **2**: 193.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**: 3133-3157.

Cutter AD, Payseur BA. 2003. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Molecular Biology and Evolution* **20**: 665-673.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* **14**: 262-274.

Doughty R. 2000. *The Eucalyptus: a natural and commercial history of the gum tree*. Baltimore, MD: John Hopkins University Press.

Dutkowski GW, Potts BM. 1999. Geographic patterns of genetic variation in *Eucalyptus globulus* ssp. *globulus* and a revised racial classification. *Australian Journal of Botany* **47**: 237-263.

Eckert AJ, Bower AD, González-Martínez SC, Wegrzyn JL, Coop G, Neale DB 2010. Back to nature : ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology* **19**: 3789-3805.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* **29**: 51-63.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A *et al*. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**: 756-760.

Faria R, Renaut S, Galindo J, Pinho C, Melo-Ferreira J, Melo M, Jones F, Salzburger W, Schluter D, Butlin R. 2014. Advances in ecological speciation: an integrative approach. *Molecular Ecology* **23**: 513-521.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**: 557–564.

Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD. 2012. Natural selection in gene dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular Biology and Evolution* **29**: 675-687.

Foll M, 2010. *Bayescan v2.0 User Manual*. Bern, Switzerland: Swiss Institute of Bioinformatics, Universität Bern.

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977-993.

Grattapaglia D, Bradshaw HD. 1994. Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Canadian Journal of Forest Research* **24**: 1074-1078.

Grattapaglia D, Kirst M. 2008. Tansley review: *Eucalyptus* applied genomics: from gene sequences to breeding tools. *New Phytologist* **179**: 911-929.

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Kulheim C, Potts BM, Myburg AA. 2012. Progress in *Myrtaceae* genomics: *Eucalyptus* as the pivotal genus. *Tree Genetics & Genomes* **8**: 463-508.

Griffin AR, Burgess IP, Wolf L. 1988. Patterns of natural and manipulated hybridization in the genus *Eucalyptus* L'Herit.: a review. *Australian Journal of Botany* **36**: 41-66.

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

Harr B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Research* **16**: 730-737.

Hudson CJ, Kullan ARK, Freeman JS, Faria D, Grattapaglia D, Kilian A, Myburg A, Potts BM, Vaillancourt RE. 2012. High synteny and colinearity among *Eucalyptus* genomes revealed by high-density comparative genetic mapping. *Tree Genetics & Genomes* **8**: 339-352.

Hurst LD, Pál C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics* **5**: 299-310.

Iglesias-Trabado G, Wilstermann D 2008. *Eucalyptus universalis*. Global cultivated eucalypt forests map 2008. Version 1.0.1.In. *GIT Forestry Consulting's EUCALYPTOLOGICS.*

Kremer A, Caron H, Cavers S, Colpaert N, Gheysen G, Gribel R, Lemes M, Lowe AJ, Margis R, Navarro C, Salgueiro F. 2005. Monitoring genetic diversity in tropical trees with multilocus dominant markers. *Heredity* **95**: 274-280.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**: 981-994.

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

McKinnon GE, Vaillancourt RE, Steane DA, Potts BM. 2004. The rare silver gum, *Eucalyptus cordata*, is leaving its trace in the organellar gene pool of *Eucalyptus globulus*. *Molecular Ecology* **13**: 3751-3762.

Meirmans PG, Hedrick PW. 2011. Assessing population structure: F-ST and related measures. *Molecular Ecology Resources* **11**: 5-18.

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.

Myburg AA, Vogl C, Griffin AR, Sederoff RR, Whetten RW. 2004. Genetics of postzygotic isolation in *Eucalyptus*: Whole-genome analysis of barriers to introgression in a wide interspecific cross of *Eucalyptus grandis* and *E. globulus*. *Genetics* **166**: 1405-1418.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R *et al*. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**: e196.

Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1587): 332-342.

Nosil P, Funk DJ, Ortiz-Barrientos D. 2009. Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**: 375-402.

Ohno S. 1970. Evolution by gene duplication. New York, NY, USA: Springer-Verlag.

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 185-205.

Pannell JR, Fields PD. 2014. Evolution in subdivided plant populations: concepts, recent advances and future directions. *New Phytologist* **201**: 417-432.

Payn K, Dvorak W, Janse B, Myburg A. 2008. Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*; endemic to seven islands in eastern Indonesia. *Tree Genetics & Genomes* **4**: 519-530.

Payseur BA, Nachman MW. 2002. Gene density and human nucleotide polymorphism. *Molecular Biology and Evolution* **19**: 336-340.

Peakall R, Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**: 288-295.

Perez-Figueroa A, Garcia A-Pereira MJ, Saura M, Roian-Alvarez E, Caballero A. 2010. Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology* **23**: 2267-2276.

Petroli CD, Sansaloni CP, Carling J, Steane DA, Vaillancourt RE, Myburg AA, da Silva OB, Jr., Pappas GJ, Jr., Kilian A, Grattapaglia D. 2012. Genomic characterization

of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus g*enome. *PLoS ONE* **7**: e44684.

Potts BM, Dungey HS. 2004. Interspecific hybridization of *Eucalyptus:* key issues for breeders and geneticists *New Forests* **27**: 115-138.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**: R208-R215.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications* **4:1827**: DOI:10.1038/ncomms2833.

Rieseberg LH, Blackman BK. 2010. Speciation genes in plants. *Annals of Botany* **106**: 439-455.

Roesti M, Hendry AP, Salzburger W, Berner D. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology* **21**: 2852-2862.

Rymer PD, Manning JC, Goldblatt P, Powell MP, Savolainen V. 2010. Evidence of recent and continuous speciation in a biodiversity hotspot: a population genetic approach in southern African gladioli (*Gladiolus*; *Iridaceae*). *Molecular Ecology* **19**: 4765-4782.

Sansaloni C, Petroli C, Carling J, Hudson C, Steane D, Myburg A, Grattapaglia D, Vaillancourt R, Kilian A. 2010. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. *Plant Methods* **6**: 16.

SAS Institute Inc. 2010. SAS 9.2 for windows. SAS Institute, Cary, North Carolina, USA.

Savolainen V, Anstett M-C, Lexer C, Hutton I, Clarkson JJ, Norup MV, Powell MP, Springate D, Salamin N, Baker WJ. 2006. Sympatric speciation in palms on an oceanic island. *Nature* **441**: 210-213.

Scotti-Saintagne C, Mariette S, Porth I, Goicoechea PG, Barreneche T, Bodenes C, Burg K, Kremer A. 2004. Genome scanning for interspecific differentiation between two closely related oak species (*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.). *Genetics* **168**: 1615-1626.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre G-P, Bank C, Brannstrom A *et al*. 2014. Genomics and the origin of species. *Nature Reviews Genetics* **15**: 176-192.

Slotte T. 2014. The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics* **13**: 268-275.

Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE. 2011. Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (*Myrtaceae*) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics and Evolution* **59**: 206-224.

Steane DA, Potts BM, McLean E, Prober SM, Stock WD, Vaillancourt RE, Byrne M. 2014. Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology* **23**: 2500-2513.

Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**: 671-688.

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. 2012. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**: 364-373.

Tibbits J, Spokevicius A, Rigault P. submitted. Genome size difference between *Eucalyptus globulus* Labill. (X46) and *E. grandis* (Brasuz1) is explained by numerous insertions and deletions and not TE proliferation. *New Phytologist*.

Turner TL, von Wettberg EJ, Nuzhdin SV. 2008. Genomic analysis of differentiation between soil types reveals candidate genes for local adaptation in *Arabidopsis lyrata*. *PLoS ONE* **3**: e3183.

Wang J, Tian L, Lee H-S, Chen ZJ. 2006. Nonadditive regulation of *FRI* and *FLC* loci mediates flowering-time variation in *Arabidopsis* allopolyploids. *Genetics* **173**: 965-974.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Science of the United States of America* **110**: E1743-E1751.

Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Research* **21**: 725-733.

Zhou L, Zhang J, Yan J, Song R. 2011. Two transposable element insertions are causative mutations for the major domestication gene *teosinte branched1* in modern maize. *Cell Research* **21**: 1267-1270.