New Phytologist

Research

Methods

A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species

Orzenil B. Silva-Junior[1,2], Danielle A. Faria[3] and Dario Grattapaglia[2,3]

[1]Laboratório de Bioinformática, EMBRAPA Recursos Genéticos e Biotecnologia, PqEB, 70770-970 Brasilia, DF, Brazil; [2]Programa de Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, SGAN 916, 70790-160 Brasilia, DF, Brazil; [3]Laboratório de Genética Vegetal, EMBRAPA Recursos Genéticos e Biotecnologia, PqEB, 70770-970 Brasilia, DF, Brazil

Author for correspondence:
Dario Grattapaglia
Tel: +55 61 34484652
Email: dario.grattapaglia@embrapa.br

Received: *22 August 2014*
Accepted: *2 January 2015*

New Phytologist (2015)
doi: 10.1111/nph.13322

Key words: *Myrtaceae*, pooled resequencing, population structure, single-nucleotide polymorphism (SNP) ascertainment bias, trans-species SNPs.

Summary

- We used whole genome resequencing of pooled individuals to develop a high-density single-nucleotide polymorphism (SNP) chip for *Eucalyptus*. Genomes of 240 trees of 12 species were sequenced at 3.5× each, and 46 997 586 raw SNP variants were subject to multivariable filtering metrics toward a multispecies, genome-wide distributed chip content.
- Of the 60 904 SNPs on the chip, 59 222 were genotyped and 51 204 were polymorphic across 14 *Eucalyptus* species, providing a 96% genome-wide coverage with 1 SNP/12–20 kb, and 47 069 SNPs at ≤ 10 kb from 30 444 of the 33 917 genes in the *Eucalyptus* genome.
- Given the EUChip60K multi-species genotyping flexibility, we show that both the sample size and taxonomic composition of cluster files impact heterozygous call specificity and sensitivity by benchmarking against 'gold standard' genotypes derived from deeply sequenced individual tree genomes. Thousands of SNPs were shared across species, likely representing ancient variants arisen before the split of these taxa, hinting to a recent eucalypt radiation. We show that the variable SNP filtering constraints allowed coverage of the entire site frequency spectrum, mitigating SNP ascertainment bias.
- The EUChip60K represents an outstanding tool with which to address population genomics questions in *Eucalyptus* and to empower genomic selection, GWAS and the broader study of complex trait variation in eucalypts.

Introduction

High-throughput, high-precision, low-cost genotyping systems constitute an essential tool to understand the patterns and dynamics of genetic variation in natural populations, and advance selective breeding of domesticated plants and animals. Genome-wide SNP genotyping has become more accessible in recent years, due to dramatic progress in large-scale next generation sequencing (NGS) and dropping prices from increased competition among array platforms. Discovery of SNPs was initially performed on EST libraries (Novaes *et al.*, 2008) but soon moved to using different reduced genomic representation strategies either based on restriction enzymes, such as RAD sequencing or genotyping by sequencing (GbS) (Davey *et al.*, 2011), or selective sequence capture on arrays or in solution (Mamanova *et al.*, 2010). Such methods have also been useful for sequence-based genotyping, mostly in inbred crops, where genotype imputation of the large proportions of missing data (Poland & Rife, 2012; Glaubitz *et al.*, 2014), resulting from the inherent limitations of GbS methods (Miller *et al.*, 2012), can be performed. For large-scale SNP genotyping of highly heterozygous genomes with rare identical by descent (IBD) segments, no reference haplotypes are available and thus little room exists for imputation. With current technologies, heterozygous genomes therefore require a much higher sequence depth to reach acceptable marker call rates and genotype accuracy (Beissinger *et al.*, 2013; Schilling *et al.*, 2014), and more so when trying to genotype the same SNPs across species, challenging the cost effectiveness publicized for such sequence-based genotyping methods.

The eucalypts are the most widely planted hardwoods in the world due to their outstanding ability to adapt, grow and provide quality wood for multiple applications (Myburg *et al.*, 2007). Amongst the now catalogued 894 taxa of *Eucalyptus* L'Hér. (*Myrtaceae*), the 'Big Nine' species of subgenus *Symphyomyrtus* account for > 95% of the world's planted eucalypts (Harwood, 2011). These include *E. grandis, E. urophylla, E. saligna* and *E. pellita*, members of section *Latoangulatae*, broadly planted in tropical areas due to their fast growth and disease resistance; *E. globulus, E. nitens* and *E. dunnii*, members of section *Maidenaria,* species of choice in temperate regions with distinctive wood chemical and physical properties for pulp production; and *E. camaldulensis* and *E. tereticornis*, members of section

Exsertaria, known for their drought tolerance and rapid growth (Myburg *et al.*, 2007). The wide intra- and interspecific diversity and sexual compatibility across species of *Symphyomyrtus* has been a major advantage to breeders. Blending of independently evolved gene pools by interspecifc hybridization and backcrosses has resulted in highly adapted hybrid planting material (Grattapaglia & Kirst, 2008). The extensive opportunities to exploit both intra- and interspecific variation has posed an additional requirement on genotyping technologies in support of population genetics studies and breeding practice. Not only do markers have to provide robust performance in a single species, but they are also expected to be informative across a wider phylogenetic range. By and large microsatellites were found to have this attribute and became the main working tool for genetic analysis of *Eucalyptus* (Grattapaglia *et al.*, 2012). In recent years, however, hybridization-based DArT arrays for eucalypts have supplied on average 3000 informative markers across species, increasing by an order of magnitude the ability to query polymorphisms across the genome (Sansaloni *et al.*, 2010). This has inaugurated genome-wide mapping (Hudson *et al.*, 2012; Petroli *et al.*, 2012), QTL detection (Freeman *et al.*, 2013), association genetics (Cappa *et al.*, 2013) and genomic selection in eucalypts (Resende *et al.*, 2012). Nevertheless, the dominant behavior of DArT markers, together with limitations in expanding their number and distributing them equally across the genome, became concerns when considering large-scale, fast turnaround genotyping for operational molecular breeding.

Our initial SNP assay development in eucalypts clearly showed that a large number of SNPs can be successfully genotyped in a genome with high nucleotide diversity, given that systematic SNP discovery with sequence context ascertainment is adopted (Grattapaglia *et al.*, 2011b). We also concluded that the development of a much larger array of informative SNPs across multiple *Eucalyptus* species would require a large and representative collection of sequences from all target species. With the recent improvements in sequencing yields, it has now become possible to move beyond reduced genomic representations for SNP discovery in moderately sized plant genomes (< 1 Gbp). Shallow resequencing data of whole genomes of several individuals in pools, allows capture of low-frequency variants (Druley *et al.*, 2009; Marroni *et al.*, 2011) and enables good estimates of allele frequencies (Gautier *et al.*, 2013; Schlotterer *et al.*, 2014). Following recalibration of base quality to eliminate the inherent biases of NGS (DePristo *et al.*, 2011), this approach may mitigate SNP ascertainment bias, maximize genome coverage and avoid gaps or biases derived from the unequal distribution of restriction enzyme-cut sites or capture probes. Parallel to improvements in sequencing technologies, genotyping of thousands of SNPs for thousands of samples has now become considerably more accessible given flexible multiplex levels and chip construction formats. In routine analyses such as those required for operational genomic selection in tree breeding, such 'SNP chip' platforms are currently the only ones that meet the requirements of high data reproducibility (> 99%) across independent experiments and laboratories. Furthermore, data for the exact same set of SNP markers can be easily shared across independent studies, enabling meta-analyses to be carried out effortlessly; such a task would be difficult or likely impossible to carry out with sequence-based genotyping data in highly heterozygous genomes. The clear advantages of SNP chips have boosted their development for the major grain (Ganal *et al.*, 2011; Bekele *et al.*, 2013; Song *et al.*, 2013; Wang *et al.*, 2014), vegetable (Felcher *et al.*, 2012; Sim *et al.*, 2012), fruit (Chagne *et al.*, 2012; Verde *et al.*, 2012) and tree (Chancerel *et al.*, 2013; Geraldes *et al.*, 2013; Pavy *et al.*, 2013) species, following those built for the main domestic animals (Matukumalli *et al.*, 2009; Ramos *et al.*, 2009; Groenen *et al.*, 2011; Tosser-Klopp *et al.*, 2014).

After the landmark publication of the *Eucalyptus grandis* genome sequence (Myburg *et al.*, 2014), we reasoned that the development of a high throughput SNP genotyping platform for the eucalypts would be a key contribution to enhance both basic and applied genetics research for species of the genus. Besides facilitating low cost, high marker density, polymorphism and speed of data generation, the platform would satidfy the essential requirements of high genotype call accuracy and reproducibility, and full public access to the SNP content. Furthermore, to make the chip widely useful to the breadth of the international community, we set forth an additional goal, not yet deliberately achieved in any other plant or animal species, to the best of our knowledge. A truly useful eucalypt SNP chip had to provide high-quality, genome-wide data for all 'big nine' *Eucalyptus* species and possibly for related *Myrtaceae* taxa. In this work we describe the experimental and analytical steps taken to reach these goals. We successfully developed the EUChip60K, a multi-species *Eucalyptus* chip with 59 222 SNPs. It provides data for thousands of SNPs in all the major eucalypt species and some related taxa. We carried out a detailed analysis of the accuracy of genotype calling as a function of the composition of the cluster file used to call genotypes by benchmarking against 'gold standard' genotypes, and assessed the level of ascertainment bias of the chip content. This large validated SNP collection provides a powerful tool for molecular breeding and population genetics investigation within and across *Eucalyptus* species.

Materials and Methods

Plant material and DNA samples

Twenty genomic DNA pools composed of 12 unrelated individuals each, totaling 240 genetically unrelated trees, were assembled to provide a representative sample of the main plantation species of *Eucalyptus* covering three subgenera and the related genus *Corymbia*. Each pool was composed of equimolar amounts of DNA as measured using a Q-bit 2.0 (Invitrogen). A single library for each DNA pool was prepared and shotgun sequenced (2 × 100) across 20 lanes in three different flow cells on a Hiseq 2000. Sequencing was carried out at the Biotechnology Center of the University of Illinois Urbana-Champaign (USA). Subgenus *Symphyomyrtus*, the largest one in the genus *Eucalyptus* was the most highly represented with 10 species. Samples of *E. pilularis* (subgenus *Eucalyptus*), *E. cloeziana* (subgenus *Idiogenes*) and the

related genus *Corymbia* were also sequenced. The 240 trees sequenced and the 1498 trees later genotyped for chip validation, were sampled either from wild or elite breeding populations (Supporting Information Table S1).

SNP discovery and ascertainment

Version 1.1 of the *Eucalyptus grandis* genome sequence (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Egrandis) was used as for read alignment. Sequence data were subject to an analytical pipeline involving specialized softwares (Fig. S1). GATK (McKenna *et al.*, 2010) was used for simultaneous SNP variant discovery and genotyping using datasets of five flagship *Eucalyptus* species (GRA, GLO, URO, CAM, TER) (Table S1). Variant quality score recalibration (DePristo *et al.*, 2011) was carried out using a high confidence set of 162 244 SNP variants derived from ultra-deep RAD-Sequencing (Grattapaglia *et al.*, 2011a). A 90.0 truth sensitivity threshold was selected by inspection of the expected Ti/Tv ratio and VQSLOD score and likely false positives SNPs marked to be filtered. SNP variants were annotated with Vcftools (Danecek *et al.*, 2011) to mark clusters of two or more SNPs in a 30-bp sequence run. Genotype filters were built using GATK SelectVariants tool to retain variant SNPs displaying polymorphism in the largest number of species. To be retained, a SNP had to provide accurate genotype call supported by >4 high-quality reads of each alternative base in each species. SNP variants were assigned scores into five categories used to prioritize the final set of SNPs (Methods S1).

Genotype data acquisition

EUChip60K manufacture and intensity data (.idat files) were obtained through GENESEEK (Lincoln, NE, USA). SNP genotypes were called using GenomeStudio (Illumina Inc., San Diego, CA, USA) following the standard genotyping and quality control procedures (Illumina, 2010). Poorly performing samples identified based on scatter plot analysis with call rate $<90\%$ and GenCall10 <0.2 were excluded. SNPs were re-clustered using 100 retained samples and filtered for quality. Only SNPs that passed the following multi-variable metrics criteria were retained: (1) $\geq 80\%$ samples with GenCall >0.15; (2) genotype clusters separation >0.3; (3) mean normalized intensity (R) value of the heterozygote cluster >0.2; (4) mean normalized theta of the heterozygote cluster between 0.2 and 0.8; (5) Mendelian allelic inheritance concordance $>95\%$ (one inconsistency allowed in 24 allelic transmissions in 12 parent-offspring tests); (6) 100% reproducibility across four replicated samples. SNP that did not pass these cutoff criteria were zeroed from further analyses. No manual editing of clusters was attempted to avoid introduction of subjective bias into the dataset. SNP and sample statistics following this procedure were: average SNP call frequency across all samples $>90\%$ and sample call rates across all SNPs $>97\%$. The cluster file built following these SNP quality filtering steps was exported and used in the subsequent analyses. These same stringent filtering steps to eliminate poorly performing samples and SNPs were used to build all species-specific, section-specific and multi-section cluster files.

Validation of SNP chip genotypes against sequence-based 'gold standard' SNP genotypes

A key aspect of the overall performance of chip-based SNP genotyping is the task of deriving genotype classes by clustering raw intensity data. A training set of reference samples for the species or population to be genotyped is used to generate a cluster file, and this file is then used to genotype samples coming from that species or population. Although for human SNPs genotyping the recommendation to build a cluster file is to use *c.* 100 samples (Illumina, 2010), no recommendations exist for genomes with much higher nucleotide diversity. Given the multi-species genotyping flexibility intended for the *Eucalyptus* chip, we therefore assessed the impact of the sample size and taxonomic composition of the DNA samples used to build the cluster files. Three validation parameters were used: the number of SNPs genotyped, the genotype call rate and the accuracy of the declared SNP genotypes. Validation of these three parameters was performed by comparing chip-based SNP genotypes to their correspondent sequence-based SNP genotypes, herein adopted as 'gold standard'. These 'gold standard' SNP genotypes were obtained from the analysis of deep ($\geq \times 20$) whole-genome resequencing data of seven *E. grandis* individuals (Methods S2).

SNP conversion and genome annotation

SNP conversion rates, that is, the overall call frequency across samples, with a minimum set at $\geq 90\%$ and the Minimum Allele Frequency (MAF) set at ≥ 0.01, were estimated using a validation panel of 1498 genetically unrelated samples of 14 *Eucalyptus* species, three hybrid breeding populations and two species of different genera of *Myrtaceae*, *Corymbia* and *Psidium*. Converted SNP loci were annotated based on their genomic most probable placement, in order to categorize the effects resulting from base substitutions as compared to the annotated reference genome using SnpEff (Cingolani *et al.*, 2012). A variant file in VCF was built based on the EUChip60K manifest file and used as input to SnpEff with the *Eucalyptus grandis* annotation v2.1 downloaded from the SnpEff website.

Assessment of the EUChip60K ascertainment bias (AB)

In order to assess the extent of AB of the EUChip60K we compared its site frequency spectrum(SFS) to the whole-genome SFS derived from the pooled sequencing data of 36 *E. grandis* individuals (72 chromosomes), a sample size close to the adequate range for good estimates of allele frequency (Schlotterer *et al.*, 2014). To estimate the SFS of the whole-genome pooled sample, we ran SNAPE-POOLED (Raineri *et al.*, 2012) on a pileup formatted file built from the alignments of the sequences on the *Eucalyptus grandis* genome using *samtools mpileup.* SNAPE-POOLED options used were: divergence 0.01, prior nucleotide diversity 0.02 and folded spectrum, using the Bayesian method to

compute the posterior distribution of allele frequency based on a flat prior. A posterior probability that alleles are actually segregating at a particular site was calculated and a threshold of $P \geq 0.9$ was used to retain SNPs in the pool. The extent of AB of the SFS of the EUChip60k was assessed by testing the equality of the one-dimensional distribution of SNP proportions in each 0.025 MAF class above MAF> 0.05 against the SFS of the whole-genome SNP set using a nonparametric Kolmogorov–Smirnov (KS) test. Additionally a KS test was used to test the equality of the distributions of the direct SNP counts between the EUChip60K and the average SNP counts of 1000 equally chip-sized samples of SNPs randomly extracted from the entire whole-genome SNP set.

SNP-based population structure analysis

Population samples of unrelated trees of two provenances each for *E. grandis* ($n = 23$ for Atherton and $n = 23$ for Coffs Harbor), *E. globulus* ($n = 22$ for Jeeralang and $n = 12$ for Flinders Island) and *E. camaldulensis* ($n = 10$ for Walsh River and $n = 12$ for Kennedy River) were used for population structure analyses. For *E. urophylla* a set of $n = 12$ from Flores Island (Indonesia) and 24 elite breeding trees considered as pure *E. urophylla* were also studied. To assess the EUChip60K information content for species and provenance differentiation, F_{st} was estimated for all genotyped SNPs with call rate > 95% and MAF> 0.01. The power of the EUChip60K SNP content to assign individual trees to species and provenances was assessed using STRUCTURE v2.3.1 (Pritchard *et al.*, 2000) using a subset of 600 evenly spaced SNPs randomly taken at a rate of 1 SNP/Mb. STRUCTURE was run applying a burn-in period of 100 000 and 200 000 iterations for data collection with K ranging from two to eight inferred clusters, performed with ten independent runs each. The most probable value of K was defined by ΔK (Evanno *et al.*, 2005) and displays of population structure were implemented using Structure Harvester (Earl & Vonholdt, 2012).

Results

Sequence analysis and SNP selection for chip manufacture

A total of *c.* 920.9 billion bp of raw sequence data (1522 *E. grandis* genome equivalents) were obtained, and 511.9 billion high-quality bases (846 genome equivalents) aligned for SNP discovery (Table 1). Average aligned sequence coverage for each one of the 240 trees was *c.* ×3.5, and coverage per species varied from *c.* ×42 for secondary species up to ×84 for *E. globulus* and ×210 for *E. grandis* (Table S1). A total of 46 997 586 raw SNP variants were discovered and following variant recalibration with 162 244 high-confidence SNPs, 20 043 471 SNPs were retained and submitted to flanking sequence filtering, a criterion earlier found to significantly drive successful SNP conversion (Grattapaglia *et al.*, 2011b). This filtering step discarded 89% of the SNPs, retaining 2 247 471 SNPs polymorphic both between and within species – that is, including those fixed within any one species but polymorphic in relation to another. When a variant position was required with no additional SNPs within a 30-bp flanking window, only 20.5% of the SNPs were retained (461 028), and only 60 523 of them had the Illumina-recommended 60-bp flanking window free of SNPs. As additional SNP filtering steps were to come, we had to keep all 461 028 SNPs, out of which 208 834 were type II Infinium SNPs (A/C; A/G; T/C; T/G), which require a single bead type, making the chip more cost-effective, and 194 152 of them had an Illumina Assay Design Tool (ADT) score ≥ 0.6. When the final filtering parameter of genome-wide distribution was applied (≥ 3 SNPs for every 43.5 kb genomic bin), only 52 402 SNPs were retained. To meet the target of having a final set of 60 000 assayed SNPs on the chip, we therefore had to allow the inclusion of 18 080 SNPs that failed the soft GATK recalibration filter but were still classified as D30 and were polymorphic within species. A total of 70 482 SNPs were eventually selected, filling 13 421 genomic bins, (96% bin coverage) with only 692 bins (4%) left empty, most of them corresponding to regions of still undetermined sequence (long stretches of Ns) in the *Eucalyptus* genome assembly. For the 4518 genomic bins (32%) where only one SNP could be selected following the filtering criteria, the SNP was listed twice for manufacture to increase its probability of eventually being present on the chip. Thus, a final list of 70 482 unique SNPs and 4518 replicated SNPs, totaling 75 000 SNPs, was sent to chip manufacture (Table 1).

Table 1 Summary of the number of single-nucleotide polymorphisms (SNPs) recovered following the consecutive analytical steps aimed at discovery, ascertainment and selection of SNPs, to populate and optimize the construction of the EUChip60K

Total bases of raw sequence bases used for SNP discovery	920 872 681 000
Total high quality aligned bases used for SNP discovery	511 906 738 694
Raw SNP variants discovered using GATK default parameters	46 997 586
SNPs that passed GATK score following recalibration with RAD sequencing data at ts = 90.00 (E category)	20 043 471
SNPs with ≥ 30 bp flanking windows free of secondary SNP (D30 score)	2247 480
Polymorphic SNP in ≥ 1 species with D30 score	461 028
Infinium II SNPs (only A/G, A/C, T/G and T/C variants)	208 834
Illumina Infinium Assay Design Tool SNPs with (ADT) score ≥ 0.6	194 152
Unique SNPs selected for chip manufacture	70 482
Replicated SNPs targeting genomic bins that had only 1 SNP/bin	4518
Total number of SNPs sent to chip manufacture	75 000
SNP effectively manufactured on the chip (includes 3735 replicated SNPs)	64 639
Unique SNPs effectively manufactured on the chip	60 904
Successfully genotyped SNPs in a multi-species panel of samples	59 222
Converted SNPs (MAF ≥ 0.01) within one or more 14 *Eucalyptus* species	51 204
Converted SNPs located at ≤ 10 kb from 30 444 annotated gene models	47 069
Converted SNPs located inside 14 116 annotated gene models	26 346

Following chip manufacture steps the four last entries correspond to the statistics of SNP validation and conversion in the 14 *Eucalyptus* species studied.

Optimizing the size and taxonomic composition of genotyping cluster files

We tested the impact of using progressively smaller training sample sets ($n = 100$, 90, 80, etc. down to 10) to generate a cluster file, using *E. grandis* as a test case for which we had 'gold standard' samples as a benchmark. A reduction of the total number of SNPs genotyped was observed as more samples were used to build cluster files, going from *c.* 56 000 down to *c.* 50 000. However, both the SNP call rate and the genotype concordance rate with the 'gold standards' increased as more samples were included, and reached a maximum at 97% and 94.4%, respectively, when $n = 60$ samples were used to train the clustering algorithm, and did not improve significantly afterwards. These results indicated that for a high-diversity genome a sample size of 60–70 appears to be optimal to build cluster files that provide the best compromise between the number of SNPs genotyped, the genotype call rate and genotype accuracy (Fig. 1a). Furthermore, the analysis of genotype discordances showed that the number of discordant genotypes in the homozygous (sequence) vs heterozygous (chip) comparison decreased by 42% when going from $n = 10$ to $n = 70$ samples with little reduction afterwards, whereas the other two genotype comparisons showed only a small change with increasing sample size (Table 2). Based on these results we adopted a sample size $n = 60$ whenever available, to build cluster files when validating SNPs for all other species studied. Given that the chip is intended to genotype additional *Eucalyptus* species besides the ones contemplated in this study, we also evaluated the effect of the taxonomic composition of the cluster file. Results showed that the *E. grandis*-specific, *Latoangulatae*-specific and multi-section cluster files provided essentially the same call rates and overall genotype concordance in *E. grandis*. When cluster files built with samples of other sections were used, genotype call rates were only slightly reduced, from 97% to 94%, although a small loss of *c.* 1000 SNPs was observed (Fig. 1b; Table 3). Heterozygous call specificity and sensitivity showed only a minor change when cluster files with different taxonomic compositions were used (Table S2).

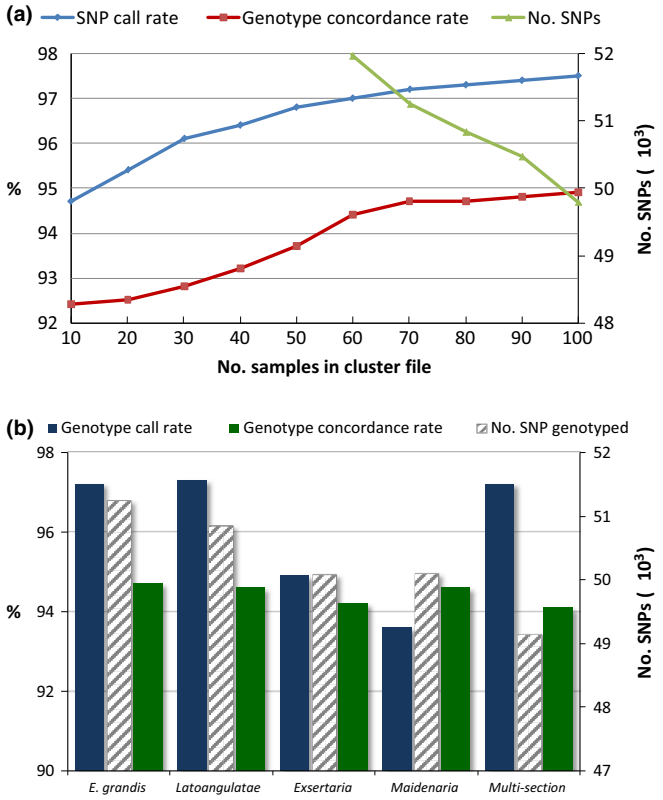

Fig. 1 (a) Impact of the number of samples used to build a genotyping cluster file on the average call rate and genotype concordance (left *y*-axis) and total number of single-nucleotide polymorphisms (SNPs) genotyped (right *y*-axis). (b) Impact of the cluster file samples' genetic composition on average SNP call rate, genotype concordance and total number of SNPs genotyped (right *y*-axis) in the 'gold standard' *Eucalyptus grandis* samples (see the Materials and Methods section).

EUChip60K genotyping performance across *Eucalyptus* species and genera of *Myrtaceae*

Of the 75 000 total SNPs sent for chip manufacture, 64 639 SNPs were effectively assembled (60 904 unique and 3735 replicated), that is, 86% of the total sent; this is well within the

Table 2 Statistics of the impact of the number of samples used to build a cluster file on the numbers of single-nucleotide polymorphisms (SNPs) genotyped, genotype call and genotype concordance rates provided by the EUChip60K for *Eucalyptus grandis*

No. samples	No. SNP genotyped	Genotype call rate (%)	No. SNPs compared	No. genotypes compared	No. discordant genotypes Het (seq) × Homo (chip)	No. discordant genotypes Homo (seq) × Het (chip)	No. discordant genotypes Homo (seq) × Homo (chip)	Concordance rate (%)
10	56 907	94.7	27 248	181 526	3367	8557	1911	92.4
20	56 631	95.4	27 218	182 313	3116	8602	1923	92.5
30	56 002	96.1	27 003	181 625	2851	7920	2231	92.8
40	55 065	96.4	26 681	179 930	2773	7176	2308	93.2
50	53 764	96.8	26 253	177 623	2664	6072	2368	93.7
60	51 958	97.0	25 804	174 886	2573	4912	2310	94.4
70	51 248	97.2	25 566	173 560	2447	4564	2249	94.7
80	50 826	97.3	25 407	172 665	2384	4493	2197	94.7
90	50 459	97.4	25 250	171 778	2329	4381	2175	94.8
100	49 794	97.5	24 938	169 872	2274	4193	2192	94.9

The number of SNPs and genotypes that were compared between the chip and the 'gold standard' sequence-based data are listed, as well as the numbers of discordant genotypes in the three possible comparisons.

New Phytologist

Table 3 Statistics of the impact of genetic composition of the samples used to build a cluster file on the numbers of single-nucleotide polymorphisms (SNPs) genotyped, genotype call and genotype concordance rates provided by the EUChip60K for *Eucalyptus grandis*

Genetic composition	No. SNP genotyped	Genotype call rate	No. SNP compared	No. genotypes compared	No. discordant SNP Het (seq) × Homo (chip)	No. discordant SNP Homo (seq) × Het (chip)	No. discordant SNP Homo (seq) × Homo (chip)	Concordance rate (%)
E. grandis	51 248	97.2	25 566	173 560	2447	4564	2249	94.7
Latoangulatae	50 852	97.3	25 345	163 065	2338	4939	2002	94.6
Multi-taxa	49 152	97.2	24 530	156 794	2083	6288	1496	94.1
Exsertaria	50 074	94.9	24 518	155 376	2196	5715	1574	94.2
Maidenaria	50 096	93.6	24 603	154 967	2243	5040	1640	94.6

The number of SNPs and genotypes that were compared between the chip and the 'gold standard' sequence-based data are listed, as well as the numbers of discordant genotypes in the three possible comparisons.

Illumina announced rates, consistent with the random nature of the assembly of beads into wells and the redundancy threshold required for each SNP to pass Illumina manufacture QC. Out of the 60 904 unique SNPs (Table S3), 59 222 (97.2%) passed the multi-variable metrics criteria, and 51 204 were polymorphic within one or more of the 14 *Eucalyptus* species, a conversion rate of 84.1% (Table S4). The 51 204 species-wide converted SNPs provide an average genome-wide coverage of 1 SNP every 11.8 kb. At the section level the numbers of converted SNPs in the validation samples were 37 932 for *Latoangulatae*, 37 793 for *Exsertaria* and 29 334 for *Maidenaria*, with corresponding average densities of 1 SNP every 16 kb for the first two sections and every 20 kb for *Maidenaria* (Fig. 2). More interestingly, however, is the fact that 90% of the effective inter-marker distances were smaller than 10 kb (Fig. S2A). Besides a homogeneous genome coverage, the EUChip60K efficiently interrogates variation in the gene space of the *Eucalyptus* genome as revealed by a strong and significant linear correlation ($r = 0.782$; $P < 0.0000$) between the number of converted SNPs and the number of genes in every 500-kb genomic segment (Fig. S2B), while supplying 47 069 SNPs located at ≤ 10 kb from 30 444 of the 33 917 (89.7%) annotated gene models in the 11 *Eucalyptus* chromosomes and 26 343 of them inside 14 166 genes (Table 1).

The EUChip60K consistently genotyped on average of 53 031 to 56 022 unique SNPs in all 12 species of *Symphyomyrtus* (Table 4), with average 99.1% reproducibility between replicated SNPs on the chip (Table S5). For species of different *Eucalyptus* subgenera (*E. pilularis* and *E. cloeziana*) between 46 000 and 48 000 SNPs were genotyped. For phylogenetically more distant taxa, the numbers of successfully genotyped SNPs were considerably lower (27 876 for *Corymbia* and 18 006 for *Psidium*) but still satisfactory considering that these species were not included in

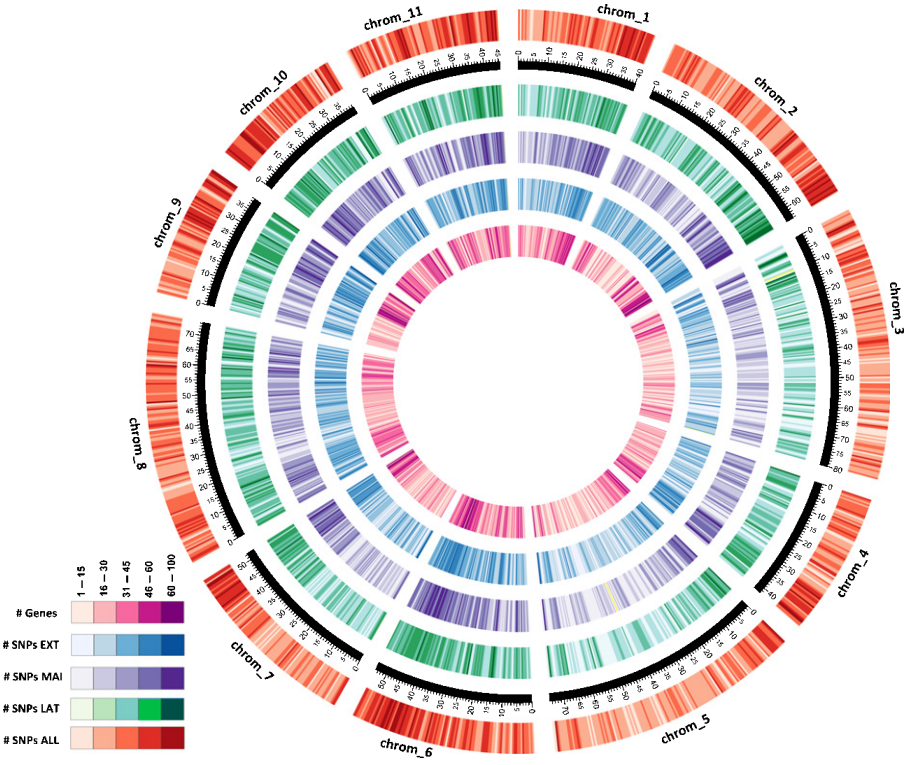

Fig. 2 Heatmap of the density distribution of the 51 204 converted single-nucleotide polymorphisms (SNPs) along the 11 *Eucalyptus* chromosomes in the three main phylogenetic sections of subgenus *Symphyomyrtus* (LAT, *Latoangulatae*; MAI, *Maidenaria*; EXT, *Exsertaria*) and all 14 *Eucalyptus* species evaluated, with the corresponding density of annotated gene models in the *E. grandis* reference genome.

Table 4 Genotyping performance of the EUChip60K as measured by call rate, total numbers of single-nucleotide polymorphisms (SNPs) successfully genotyped, total numbers of polymorphic SNPs (MAF > 0.01) and average observed heterozygosity across 19 taxa, including 14 pure *Eucalyptus* species, three hybrid populations and species of two different genera (*Corymbia* and *Psidium*) based on variable genetic compositions of the genotyping cluster files used to call genotypes from raw intensity data

		Species-specific cluster				Section-specific cluster				Multi-section cluster			
Taxon	N	% Call rate	No. SNP called	No. SNP MAF > 0.01	Aver.H_{obs}	% Call rate	No. SNP called	No. SNP MAF > 0.01	H_{obs}	% Call rate	No. SNP called	No. SNP MAF > 0.01	H_{obs}
E. grandis	79	98.6	51 028	30 040	0.274	98.3	52 593	31 116	0.282	98.5	50 880	30 592	0.304
E. urophylla	297	98.6	50 011	30 196	0.319	98.1	50 933	30 421	0.317	98.3	50 219	30 612	0.333
E. saligna	12	98.8	56 022	28 105	0.379	99.0	49 151	22 331	0.296	99.0	49 870	23 423	0.317
E. pellita	300	98.8	51 152	19 360	0.253	98.4	50 539	18 867	0.257	98.6	49 037	18 721	0.273
E. camaldulensis	42	98.5	52 621	39 525	0.272	98.7	49 975	37 049	0.243	98.5	50 612	37 258	0.236
E. tereticornis	12	98.0	54 118	23 726	0.412	98.4	49 076	19 347	0.352	97.4	49 686	19 647	0.336
E. brassiana	7	98.7	55 443	31 942	0.394	99.3	47 929	28 366	0.312	99.0	49 152	28 830	0.326
E. globulus	131	98.9	50 851	19 299	0.276	98.8	51 630	19 921	0.289	98.6	49 578	19 124	0.300
E. benthamii	558	98.8	51 154	12 048	0.305	98.6	51 237	11 769	0.319	98.6	48 355	11 102	0.334
E. dunnii	12	98.8	54 867	17 014	0.462	98.9	49 037	12 739	0.349	98.8	48 718	12 534	0.361
E. viminalis	12	98.5	54 859	22 238	0.386	98.4	48 960	17 716	0.271	98.3	48 854	17 756	0.284
E. nitens	12	98.7	54 248	18 172	0.360	98.3	48 183	12 674	0.231	98.2	47 899	12 860	0.243
Hybrid ARA-B	956	98.5	51 463	34 966	0.312	98.1	53 133	34 980	0.311	98.6	48 556	33 447	0.328
Hybrid ARA-C	914	98.3	50 722	34 375	0.331	97.5	53 133	35 394	0.328	98.4	48 462	33 617	0.346
Hybrid CEN	763	98.6	52 699	26 069	0.366	97.7	53 133	26 115	0.364	98.9	47 967	24 468	0.387
E. pilularis	12	98.5	48 140	6380	0.458	–	–	–	–	97.9	38 517	3266	0.292
E. cloeziana	12	99.4	46 938	4600	0.587	–	–	–	–	99.3	37 203	1864	0.398
Corymbia	12	98.0	27 876	11 785	0.139	–	–	–	–	97.3	19 220	6344	0.075
Psidium sp.	11	98.7	18 006	4024	0.205	–	–	–	–	95.1	8893	1150	0.150

the SNP discovery panel and no SNP resources exist for these 'orphan' taxa. For these distant genera, however, the use of a species-specific cluster file is mandatory to maximize the output of high-quality SNP data. The number of polymorphic SNPs varied across species and a decline in the number of polymorphic SNPs was observed when cluster files with a progressively more distant genetic composition from the target species being genotyped were used, further highlighting the importance of adopting species-specific cluster files (Table 3).

It is important to note that the numbers of called and polymorphic SNPs reported using species-specific cluster files built with < 60 individuals should be viewed with caution. Numbers of SNPs listed for *E. camaldulensis* with $n = 42$ and those for all species with $n = 12$ tend to be slightly overestimated and, more importantly, suffer from a lower genotype concordance rate as observed in a sensitivity analysis carried out for *E. grandis* (Fig. 1). *For E. camaldulensis*, whereas 52 621 SNPs were called using a species-specific cluster file built with $n = 42$ samples, only 49 975 were called when using a section-specific cluster file that involved $n = 61$ samples. A similar trend was observed for *E. nitens*, *E. dunnii* and *E. viminalis*, all of them showing a reduction from *c.* 54 000 to *c.* 48 000 SNPs (Table 4). It is expected, however, that such reductions of SNP numbers will be accompanied by an increase in genotype concordance rate, as seen for *E. grandis* (Fig. 1). For the taxa for which no section-specific cluster file could be built (*E. pilularis, E. cloeziana, Corymbia* and *Psidium*), the multi-section cluster file provides the best approximation of the EUChip60K performance, although taxa-specific cluster files and benchmarking against 'gold standard' genotypes should be used to obtain an accurate estimate of the performance of this genotyping platform for these species.

The estimates of the numbers of polymorphic SNPs provided by the EUChip60K (Table 4) should be considered as low-end estimates, reflecting the partial intraspecific variation sampled so far. As more individuals are genotyped and more intra-population and inter-provenance variation gets sampled for any particular species, it is likely that more SNPs will turn out to be polymorphic. Furthermore, the numbers of polymorphic SNPs depend on the evolutionary history of each species. For example, although 75% of the genotyped SNPs were polymorphic in the most widespread eucalypt in the Australian continent, *E. camaldulensis* (Butcher *et al.*, 2009), with only $n = 42$ individuals genotyped, only 23.5% of the SNPs were polymorphic in *E. benthamii*, with $n = 558$ genotyped, a rare species known for its restricted occurrence and genetic vulnerability (Butcher *et al.*, 2005). As expected, hybrid populations ARA-B and ARA-C, that involve up to four different species in their composition, showed the largest proportions of polymorphic SNPs, likely capturing those additional SNPs fixed within species but polymorphic in hybrid individuals. A considerably smaller proportion of polymorphic SNPs (10–13%) was observed for species of different subgenera, consistent with a decline in the rate of common variant SNP positions with increased phylogenetic distance. Interestingly, *Corymbia* showed a high (42%) proportion of polymorphic SNPs, but most of them had MAF < 0.15, a result that will require further scrutiny using a species-specific cluster file and validation using 'gold standard' sequence-based genotypes. In the joint analysis of species and sections, 42 755 SNPs were successfully genotyped and 10 079 of them were

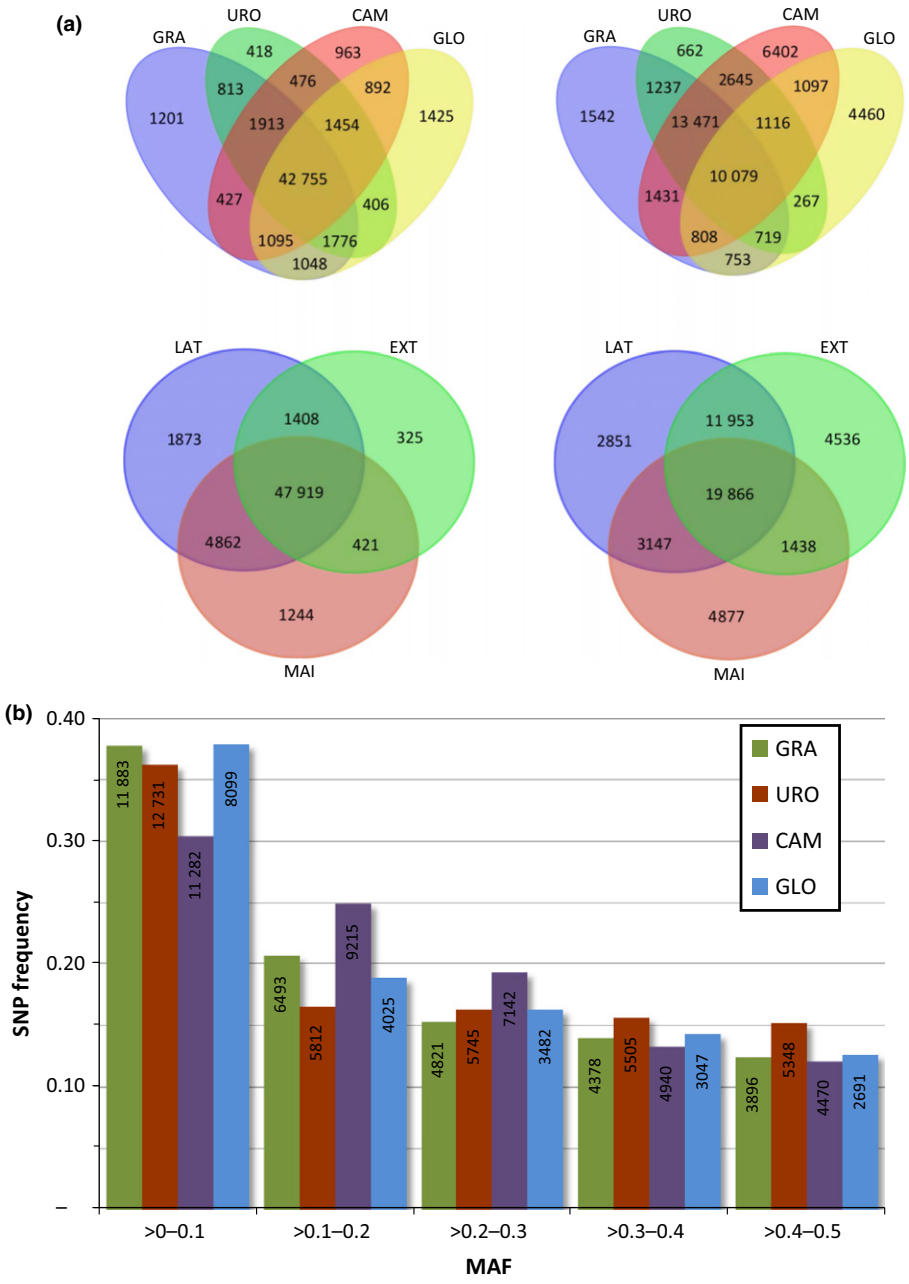

Fig. 3 (a) Venn diagrams of the successfully genotyped (call rate > 90%) (right) and converted (call rate> 90%; MAF> 0.01) (left) single-nucleotide polymorphisms (SNPs) in the four main *Eucalyptus* species (top), and three main sections (bottom) of subgenus *Symphyomyrtus*. (b) Distribution of SNP frequency into minimum allele frequency (MAF) classes in the four main *Eucalyptus* species with corresponding SNP counts inside histogram bars (GRA, *E. grandis;* URO, *E. urophylla;* GLO, *E. globulus;* CAM, *E. camaldulensis*).

simultaneously polymorphic in the four main *Eucalyptus* species. At the section level within *Symphyomyrtus*, 47 919 were genotyped and 19 890 were simultaneously polymorphic in the three sections, with several thousand more informative SNPs when any two or three species were considered concurrently (Fig. 3a). The distribution of MAF classes showed similar enrichment for rare SNPs (MAF > 0–0.1) in all four main species and decreasing frequencies of SNPs toward higher MAF (Fig. 3b).

Ascertainment bias in *E. grandis*

No significant difference was seen between the site frequency spectra of the 24 035 EUChip60K SNPs and the whole-genome set of 19 432 790 SNP (MAF> 0.05) discovered in *the E. grandis* pooled sample ($P = 0.709$; KS test). Likewise no difference was seen when the SFS comparison was carried out between direct SNP counts and the average counts of 1000 random SNP sets ($P = 0.710$) (see data in Table S6). These results and the equivalent patterns of the SFS (Fig. 4) provide evidence for a considerable reduction of the ascertainment bias when genotyping *E. grandis* with the EUChip60K, although slight differences in the SNP proportions are observed at the extremes of the distributions.

Annotation of the EUChip60K SNP effects

The snpEff annotation of the 59 222 genotyped SNP resulted in a substitution rate of 1/10 230 bp, a Ti/Tv ratio (47 166/12 056) of 3.9 with transitions corresponding to *c.* 80% of the substitutions. Such a high Ti/Tv ratio is taken as a general indication of

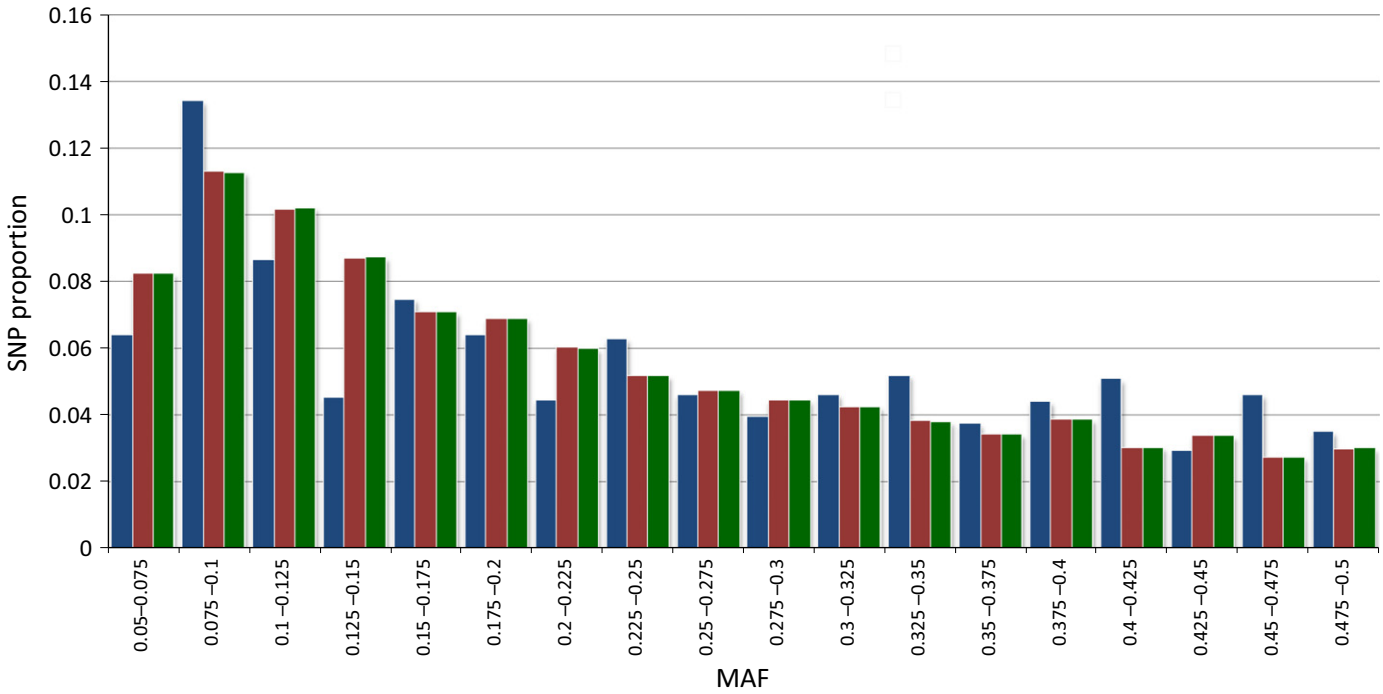

Fig. 4 Site frequency spectra (SFS) of 24 035 polymorphic single-nucleotide polymorphisms (SNPs) (minimum allele frequency (MAF) > 0.05) in *Eucalyptus grandis* genotyped with the EUChip60k (blue bars), the set of 19 432 790 SNP (MAF> 0.05) discovered in a whole-genome pooled sample of 36 *E. grandis* individuals (green bar) and the average SFS of 1000 random samples of an equivalent number of SNPs extracted from all 19.4 million SNPs (red bars). No significant difference was seen among the three SFS based on a Komolgorov-Smirnov nonparametric test (see text for details and Supporting Information Table S4 for the data).

high accuracy in variant calling (Liu *et al.*, 2012). Synonymous and nonsynonymous SNPs were 10 127 and 4610 (respectively), the latter ones leading to amino acid changes in 4046 predicted genes. These substitutions would, in turn, lead to 14 250 and 6020 predicted effects (respectively), considering that a single SNP variant may lead to multiple effects depending on the number of transcripts predicted for the gene. SNPs were also annotated in other categories. Variants annotation also was done based on genomic locations, showing a total of 26 346 SNPs located within 14 116 predicted protein coding genes and 32 876 intergenic SNPs. The total number of changes affecting the whole genome was estimated at 122 028, acknowledging that these effects were annotated assuming that all predicted genes in the genome annotation are truly protein coding genes (Table 5).

Population structure analysis

Estimates of F_{st} were obtained for a subset of 27 985 SNPs that were polymorphic (MAF > 0.05) across the consolidated set of 114 trees of the four main species in their centers of origin. Average estimates of F_{st} for the six pair-wise species comparisons were generally high as expected and consistent with their current phylogenetic standings, with GRA vs GLO showing the highest F_{st} at 0.376, followed by URO vs GLO 0.364, CAM vs GLO 0.300, CAM vs GRA 0.297, URO vs GRA 0.274, CAM vs URO 0.258. Within species a marked difference was observed in the frequency spectrum of F_{st} estimates between provenances (Fig. 5). Although the majority of SNPs displayed a low F_{st} between the two CAM provenances, the opposite was seen for GLO. The STRUCTURE analysis resolved the four species and detected a considerable hybrid composition between *E. urophylla* and *E. grandis* in the set of 24 elite breeding trees regarded as being pure *E. urophylla* (Fig. 5c). The same 600 SNPs successfully resolved the provenance variation in *E. grandis* and *E. globulus* but not in CAM, consistent with the spectrum of F_{st} estimates. A larger set of 6000 SNPs did not improve the provenance separation in CAM (data not shown), corroborating the close genetic proximity of these two provenances.

Discussion

The EUChip60K provides the highest SNP genotyping density and best genome-wide distribution for any forest tree genome to date, while matching existing high-density SNP chips developed for mainstream crops such as maize (Ganal *et al.*, 2011) and soybean (Song *et al.*, 2013), although higher density chips are rapidly becoming available for such species (Unterseer *et al.*, 2014). More importantly, however, this high throughput genotyping tool provides unprecedented flexibility to genotype multiple species of the same genus, and demonstrates the technical feasibility and advantages of deliberately developing a multi-species SNP genotyping chip based on whole-genome pooled sequencing. The success of our strategy was highly dependent on generating a large amount of whole-genome sequence data for a large, diverse and representative assembly of germplasm, coupled to a customized multi-step SNP discovery pipeline with stringent

Table 5 Prediction of the EUChip60K single-nucleotide polymorphism (SNP) variant effects in *Eucalyptus grandis* according to the terminology adopted by the SnpEff pipeline and the corresponding standardized terminology used by Sequence Ontology (SO) for assessing sequence changes

Sequence Ontology vocabulary term of the effect	Note	Impact	No. of effects
downstream_gene_variant	Downstream of a gene (up to 5Kb)	MODIFIER	27 629
intergenic_region	Variant is in an intergenic region	MODIFIER	32 876
intron_variant	Variant hits an intron	MODIFIER	15 446
missense_variant	Variant causes a codon that produces a different amino acid	MODERATE	6020
splice_acceptor_variant	Variant hits a splice acceptor site	HIGH	17
splice_donor_variant	variant hits a splice donor site	HIGH	14
5_prime_UTR_premature_start_codon_gain_variant	Variant hits 5′UTR region and produces a three base sequence that can be a START codon	LOW	173
start_lost	Variant causes start codon to be mutated into a nonstart codon	HIGH	10
stop_gained	Variant causes a STOP codon	HIGH	37
stop_lost	Variant causes stop codon to be mutated into a nonstop codon	HIGH	11
synonymous_variant	Variant causes a codon that produces the same amino acid	LOW	14 250
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon	LOW	11
upstream_gene_variant	Upstream of a gene (up to 20Kb)	MODIFIER	21 932
3_prime_UTR_variant	Variant hits 3′UTR region	MODIFIER	2597
5_prime_UTR_variant	Variant hits 5′UTR region	MODIFIER	1005
Total			122 028

Impact categories do not predict whether a variant is producing phenotype changes but only represent significant variants in the context of gene level changes.

recalibration modeling. Such an approach to chip development has now become possible with relatively modest budgets and efficiently enhanced by the use of pooled sample sequencing that substantially reduces library construction costs, while allowing much larger numbers of individuals to be included with much wider and deeper SNP sampling (Schlotterer *et al.*, 2014). The rapidly diminishing numbers of SNPs retained following each one of the filtering steps (Table 1), further emphasizes the need to start with a large and diverse sample of individuals for a successful outcome in SNP genotyping and conversion. We believe that this report, detailing all the SNP discovery, ascertainment and validation strategies adopted, should provide a valuable roadmap for future developments of large-scale and flexible SNP genotyping platforms for highly heterozygous genomes.

Shared SNPs across species hint to a recent species radiation of the eucalypts

The availability of a high-quality reference genome coupled to stringent SNP discovery parameters to resolve unique alignments, despite the extensive tandem duplications in *Eucalyptus* (Myburg *et al.*, 2014), positively contributed to robust SNP genotyping metrics. Only 1692 out of the 60 904 on the chip (2.8%), could not be genotyped in any species, mostly due to diffuse clustering patterns possibly indicating paralogous SNPs that passed the mapping and filtering criteria, although the occasional deviations from HWE assessed with GenomeStudio involved excess homozygous and not heterozygous genotypes. On average *c.* 53 000 SNPs were genotyped with an overall call rate across samples > 97% in all *Eucalyptus* species (Table 2). SNP transferability and polymorphism across sections and species was high (Fig. 3a). When more distantly related species are considered, several thousand SNPs are also found to be simultaneously polymorphic across subgenera (2662 GRA/PIL; 1579 GRA/CLO) and across genera (5895 GRA/CIT; 1759 GRA/PSI) (estimates extracted from MAF data in Table S4). Not surprisingly, the multi-species SNP discovery strategy based on a large sample of 846 resequenced genome equivalents, almost doubled the rate of polymorphic SNP transferable across species beyond our previous estimates (Grattapaglia *et al.*, 2011b). For example, the number of shared SNPs between GRA and CAM went from 27.3% in that previous study to 55.8% now, whereas for GRA vs GLO, it went from 16.5% to 26.4%. These estimates are considerably higher than the few published estimates of SNP transferability across plant (Vezzulli *et al.*, 2008; Pavy *et al.*, 2013) and animal species (Haynes & Latch, 2012; Hoffman *et al.*, 2013). These trans-species SNPs in eucalypts most likely represent ancient variants that arose before the split of these taxa and persisted in separate lineages due to the presumably high effective population sizes of these preferentially outcrossed and widespread tree species. The relationship between successful transferability of SNPs and divergence times between species has been investigated in domestic animals (Miller *et al.*, 2012). A linear decrease of 1.5% in SNP call rate per million years and an exponential decay of retention of polymorphisms was seen, with < 10% of SNPs with retained polymorphism when the time to the last common ancestor was < 5 Myr. The genome-wide SNP persistence we observed across the eucalypt species and subgenera sampled may therefore shed some light on the current debate about the dating of eucalypt divergence. Although evidently some level of ascertainment bias does exist for polymorphism, the high rates of commonly genotyped sites across species (> 90%) and the high proportions of them (25–50%) with retained polymorphism, better fit the proposed hypothesis of a recent species (< 2 Myr ago) and section (5–10 Myr ago) radiation (Ladiges *et al.*, 2003), rather than older radiation dates based exclusively on ITS data (Crisp *et al.*, 2004).

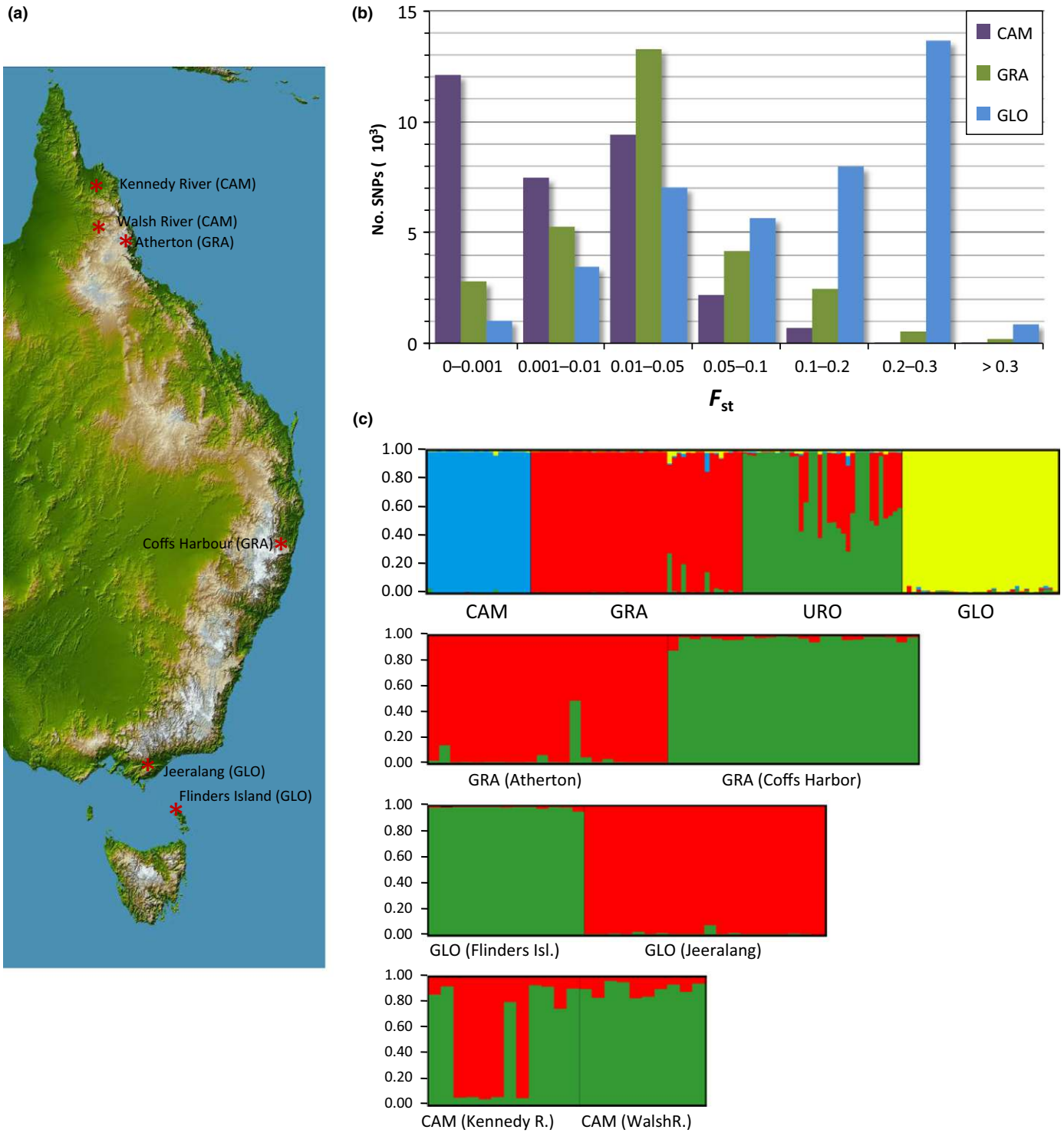

Fig. 5 Population structure analyses of *Eucalyptus* species and provenances. (a) Sampling locations of the *Eucalyptus* species and provenances assessed for population structure (GRA, *E. grandis;* URO, *E. urophylla;* GLO, *E. globulus;* CAM, *E. camaldulensis*). (b) Distribution of F_{st} estimates capturing the provenance variation in three *Eucalyptus* species. (c) From top to bottom, STRUCTURE results using the same set of 600 genome-wide single-nucleotide polymorphisms (SNPs) spaced at *c*. 1 Mb to detect interspecific variation hybrid composition of URO individuals using K = 4, and within-species provenance differentiation in GRA, GLO and CAM, with K = 2 (see text for details). Each column along the *x*-axis represents an individual tree and the *y* axis shows Q, the estimated membership coefficient for each individual.

The large-scale whole-genome sequence-based SNP data we have gathered will be valuable to help elucidate this issue further, including the potential discovery of sites under balancing selection (Delph & Kelly, 2014). From the applied standpoint, these shared SNPs will be valuable for identity and parentage analysis in breeding programs.

Multi-species SNP discovery mitigated the SNP ascertainment bias of the EUChip60K content

A commonly raised issue about fixed content SNP chips has been the degree of SNP ascertainment bias (AB) introduced by the discovery process, depending on how limited and genetically distant the discovery panel was from the individuals to be genotyped. Under such circumstances the discovery panel favors the detection of high-MAF SNPs, a problem which may compromise diversity metrics that depend on allele frequency (Albrechtsen *et al.*, 2010). This becomes even more severe when cross-species genotyping is attempted (Garvin *et al.*, 2010). We assessed the extent of AB in the EUChip60K by comparing the site frequency spectrum (SFS) of the 24 035 polymorphic SNPs (MAF > 0.05) in *E. grandis* with the SFS of the entire set of 19.4 million SNPs discovered in the pooled sample of this same species (Table S6). The lack of significant difference between these two distributions suggests that the relative proportions of SNPs among MAF classes > 0.05 in the EUChip60K fit the distribution of randomly sampled SNPs in the *E. grandis* genome (Fig. 4). Furthermore, the SFS observed for the EUChip60k, enriched for rare variants, is consistent with the distribution of random SNPs in humans (Albrechtsen *et al.*, 2010) and in poplar (Marroni *et al.*, 2011), while markedly different from the SFS of chips lacking rare variants developed from small discovery panels (Groenen *et al.*, 2011; Sim *et al.*, 2012). When the comparison was carried out between the direct SNP counts in the chip and the average SNP counts of 1000 random samples of 24 035 SNPs, again no difference was seen. This second test effectively corresponded to comparing the EUChip60K SNP content to 1000 'simulated chips' built using random subsets of all SNPs discovered in the pooled sample without any prior selection as far as SNP position, sequence context or polymorphism level besides MAF> 0.05. Our results therefore suggest that the large and diverse pooled sample used for SNP discovery, together with the variable SNP selection constraints, alleviated strong selection on common SNPs favoring sampling SNPs across most of the frequency range for *E. grandis*. As for the other species, although our data do not allow any valid test due to lack of sufficient size in the pooled samples, the similar patterns of enrichment for rare SNPs observed (Fig. 3b) also suggest a potential reduction of AB.

EUChip60K, a powerful tool for Eucalyptus genetics, genomics and breeding

Although the main intended use for the EUChip60K is operational Genomic Selection and gene discovery by GWAS, we believe that the chip should also be valuable for population genomics. Nevertheless, because the EUChip60k cannot be considered AB-free, caution should be taken when using it to compare SFS patterns across species or using them as a benchmark for neutrality tests. Still, as a prelude to future population genetic studies, the observed distribution of SNPs F_{st} among provenances (Fig. 5b), clearly show that opportunities exist to use EUChip60K data for genome-wide scans for signatures of selection given the excellent gene-space coverage it provides. Additionally, highly informative ancestry-informative marker panels could be derived from the several hundred SNPs that are privately polymorphic or fixed in species and provenances, as indicated by their high F_{st} both at the inter- and intraspecific levels. Such SNP panels would provide powerful systems to identify and quantify introgression and hybrid composition at the single-chromosome level in individuals of wild and breeding populations, and to understand patterns of species diversification at the whole-genome level. The distribution of F_{st} estimates seen between eucalypt provenances was consistent with their geographical origin and life history (Fig. 5). An $F_{st} < 0.05$ was estimated for 29 020 of the 31 931 (91%) SNPs between the two CAM provenances located at a relatively close distance in northern Queensland, consistent with earlier reports based on RFLP and SSR, showing that geographic proximity and not river system is the main determinant of genetic similarity (Butcher *et al.*, 2009). On the other extreme, 28 138 out of 39 692 SNPs (71%) had $F_{st} > 0.05$ between two GLO provenances that although geographically close, are separated by the Bass Strait in southeastern Australia, confirming earlier microsatellite studies that showed a highly structured genetic architecture of *E. globulus* populations in southeastern Australia (Steane *et al.*, 2006). Finally, even a very modest randomly sampled set of 600 SNPs was much more powerful than microsatellites (Faria *et al.*, 2011) for detecting provenance variation and hybrid composition. From the applied breeding standpoint we are currently on the brink of operational implementation of Genomic Selection in eucalypts. Chip-based genotyping provides breeder-friendly, highly reproducible data within and between laboratories, for the same SNPs across species, with high call rates and no need for specialized bioinformatics infrastructure and personnel. These are absolutely essential conditions for considering the truly operational adoption of molecular tree breeding. The powerful, flexible, user-friendly and cost-effective genotyping platform provided by the EUChip60K, represents a vital component of this process and should prove to be easily integrated into routine breeding practice.

Acknowledgments

We thank FAP-DF grant 'NEXTREE' and EMBRAPA project 03.11.01.007.00.00. O.B.S-J. had an EMBRAPA Doctoral scholarship, D.A.F. a CNPq postdoctoral fellowship and D.G. a CNPq research fellowship. We acknowledge the computing infrastructure assistance of Roberto Togawa and the joint pre-competitive vision and support of the Brazilian forest-based companies that made possible the EUChip60K manufacturing. We also wish to thank the three anonymous reviewers for their insightful critiques and suggestions.

References

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* **27**: 2534–2547.

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N. 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* **193**: 1073–1081.

Bekele WA, Wieckhorst S, Friedt W, Snowdon RJ. 2013. High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal* **11**: 1112–1125.

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

Butcher PA, Skinner AK, Gardiner CA. 2005. Increased inbreeding and inter-species gene flow in remnant populations of the rare *Eucalyptus benthamii*. *Conservation Genetics* **6**: 213–226.

Cappa EP, El-Kassaby YA, Garcia MN, Acuna C, Borralho NMG, Grattapaglia D, Poltri SNM. 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**: e81267.

Chagne D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A *et al.* 2012. Genome-wide SNP detection, validation, and development of an 8k SNP array for apple. *PLoS ONE* **7**: e31745.

Chancerel E, Lamy JB, Lesur I, Noirot C, Klopp C, Ehrenmann F, Boury C, Le Provost G, Label P, Lalanne C *et al.* 2013. High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology* **11**: 368.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**: 80–92.

Crisp M, Cook L, Steane D. 2004. Radiation of the australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present-day communities? *Philosophical Transactions of the Royal Society of London B-Biological Sciences* **359**: 1551–1571.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al.* 2011. The variant call format and vcftools. *Bioinformatics* **27**: 2156–2158.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**: 499–510.

Delph LF, Kelly JK. 2014. On the importance of balancing selection in plants. *New Phytologist* **201**: 45–56.

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**: 491–498.

Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS *et al.* 2009. Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods* **6**: 263–265.

Earl DA, Vonholdt BM. 2012. Structure Harvester: a website and program for visualizing Structure output and implementing the Evanno method. *Conservation Genetics Resources* **4**: 359–361.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Faria DA, Mamani EMC, Pappas GJ, Grattapaglia D. 2011. Genotyping systems for eucalyptus based on tetra-, penta-, and hexanucleotide repeat EST microsatellites and their use for individual fingerprinting and assignment tests. *Tree Genetics & Genomes* **7**: 63–77.

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS. 2012. Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE* **7**: e36347.

Freeman JS, Potts BM, Downes GM, Pilbeam D, Thavamanikumar S, Vaillancourt RE. 2013. Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in *Eucalyptus globulus*. *New Phytologist* **198**: 1121–1134.

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J *et al.* 2011. A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. *PLoS ONE* **6**: e28334.

Garvin MR, Saitoh K, Gharrett AJ. 2010. Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**: 915–934.

Gautier M, Foucaud J, Gharbi K, Cezard T, Galan M, Loiseau A, Thomson M, Pudlo P, Kerdelhue C, Estoup A. 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology* **22**: 3766–3779.

Geraldes A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N *et al.* 2013. A 34k SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other populus species. *Molecular Ecology Resources* **13**: 306–323.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9**: e90346.

Grattapaglia D, de Alencar S, Pappas GJ. 2011a. Genome-wide genotyping and SNP discovery by ultra-deep restriction-associated DNA (RAD) tag sequencing of pooled samples of *E. grandis* and *E. globulus*. *BMC Proceedings* **5**(Suppl 7): P45.

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

Grattapaglia D, Silva OB, Kirst M, de Lima BM, Faria DA, Pappas GJ. 2011b. High-throughput SNP genotyping in the highly heterozygous genome of eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biology* **11**: 65.

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

Groenen MAM, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, Vereijken A, Okimoto R, Muir WM, Cheng HH. 2011. The development and characterization of a 60k SNP chip for chicken. *BMC Genomics* **12**: 274.

Harwood C. 2011. New introductions – doing it right. In: Walker J, ed. *Developing a eucalypt resource: learning from Australia and elsewhere: University of Canterbury*. Christchurch, New Zealand: Wood Technology Research Centre, 43–54.

Haynes GD, Latch EK. 2012. Identification of novel single nucleotide polymorphisms (SNPs) in deer (*Odocoileus* spp.) using the bovineSNP50 beadchip. *PLoS ONE* **7**: e36536.

Hoffman JI, Thorne MAS, McEwing R, Forcada J, Ogden R. 2013. Cross-amplification and validation of SNPs conserved over 44 million years between seals and dogs. *PLoS ONE* **8**: e68365.

Hudson CJ, Freeman JS, Kullan AR, Petroli CD, Sansaloni CP, Kilian A, Detering F, Grattapaglia D, Potts BM, Myburg AA *et al.* 2012. A reference linkage map for eucalyptus. *BMC Genomics* **13**: 240.

Illumina 2010. Infinium genotyping data analysis – a guide for analyzing infinium genotyping data using the genomestudio genotyping module. Illumina Inc. [WWW document] URL http://res.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf [accessed 20 August 2014].

Ladiges PY, Udovicic F, Nelson G. 2003. Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *Journal of Biogeography* **30**: 989–998.

Liu Q, Guo Y, Li J, Long JR, Zhang B, Shyr Y. 2012. Steps to ensure accuracy in genotype and SNP calling from illumina sequencing data. *BMC Genomics* **13**: S8.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**: 111–118.

Marroni F, Pinosio S, Di Centa E, Jurman I, Boerjan W, Felice N, Cattonaro F, Morgante M. 2011. Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation ecotilling. *Plant Journal* **67**: 736–745.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS *et al.* 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **4**: e5350.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Miller JM, Kijas JW, Heaton MP, McEwan JC, Coltman DW. 2012. Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Molecular Ecology Resources* **12**: 1145–1150.

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**: 356–362.

Myburg AA, Potts BM, Marques CM, Kirst M, Gion JM, Grattapaglia D, Grima-Pettenati J. 2007. Eucalyptus. In: Chittaranjan K, ed. *Genome mapping and molecular breeding in plants*. New York, NY, USA: Springer, 115–160.

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.

Pavy N, Gagnon F, Rigault P, Blais S, Deschenes A, Boyle B, Pelgas B, Deslauriers M, Clement S, Lavigne P *et al.* 2013. Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources* **13**: 324–336.

Petroli CD, Sansaloni CP, Carling J, Steane DA, Vaillancourt RE, Myburg AA, da Silva OB, Pappas GJ, Kilian A, Grattapaglia D. 2012. Genomic characterization of dart markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. *PLoS ONE* **7**: e44684.

Poland JA, Rife TW. 2012. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **5**: 92–102.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Perez-Enciso M. 2012. SNP calling by sequencing pooled samples. *BMC Bioinformatics* **13**: 239.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P *et al.* 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* **4**: e6524.

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA *et al.* 2012. Genomic selection for growth and wood quality in eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist* **194**: 116–128.

Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A. 2010. A high-density diversity arrays technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. *Plant Methods* **6**: 16.

Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, Mock KE. 2014. Genotyping-by-sequencing for populus population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS ONE* **9**: e95292.

Schlotterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature reviews. Genetics* **15**: 749–763.

Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S *et al.* 2012. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS ONE* **7**: e40563.

Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB. 2013. Development and evaluation of SoySNP50k, a high-density genotyping array for soybean. *PLoS ONE* **8**: e54985.

Steane DA, Conod N, Jones RC, Vaillancourt RE, Potts BM. 2006. A comparative analysis of population structure of a forest tree, *Eucalyptus globutus* (Myrtaceae), using microsatellite markers and quantitative traits. *Tree Genetics & Genomes* **2**: 30–38.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HCM, Jamli S *et al.* 2014. Design and characterization of a 52k SNP chip for goats. *PLoS ONE* **9**: e86227.

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H *et al.* 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* **15**: 823.

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E *et al.* 2012. Development and evaluation of a 9k SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS ONE* **7**: e35668.

Vezzulli S, Micheletti D, Riaz S, Pindo M, Viola R, This P, Walker MA, Troggio M, Velasco R. 2008. A SNP transferability survey within the genus *Vitis*. *BMC Plant Biology* **8**: 128.

Wang SC, Wong DB, Forrest K, Allen A, Chao SM, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L *et al.* 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal* **12**: 787–796.

Supporting Information

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

Fig. S1 Flowchart of SNP discovery and ascertainment pipeline.

Fig. S2 Frequency distribution of the inter-SNP distances in kb and correlation between converted SNPs and gene models.

Table S1 Sample sizes and origin of the 240 eucalypt trees resequenced in pools

Table S2 Impact of the cluster file samples' genetic composition on the sensitivity and specificity of the EucHIP60k heterozygous calls

Table S3 Infinium SNP probe sequence, ADT score and genome coordinate of the 60 904 SNPs of the EUChip60k

Table S4 List of the 59 222 successfully genotyped SNPs in 14 *Eucalyptus* species and related taxa with their MAF estimates

Table S5 Genotype concordance analysis of replicated SNPs in *E. grandis*

Table S6 Comparative site frequency spectra counts of the EUChip60K and the whole-genome pooled sample

Methods S1 SNPs variants categorization adopted to select the EUChip60K SNP set.

Methods S2 Assessment of sample size and taxonomic composition of the DNA samples used to build cluster files.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.