New Phytologist

Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*

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

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

Author for correspondence:
Dario Grattapaglia
Tel: +55 61 99712142
Email: dario.grattapaglia@embrapa.br

Received: *28 November 2014*
Accepted: *6 May 2015*

New Phytologist (2015) **208:** 830–845
doi: 10.1111/nph.13505

Key words: ancestral recombination graphs, effective population size, *Eucalyptus*, linkage disequilibrium (LD), mutation rate, nucleotide diversity, population recombination rate, whole-genome pooled resequencing.

Summary

• We used high-density single nucleotide polymorphism (SNP) data and whole-genome pooled resequencing to examine the landscape of population recombination (ρ) and nucleotide diversity (Θ_w), assess the extent of linkage disequilibrium (r^2) and build the highest density linkage maps for *Eucalyptus*.

• At the genome-wide level, linkage disequilibrium (LD) decayed within *c*. 4–6 kb, slower than previously reported from candidate gene studies, but showing considerable variation from absence to complete LD up to 50 kb. A sharp decrease in the estimate of ρ was seen when going from short to genome-wide inter-SNP distances, highlighting the dependence of this parameter on the scale of observation adopted. Recombination was correlated with nucleotide diversity, gene density and distance from the centromere, with hotspots of recombination enriched for genes involved in chemical reactions and pathways of the normal metabolic processes.

• The high nucleotide diversity ($\Theta_w = 0.022$) of *E. grandis* revealed that mutation is more important than recombination in shaping its genomic diversity ($\rho/\Theta_w = 0.645$). Chromosome-wide ancestral recombination graphs allowed us to date the split of *E. grandis* (1.7–4.8 million yr ago) and identify a scenario for the recent demographic history of the species.

• Our results have considerable practical importance to Genome Wide Association Studies (GWAS), while indicating bright prospects for genomic prediction of complex phenotypes in eucalypt breeding.

Introduction

Meiotic recombination is a key process driving evolution and selective breeding (Charlesworth *et al.*, 2009). While mutation generates new allelic variants upon which natural selection acts, recombination amplifies the existing genetic variation by shuffling mutations into novel combinations. Furthermore, recombination has been underscored as an important force in driving some aspects of plant genome evolution by causing mutation and influencing the strength of natural selection (Gaut *et al.*, 2007). The recent milestone publication of the *Eucalyptus grandis* genome (Myburg *et al.*, 2014) has now opened extraordinary opportunities to advance the detailed investigation of the genomic attributes that underlie the exceptional diversity of the eucalypts. Understanding the genome-wide patterns of recombination and nucleotide diversity provides important insights into the evolutionary processes that have shaped its genetic history and has, together with demographic factors, a direct impact on the extent of linkage disequilibrium (LD), the nonrandom association of alleles at different loci. The extent of LD, in turn, determines our ability to dissect quantitative traits by linkage or association mapping and carry out accurate whole-genome prediction of complex phenotypes (Goddard & Hayes, 2009).

Genetic map construction has been the standard approach to study recombination by obtaining estimates in cM Mb^{-1} per generation. Ample variation in such rates exists among plant species, from 0.75 for maize to 4.42 for *Arabidopsis thaliana* and 6.07 for *Populus* (Henderson, 2012). In *E. grandis*, the average map-based recombination rate was estimated at 2.53 cM Mb^{-1} (Grattapaglia & Sederoff, 1994), closely matching later estimates based on higher density maps (Petroli *et al.*, 2012). Such direct measurements of recombination, however, typically have low resolution and limited representation of only a few individuals in a single generation. Evolving from the map-based approach, indirect but powerful statistical methods have been developed to extract information about the patterns of recombination from genetic variation data in samples of natural populations (Hudson, 1987; Stumpf & McVean, 2003). This approach became more

accessible with the recent advances in high-throughput genotyping and sequencing. The genetic variation in a sample of unrelated individuals derives from mutation and recombination over many generations in the ancestors of that sample. Coalescent-based methods allow the estimation of the number of crossover events that took place along the ancestry of a particular genomic segment by a quantity called 'population-scaled recombination rate', defined as $\rho = 4N_e c$, where c is the probability of crossover per base pair per generation occurring in the segment and N_e is the effective population size (McVean *et al.*, 2002). For populations in approximate drift–mutation–recombination equilibrium, N_e can then be estimated by equating the estimate of c from genetic maps to the estimate of ρ from population data (Hellenthal & Stephens, 2006). Population recombination rates are known to be highly heterogeneous along eukaryotic genomes (Petes, 2001), including the ubiquitous occurrence of recombination 'hotspots' (Mezard, 2006). Understanding the genome-wide variability in recombination rates and extent of LD has been a theme of interest in humans (McVean *et al.*, 2004), *Drosophila* (Chan *et al.*, 2012) and a few plants for which extensive genomic resources have been available, such as *Arabidopsis* (Kim *et al.*, 2007), *Medicago* (Branca *et al.*, 2011; Paape *et al.*, 2012), *Populus* (Slavov *et al.*, 2012), *Zea mays* (Bauer *et al.*, 2013) and *Mimulus* (Hellsten *et al.*, 2013). Results from such genome-wide studies have corroborated the ample genome-wide variation in recombination rates and allowed estimation of the relative importance of mutation to recombination (μ/c) in shaping genetic variation in the species.

A number of studies have estimated the extent of LD in forest tree genomes based on sampling polymorphisms in short sequence stretches along genes. While early studies in *Pinus taeda* showed LD dropping to $r^2 < 0.2$ within a *c.* 1.5 kb distance (Brown *et al.*, 2004; Neale & Savolainen, 2004), later studies in the same species and other conifers converged to a much faster decay of LD within a few hundred bp (González-Martínez *et al.*, 2011). A rapid decay of intragenic LD within < 1 kb distances was also reported in *Populus* (Neale & Ingvarsson, 2008) and *Eucalyptus* (Grattapaglia & Kirst, 2008; Denis *et al.*, 2013). This picture, taken as the consensus for outcrossed undomesticated forest trees, has started to change in the last few years, as genome-wide genotyping technologies have become available, allowing the assessment of LD from a much larger number of two-point estimates at variable SNP distances. Recent genome-wide analyses in *Populus* have shown a substantially slower decline of the average LD within 3–6 kb (Slavov *et al.*, 2012) when compared with earlier reports in short genic tracts. A significantly more extended LD (up to 110 kb) was reported in the conifer *Cryptomeria japonica* when LD was assessed across longer genomic segments (Moritsuka *et al.*, 2012). More variable LD was also reported in Norway spruce (Larsson *et al.*, 2013), loblolly pine (Eckert *et al.*, 2010) and *Fagus* (Lalague *et al.*, 2014), clearly highlighting the necessity to move beyond the standard consensus that forest trees display very low LD, and better investigate its extent at wider genomic scales.

Studies in humans (Frisse *et al.*, 2001; Pritchard & Przeworski, 2001; Tenesa *et al.*, 2007), *Drosophila* (Andolfatto & Wall, 2003) and *Arabidopsis* (Nordborg *et al.*, 2005; Kim *et al.*, 2007) have reported lower recombination but significant LD over distances longer than those predicted by standard population models, and higher recombination and less LD than would be expected in short genomic segments. In other words, the short-range LD is incompatible with the long-range pattern, with too little of the former relative to the latter for the data to be explained by a simple recombination model based exclusively on crossing-over. Clearly, such a nonlinear relationship between recombination and physical distances between sites indicates that population recombination rate and, consequently, LD depend upon the genomic scale of observation adopted. More important than averaging ρ is understanding its variance as a function of the spacing between heterozygous SNP sites over variable genomic distances (Goldstein & Weale, 2001), similarly to what occurs with the well-known variability of pairwise LD measures (Hill & Weir, 1988). Features such as demographic fluctuations, local variation in recombination rates and the impact of gene conversion contribute to this apparent discrepancy (Pritchard & Przeworski, 2001).

Nothing is known about population recombination rates in *Eucalyptus*, and there is very little information about the extent of LD. No attempt has been made to use LD data to describe past demographic events and thus contribute to the standing debate regarding the evolutionary history of the eucalypts (Ladiges & Urdovicic, 2005). The few available LD studies to date were carried out along short genic segments showing a rapid LD decay after 500–1500 bp in *E. grandis* (Grattapaglia & Kirst, 2008), *Eucalyptus urophylla* (Denis *et al.*, 2013), *Eucalyptus globulus* (Thavamanikumar *et al.*, 2011) and *Eucalyptus nitens* (Thumma *et al.*, 2005). Similarly, nucleotide diversity has only been described in *Eucalyptus* within a limited set of genes by Sanger sequencing (Poke *et al.*, 2003; Thavamanikumar *et al.*, 2011; Mandrou *et al.*, 2014) or by next-generation pooled sequencing (Novaes *et al.*, 2008; Kulheim *et al.*, 2009), revealing generally high but variable rates between 0.006 and 0.05.

In this study we characterized the genomic landscape of population-scaled recombination rate and nucleotide diversity of *E. grandis* using data from two different SNP sources, Infinium genotyping and pooled whole-genome resequencing. Genome-wide estimates of map-based recombination rates were obtained from the highest density linkage maps ever built for eucalypts. At the genome-wide level, we show that LD decays considerably more slowly than previously reported. We also examined the relationship between population recombination rates and genomic features while showing that recombination varies with the pairwise SNP distance range at which it is measured. The estimates of ρ in turn allowed inferences on the demographic history of *E. grandis*, both recent and in the more distant past. Besides providing the first genomic-based examination of the evolutionary history of *E. grandis*, our results have significant implications for the study of complex trait variation and the application of Genome Wide Association Studies (GWAS) and genomic prediction to modern breeding.

Materials and Methods

Plant material, Infinium SNP genotyping and whole-genome pooled resequencing

We studied a sample of 48 unrelated *Eucalyptus grandis* Hill ex Maiden trees, 24 from each one of two wild populations in Australia (Atherton, 17°15′S, 145°28′E; Coffs Harbor, 30°18′S, 153°07′E). Genome-wide SNP data were obtained with the Infinium EuCHIP60K (Silva-Junior *et al.*, 2015). Whole-genome resequencing data for 36 of the 48 *E. grandis* trees (18 per natural population) were obtained in three pools of 12 trees each, without sample barcoding, with an estimated *c.* 3.5× coverage of each haploid genome by shotgun sequencing (Hiseq paired-end 2 × 100; *c.* 500 bp insert size). Linkage maps were built using the reference hybrid mapping population of 189 F_1 individuals used earlier (Petroli *et al.*, 2012).

High-density linkage map construction using EUChip60K SNP data

Independent linkage maps were built for the two parents using Infinium SNPs segregating 1 : 1 following a pseudo-testcross (Grattapaglia & Sederoff, 1994) (Supporting Information Methods S1). Markers were grouped (LOD > 20.0) and genotype data phased using JoinMap v3.0 (Van Ooijen & Voorrips, 2001). Phased SNP genotype data for each linkage group separately were exported from JoinMap and entered into Record (Van Os *et al.*, 2005) for marker ordering using Kosambi's mapping function. Linkage maps were drawn using MapChart (Voorrips, 2002). The relationships between the genetic and physical positions of the mapped SNPs on the *E. grandis* genome were represented by Marey maps (Chakravarti, 1991). Linkage map-based recombination rates (in cM Mb^{-1}) were estimated for every 5 Mb windows using linear regression in R (version 3.1.0) after analyzing the residuals plot against the fitted values to remove outlier markers.

Linkage disequilibrium from genome-wide EUChip60K SNP data

Pairwise estimates of LD measured by the squared correlation of allele frequencies (r^2) were obtained using SNPs with minor allele frequency (MAF) > 0.05 and call rate > 95% (21 517 SNPs in total) for each chromosome separately. LD analysis was performed using LDcorSV (Mangin *et al.*, 2012) to estimate r^2 and r^2_{SV}, that is, corrected for population structure and relatedness. Population structure analysis was performed with 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^{-1}. The most probable value of K (K = 2) was defined by ΔK (Evanno *et al.*, 2005). A realized kinship matrix based on marker data was calculated using Synbreed (Wimmer *et al.*, 2012). Decay curves of r^2 and r^2_{SV} were fitted using a nonlinear regression of pairwise LD based on the expectation of r^2 and r^2_{SV} for drift–recombination equilibrium (Hill & Weir, 1988) using the R script by Marroni *et al.* (2011).

Population-scaled recombination rate from genome-wide EUChip60K SNP data

We used coalescent-based methods implemented by LDhat (McVean *et al.*, 2002) and Hotspotter (Li & Stephens, 2003) to estimate the population-scaled recombination rate, $\rho = 4 N_e c$, by which the underlying recombination rate is directly related to the pattern of LD considering all sites simultaneously instead of only pairs of sites. LDhat, a widely used program, uses a 'composite likelihood' and works well if there is little LD among SNP pairs. However, it has been shown not to be particularly efficient in capturing the fact that markers may be in weak LD with neighboring markers, but in strong LD with more distant ones (Li & Stephens, 2003). Hotspotter, on the other hand, develops a parsimonial method to model LD, exploring its variable patterns under different scenarios without assuming loci to be independent. Genotypic SNP data gathered with the EUChip60K were converted to VCF files which were processed in sliding windows containing a fixed number of sites (15 or 105) with adjacent windows overlapping by 5 SNPs. Phase 2.1.1 (Stephens *et al.*, 2001) was used to estimate the haplotypes within each SNP window and the programs *Interval* in LDhat version 2.2 and *Rholike* in Hotspotter were used to obtain estimates of ρ (Methods S2). Additionally, a third approach was used to estimate ρ, relying directly on the measure of r^2_{SV} obtained for pairwise distances from 0.1 kb up to 50 kb according to a model of drift–recombination equilibrium (Hill & Weir, 1988).

Estimation of nucleotide diversity and population-scaled recombination from whole-genome pooled resequencing data

Sequence alignment information to the reference genome generated by Novoalign was extracted as pileup files from the full read alignment using Samtools mpileup (Li *et al.*, 2009). We used Popoolation (Kofler *et al.*, 2011) to calculate the genome-wide nucleotide diversity, that is, the population-scaled mutation rate per site, $\Theta_w = 4 N_e \mu$ (Watterson, 1975) along chromosomes using nonoverlapping sliding windows of 100 kb with default parameters. We used mlRho (Haubold *et al.*, 2010) to estimate ρ for SNP sets at variable pairwise distances, based on maximum likelihood estimators for Θ and sequencing error. As mlRho was developed to analyze data from a single individual and our data came from a pool, we first performed an SNP identification step using Snape-pooled (Raineri *et al.*, 2012) (Methods S3). Having the profiles for the alignment positions in the pileups we used the formatPro program (with minimum profile coverage = 4) for subsequent analysis with mlRho. The scaled mutation rate $\Theta = 4 N_e \mu$ was computed for all positions looked at individually (mlRho – m0). To obtain the estimates of ρ we evaluated the zygosity (Δ) at all pairwise distances up to 50 kb along each one of the chromosomes in the reference genome.

Effective population size (N_e) and time to the most recent common ancestor (TMRCA) of *E. grandis*

Estimates of N_e were derived from the computed values of ρ by equating ρ to the recombination rate c estimated by linkage mapping ($\rho = 4\ N_e c$). The estimates of N_e, Θ_w, μ (mutation rate per generation) and the phased haplotype sequences were used to obtain estimates of the TMRCA from ancestral recombination graphs (ARGs). Because of time limitations in running this analysis, only chromosomes 8 and 10 were used, involving, nevertheless, 110 Mb of genomic sequence. ARGWEAVER (Rasmussen *et al.*, 2014) was used to sample ARGs for the genomic regions described by the 105-SNP windows in an alignment of the sequences representing the phased haplotypes (Methods S4).

Recombination hotspot detection and annotation

Phased Infinium SNP data for the sliding windows of 105 SNPs were used to infer the location of recombination hotspots from patterns of LD using LDHOT 0.4 (Auton *et al.*, 2014). By default LDHOT assesses the significance of the localized peaks within previously recorded recombination rate estimates and tests for the presence of hotspots by constructing overlapping 3 kb subwindows that are separated by 1 kb across any particular SNP window region (Methods S5). To gain initial biological insight about the gene content of putative hotspots, we used a gene ontology (GO) categories enrichment analysis to record the distribution of GO terms of the genes within those regions and searched for the high-frequency (and low-frequency) terms compared with the rest of the genome (Methods S6).

Genomic correlates of recombination

Pearson's correlations along windows of 100 kb were calculated between recombination rates and the following genomic features: nucleotide diversity (Θ_w), gene density (measured as the proportion of base pairs of the window falling into coding regions), GC content (%) and distance from the centromere to the tip of each chromosome arm (in kb). As no information exists regarding the exact position of centromeres in the *Eucalyptus* chromosomes, and no relationship has been yet established between the pseudochromosomes and the chromosomes in cytological observations, all chromosomes were assumed to be metacentric. Correlation significance was assessed by comparing the calculated values with those of 5000 permuted data sets that maintained the chromosomal order of all observations but that shuffled the relative positions of the two variables (Nordborg *et al.*, 2005) using the function 'sample' in R.

Results

Linkage map alignment to the *E. grandis* genome and map-based estimates of recombination

The two linkage maps built with independent sets of SNPs contained 4396 SNPs for *E. grandis* and 3991 for *E. urophylla*, covering similar total recombination distances (Fig. 1; Table S1; Notes S1). A total of 192 (2.3%) nonsyntenic SNPs, were excluded, that is SNPs that mapped to linkage groups different from the expected ones based on their genome position. The alignment of the *E. grandis* maps to the current genome version 1.1 revealed assembly inconsistencies on several chromosomes, more notably on chromosomes 1, 2, 4 and 8 (Figs 1, 2, S1). Map-based estimates of recombination rates were similar for the two species, 3.18 ± 1.1 and 3.55 ± 0.8 cM Mb^{-1} (Table S1). The MAREY maps (Figs 2, S1) revealed a fairly similar slope of recombination rate across all chromosomes, and only modest plateaus of recombination were seen, for example, on chromosomes 4, 9 and 10.

Extent of genome-wide linkage disequilibrium in *E. grandis*

Pairwise estimates of r^2 were obtained from haplotype probabilities for all pairwise distances among the Infinium SNPs on each chromosome. A total of 21 351 SNPs, an average of 1941 SNPs per chromosome, with pairwise distances varying from 135 bp up to several Mb, were used in the calculations, resulting in nearly two million pairwise estimates of r^2 per chromosome and over 21 million at the genome-wide scale for 72 sampled genomes of *E. grandis*. Owing to the very large number of estimates of r^2 and because the LD decay curves become an asymptote thereafter, LD decay plots are shown only up to 50 kb distances. Average genome-wide LD was $r^2 = 0.131$, decaying to < 0.2 within *c.* 5.7 kb (half-decay within 4.3 kb) while $r^2_{SV} = 0.123$, showing a slightly faster decay within *c.* 4.9 kb (half-decay within 3.7 kb) (Fig. 3). The small difference between raw and corrected r^2 is consistent with the lack of population structure between the two *E. grandis* provenances as indicated by the low $F_{st} = 0.041 \pm 0.06$ previously calculated based on 28 658 genome-wide SNPs (Silva-Junior *et al.*, 2015). LD was also estimated among the *c.* 4000 linkage mapped markers in *E. grandis* and r^2 plotted against the distance in cM. With map resolution of *c.* 0.3 cM, corresponding to *c.* 106 kb, no r^2 estimate was larger than 0.2, consistent with the decay observed at *c.* 4–6 kb (Fig. S2). No impact of the few assembly inconsistencies of the *E. grandis* genome version 1.1 was seen on the pattern of LD decay and no difference was observed when rarer SNPs with MAF > 0.01 were included in the estimation of r^2 (Fig. S3).

Population-scaled recombination rate in *E. grandis*

Three approaches to obtain genome-wide estimates of ρ were undertaken using two independent experimental data sets encompassing nearly 13 million SNPs from the pooled sequence data and over 21 000 Infinium SNPs. Chromosome-specific estimates of ρ were obtained for different genomic window sizes in terms of SNP numbers with LDHAT and HOTSPOTTER (Table 1). Estimates were computed across all chromosomes in 2402 overlapping bins of 15 SNPs (mean bin size, 250 kb; SNP density, 1/16.7 kb) and in 242 overlapping bins of 105 SNPs (mean bin size, 2500 kb; SNP density, 1/23.81 kb). Little

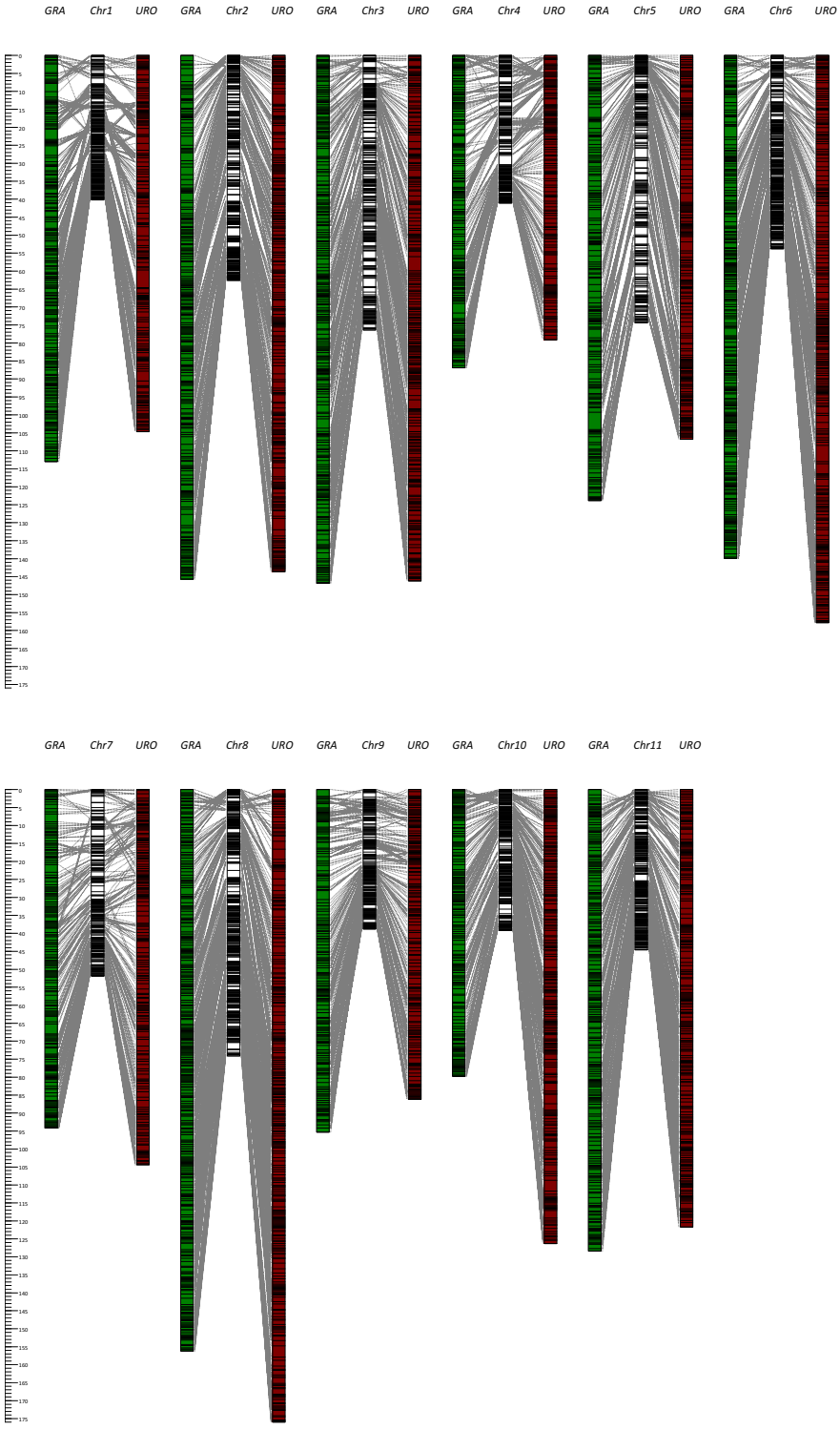

Fig. 1 Linkage maps of *Eucalyptus grandis* (green bars) (4396 single nucleotide polymorphisms (SNPs), 1310.2 cM) and *Eucalyptus urophylla* (red bars) (3991 SNPs; 1,352.8 cM) aligned to the *Eucalyptus* reference genome 1.1 (white bars). The scale on the left corresponds simultaneously to cM distances for the linkage map and to Mb of sequence for the chromosome scaffolds. SNP positions in centiMorgan along the maps and their corresponding base pair position in the genome are provided (Supplementary Information Notes S1).

variation was seen between the two estimates and across chromosomes. Estimates obtained by HOTSPOTTER were in general twice as large as those from LDHAT, probably reflecting the different assumptions regarding the LD model of the two estimation methods. Consequently, the genome-wide estimates of effective population sizes obtained by equating ρ to the recombination rate c were also twice as large with HOTSPOTTER. The genome-wide estimates of ρ obtained with Infinium SNPs converged to essentially equivalent values and magnitudes to the estimate obtained with the pooled sequencing data using mlRho as well as to estimates derived from population-level LD (Table 2). Although some degree of sampling bias intrinsic to the method seems to be present when comparing the estimates of ρ from HOTSPOTTER and mlRho and also between the estimates from LDHAT and population-level LD, the values did not differ by > 30%. Nevertheless, the latter two estimates are two to three

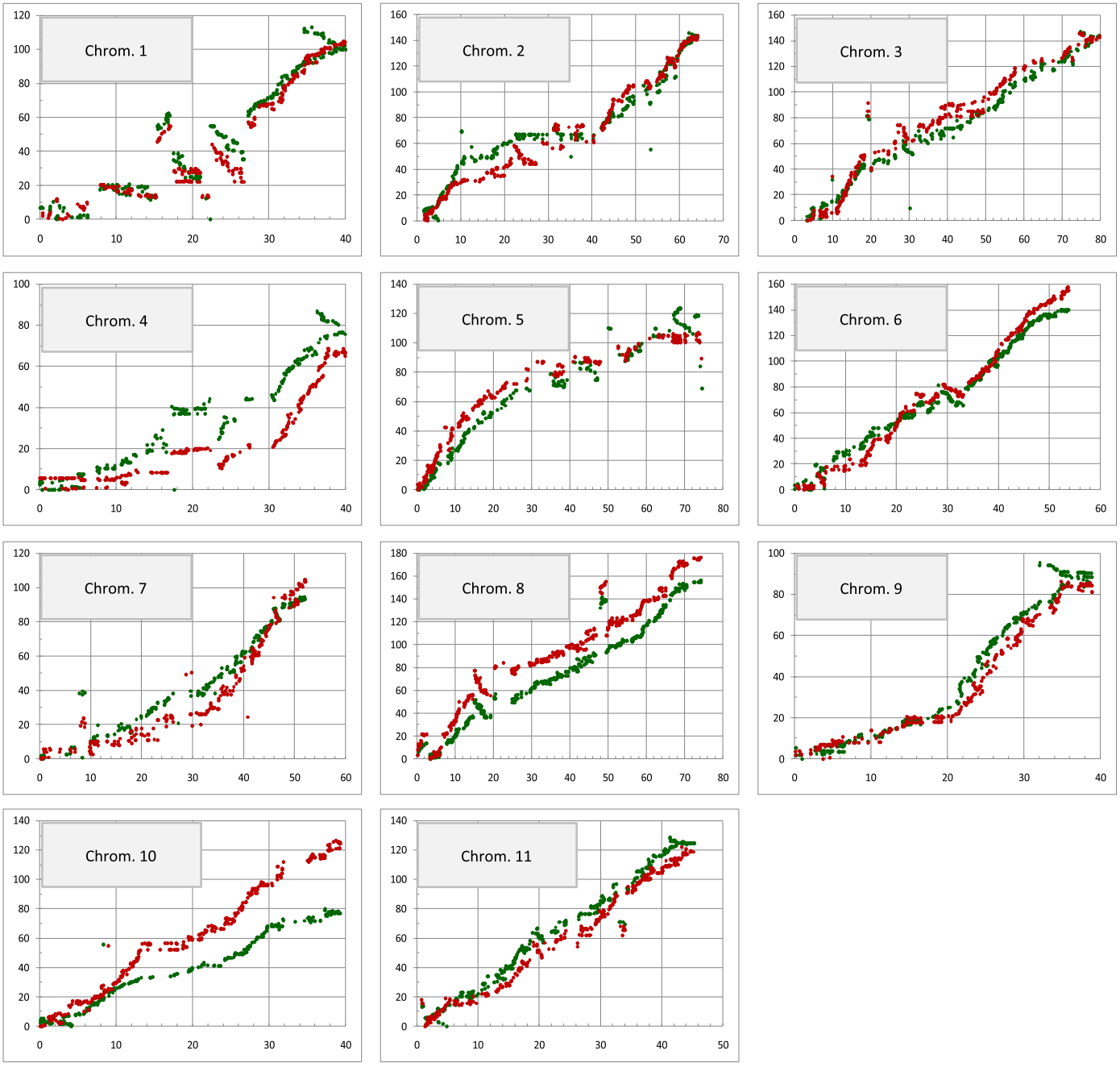

Fig. 2 Marey maps showing the correspondence between the physical position (*x*-axis, Mb) and the recombination-based position (*y*-axis, cM) on the 11 chromosomes of the *Eucalyptus* reference genome 1.1 for two independent sets of single nucleotide polymorphisms (SNPs) mapped on separate linkage maps, *Eucalyptus grandis* in green and *Eucalyptus urophylla* in red.

times smaller than the first two. All estimates of ρ have large standard deviations, reflecting the anticipated genome-wide variation in recombination, and particularly so for the estimate from sequencing data, possibly as a result of the much larger number of nucleotides inspected.

Population recombination rates at variable pairwise SNP distances

The large sampling variance of ρ observed for the sequencing data was further investigated by plotting the estimates of ρ at variable pairwise physical distances from 1 bp up to 50 kb. A rapid decay of ρ with physical distance was seen from 10^{-1} to 10^{-3} within *c.* 6–8 kb (Fig. 4). The same pattern was seen when ρ was estimated directly from LD as r^2 estimated from Infinium SNP data using Hill & Weir formula, evidently not covering distances < 200 bp as a result of the inherent limitations of SNP spacing with Infinium genotyping. We also summarized the estimates of ρ at short (0–0.1 kb), moderate (0–2 kb) and genome-wide (0–50 kb) scales (Table 3). Although little variation was seen across chromosomes, drops of orders of magnitude were seen in recombination rate going from the short to the genome-wide

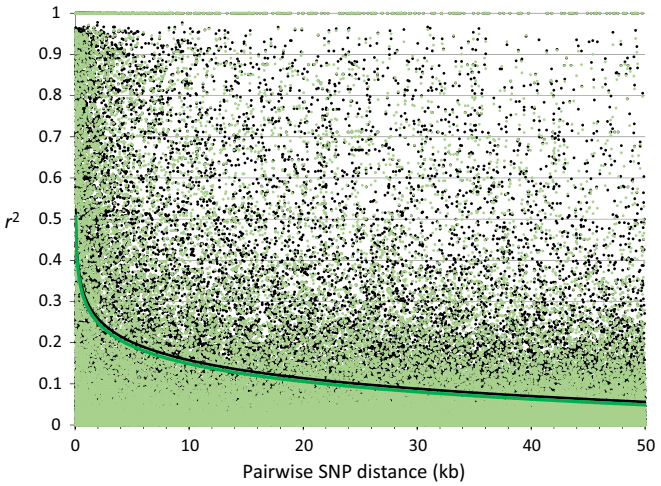

Fig. 3 Genome-wide pattern of decay of linkage disequilibrium (LD) in *Eucalyptus grandis* up to 50 kb pairwise single nucleotide polymorphism (SNP) distances. Decay curves without correction for structure and relatedness (black dots and regression line) and adjusted for structure and relatedness using LDcorSV (green dots and regression line).

range. More important, however, was the behavior of the associated variance. A considerably larger sampling variance was seen when ρ was estimated between closely spaced SNPs (0.142 ± 0.171) as compared with the estimates obtained at moderate (0.0143 ± 0.005) and genome-wide (0.0015 ± 0.00075) distance ranges. When the variable estimates of ρ obtained at different pairwise distance scales were used to calculate the effective population sizes, a corresponding variation was seen. A considerably larger estimate of $N_e = 1\,116\,352$ was obtained when recombination was measured at the short range (0–100 bp), diminishing to $N_e = 112\,421$ at moderate range (0–2 kb) and further dropping to $N_e = 11\,557$ when all distances were taken into account, this last estimate consistent with the genome-wide estimate obtained with the EuCHIP60K data using HOTSPOTTER ($N_e = 8842$).

Genome-wide nucleotide diversity in *E. grandis*

Measures of nucleotide diversity (Θ_w) with both mlRho and Popoolation were obtained from the whole set of sequence alignments that included *c.* 13 million SNPs. Genome-wide and chromosome-specific estimates of nucleotide diversity obtained by the two methods were essentially equivalent, 0.024 bp^{-1} and 0.022 bp^{-1}, respectively, corresponding to an overall SNP frequency *c.* 1 SNP/45 bp with little variation among chromosomes (Table 3). The genome-wide estimates of Θ_w obtained with POPOOLATION were used to derive estimates of mutation rate per generation, which varied between 4.96×10^{-9} and 4.8×10^{-7} bp^{-1} per generation, depending on the scale of pairwise SNP distance adopted. Assuming a generation time of 10 yr for *E. grandis* in natural environments (Eldridge *et al.*, 1993), the corresponding annual mutation rate varied between 4.96×10^{-10} and 4.8×10^{-8} bp^{-1} yr^{-1} with a more credible estimate from the moderate pairwise range (0–2 kb) at 4.93×10^{-9} bp^{-1} yr^{-1} (see later). Using the estimate in the moderate pairwise range, the ratio between mutation and recombination rates (μ/c) given by Θ_w/ρ ($4N_e\mu/4N_ec$) was $\mu/c = 1.55$ or $\rho/\Theta_w = 0.645$ (Table 3).

Inferences on the demographic history of *E. grandis*

Ancestral recombination graphs were summarized in terms of the branch length at each genomic position and the estimated posterior expected values of time to the TMRCA. Posterior mean and 95% credible intervals were also computed per genomic position (Table 4). The results derive from a data set of 66 sequences of 2 Mb length listed in terms of 11 positions along chromosome 8 of *E. grandis*. The same analysis was carried out for chromosome 10 (data not shown) delivering equivalent results. N_e was set to reflect an ancestral effective population size N_0 (138 177) and a current effective population size N_c (9658), both estimated for

Table 1 Mean population-scaled recombination rates (ρ) for the 11 *Eucalyptus grandis* chromosomes and genome-wide (GW), based on genomic windows of 15 and 105 Infinium-genotyped single nucleotide polymorphisms (SNPs) estimated using two alternative models implemented by LDHAT and HOTSPOTTER

			LDHAT				HOTSPOTTER			
Chromosome	No. of SNPs	Window size (kb)	$\rho \times 10^{-3}$ bp^{-1} 15 SNPs	$\rho \times 10^{-3}$ bp^{-1} 105 SNPs	N_e 15 SNPs	N_e 105 SNPs	$\rho \times 10^{-3}$ bp^{-1} 15 SNPs	$\rho \times 10^{-3}$ bp^{-1} 105 SNPs	N_e 15 SNPs	N_e 105 SNPs
1	2526	359.5	0.490	0.473	3852	3719	1.46	0.877	11 452	6897
2	3790	381.2	0.649	0.824	5102	6478	1.40	1.39	10 982	10 935
3	3686	485.2	0.637	0.709	5008	5574	1.24	1.22	9776	9556
4	2526	379.4	0.530	0.444	4167	3491	0.966	0.836	7593	6575
5	3515	478.8	0.424	0.398	3333	3129	0.705	0.728	5544	5723
6	3803	319.6	0.542	0.228	4261	1792	0.922	0.915	7246	7192
7	3041	375.5	0.567	0.662	4458	5204	1.43	1.13	11 279	8891
8	4657	355.6	0.600	0.538	4717	4230	1.33	1.24	10 436	9722
9	2618	329.0	0.444	0.237	3491	1863	0.870	0.962	6836	7560
10	2726	333.4	0.478	0.351	3758	2759	1.11	0.995	8713	7821
11	3132	325.0	0.459	0.355	3608	2791	0.865	0.950	6802	7467
GW	36 020	375	0.530	0.444	4167	3491	1.12	1.03	8842	8120

Corresponding estimates of effective population sizes (N_e) were calculated from $\rho = 4N_ec$, using c, the linkage map-based recombination rate c bp^{-1} per generation.

Table 2 Summary of the genome-wide estimates of population-scaled recombination rate (ρ) and corresponding SD obtained using different single nucleotide polymorphism (SNP) data sets and estimation methods for *Eucalyptus grandis*

Data set type	No. of SNPs surveyed	Estimation method	ρ ($\times 10^{-3}$ bp^{-1})	SD ($\times 10^{-3}$ bp^{-1})
Infinium SNPs	21 351	HOTSPOTTER	1.125	1.69
Infinium SNPs	21 351	LDHAT	0.530	0.460
Infinium SNPs	21 351	LD – Hill & Weir	0.690	0.156
Pooled sequencing SNPs	13 000 000	mlRho	1.470	10.0

chromosome 8 (Table 3). Both values of N_e were assumed to be constant in building the sample ARGs. Using N_0, the mean TMRCA over all positions was 2463.4 thousand yr ago (ka) with estimates at each position generating wider confidence intervals, reflecting a large variability in the ARGs. Using N_c, the mean TMRCA over all positions was *c.* 500 ka with a considerably narrow confidence interval (Table 4). The larger dispersion on the estimates of TMRCA at N_0 is expected by the algorithm used to sample ARGs, as those are impacted to a greater extent by the lower ratio $\Theta/\rho = 1.17$ than the estimates at N_c where $\Theta/\rho = 16.8$ (Rasmussen *et al.*, 2014). Besides the dependence on the ratio Θ/ρ, coding sequences show prominent reduction in ARG-based TMRCA with minimal values in exons and gradually increasing with distance from exon boundaries (Rasmussen *et al.*, 2014). Regarding this issue, we noted a negative correlation, although not significant, between the estimates of TMRCA and gene density along the analyzed positions ($r = -0.381$; $P = 0.248$). Considering these two aspects, we reasoned that the estimates for the branch 53.08–56.20 Mb ($N_0 = 3271.17$ ka; $N_c = 469.46$ ka) containing narrow credible intervals ($N_0 = 1739.18$–4803.17 ka; $N_c = 434.70$–504.23 ka) and the lowest gene density (0.002) should reflect more closely a situation in which the influence of genes on the patterns of neutral diversity was probably small and the variability in the ARG sampling process was therefore lower.

Correlates and hotspots of recombination

Correlations calculated based on 6122 windows of 100 kb revealed that ρ is significantly positively correlated, albeit in decreasing magnitude, with nucleotide diversity, gene density, distance from the centromere and GC content (Table 5). The chromosome-wide landscapes of recombination showed recombination peaks located mostly toward the extremes of the chromosomes (e.g. chromosomes 1, 3, 4 and 6), corroborated by the higher positive correlations (Table S2), although more centrally located peaks were observed on chromosomes 7 and 10 for which correlations were negative (Table S2; Fig. 5). Again no significant impact of the localized assembly inconsistencies of the *E. grandis* genome version 1.1 was seen on the genome-wide profiles of

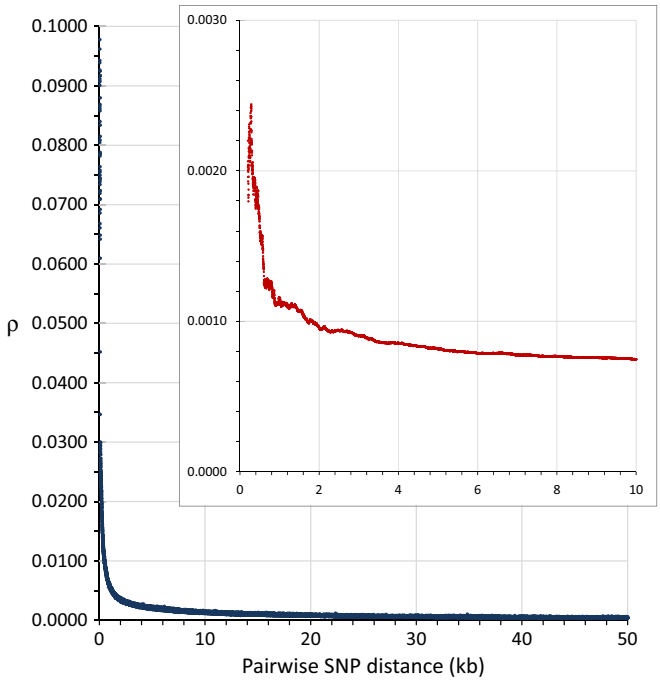

Fig. 4 Decay of population-scaled recombination rate (ρ) with increasing pairwise single nucleotide polymorphism (SNP) distance. Estimates of ρ were obtained based on whole-genome pooled sequencing SNP data using mlRho (blue curve) and EuCHIP60K Infinium SNP data by fitting the estimates of pairwise r^2 obtained with LDcorSV into the Hill & Weir (1988) regression formula (red curve).

population recombination rate (Fig. S4) and the chromosome-specific estimates of ρ (Fig. S5). The profiles of gene density also showed higher values toward the tips of the chromosomes and significant positive correlations with the distance from the centromere were seen for all chromosome except no. 9 (Table S2). In opposite to those spiky patterns, very homogeneous degrees of nucleotide diversity were seen across the entire genome and, as expected, Θ_w was strongly negatively correlated with gene density and positively correlated with GC content. A total of 179 hotspots of recombination were detected in the *E. grandis* genome. When compared with the rest of the genome, the hotspot regions were found to be significantly enriched for GO terms associated with chemical reactions and pathways that are part of the normal metabolic processes, while terms associated with regulation, signal transduction and response were significantly underrepresented (Table 6).

Discussion

A combination of whole-genome pooled resequencing and high-density SNP genotyping was used to report the first genome-wide examination of key features of the *Eucalyptus* genome that have a profound impact on the understanding of fundamental evolutionary issues and on the practice of molecular breeding: nucleotide diversity, population recombination and extent of LD. We corroborated previous results showing that *E. grandis* displays one of the highest nucleotide diversities in plants in general, and forest trees in particular ($\Theta_w = 0.022$), consistent with its exceptional

Table 3 Population recombination rates for *Eucalyptus grandis* (ρ) estimated at different inter-single nucleotide polymorphism (SNP) distances along the 11 *Eucalyptus* chromosomes and the genome-wide (GW) average using mlRho based on the pooled resequencing data

	0–100 bp				0–2 kb				0–50 kb					
Chromosome	ρ	N_e	μ per generation ($\times 10^{-8}$)	μ/c	ρ	N_e	μ per generation ($\times 10^{-8}$)	μ/c	ρ	N_e	μ per generation ($\times 10^{-8}$)	μ/c	Θ_w mlRho ($\times 10^{-2}$)	Θ_w Popool. ($\times 10^{-2}$)
1	0.1436	596 175	0.850	0.14	0.0143	59 385	8.53	1.42	0.0010	4298	117.8931	19.58	2.29	2.03
2	0.1380	1088 344	0.525	0.17	0.0142	111 987	5.10	1.61	0.0010	8273	69.02985	21.78	2.40	2.28
3	0.1499	1296 289	0.495	0.17	0.0155	134 083	4.78	1.66	0.0013	11 125	57.66067	19.95	2.54	2.57
4	0.1474	888 127	0.593	0.14	0.0146	87 952	5.99	1.44	0.0015	8819	59.70161	14.39	2.36	2.11
5	0.1592	1624 980	0.391	0.16	0.0161	164 286	3.86	1.58	0.0012	11 918	53.27282	21.74	2.72	2.54
6	0.1481	1267 723	0.372	0.13	0.0147	125 856	3.74	1.28	0.0012	10 437	45.12945	15.46	2.17	1.88
7	0.1437	1996 306	0.301	0.17	0.0142	197 222	3.05	1.69	0.0009	12 511	48.01385	26.67	2.61	2.40
8	0.1575	1121 717	0.508	0.14	0.0194	138 177	4.12	1.17	0.0014	9658	58.97111	16.80	2.53	2.28
9	0.1467	1324 377	0.373	0.13	0.0148	133 574	3.70	1.34	0.0011	10 060	49.13233	17.74	2.30	1.98
10	0.1500	1543 508	0.300	0.12	0.0147	151 235	3.06	1.26	0.0013	13 358	34.64584	14.26	2.28	1.85
11	0.1394	1213 972	0.401	0.14	0.0143	124 564	3.91	1.36	0.0013	11 603	41.99422	14.63	2.23	1.95
GW	0.1421	1116 352	0.496	0.16	0.0143	112 421	4.93	1.55	0.0015	11 557	47.94229	15.08	2.41	2.22

Corresponding estimates of effective population sizes (N_e) were estimated from $\rho = 4N_e c$ using the linkage map-based recombination rate c bp^{-1} per generation. Mutation rates (μ) per generation were estimated from the estimates of Watterson nucleotide diversity $\Theta_w = 4N_e\mu$. Estimates of nucleotide diversity were obtained with mlRho and Popoolation.

genetic and phenotypic variation (Grattapaglia *et al.*, 2012), resulting from a high estimated genome-wide mutation rate (4.9×10^{-8} bp^{-1} per generation). The genomic landscape of recombination showed that the 11 *Eucalyptus* chromosomes display a relatively similar rate of recombination, which in turn was found to be significantly positively correlated with nucleotide diversity, gene density, distance from the centromere and GC content. This picture was consistent with the fairly constant recombination rates seen along the linkage maps. Both Infinium and sequence-based SNP data converged to essentially equivalent measures of genome-wide population recombination rate ($\rho = 0.53$–1.47×10^{-3} bp^{-1}) using alternative estimation methods. Thus, at the genome-wide scale, mutation was found to be *c.* 1.5 times more important than recombination in shaping the genomic diversity of *E. grandis*. When a closer look was taken at the larger variance associated with the genome-wide estimate of ρ obtained from sequence data, a surprisingly variable pattern emerged. A progressive decrease of ρ by two orders of magnitude was seen when going from the short (0–100 bp) pairwise SNP distance range to the genome-wide scale (0–50 kb). Perhaps not so surprisingly, given recent reports showing variable extents of LD in other forest trees, we provide compelling evidence that the extent of LD in *Eucalyptus*, when assessed at the genome-wide scale, decays considerably more slowly (*c.* 4–6 kb) when compared with previous reports based on the analysis of short sequence stretches in candidate genes.

High-density SNP mapping reveals fairly similar recombination rates across chromosomes in *Eucalyptus*

The EUChip60K genotyping platform (Silva-Junior *et al.*, 2015) allowed the construction of the two highest density linkage maps for *Eucalyptus* (Fig. 1; Table S1). Both maps indicated a few, but clear, assembly inconsistencies in the current genome version, in line with our previous report (Petroli *et al.*, 2012), which also pinpointed problems especially on chromosomes 1 and 4 (see Fig. 2 of that report). A recent mapping study highlighted the same facts, and corrected the assembly inconsistencies in a version 2.0 genome available in PHYTOZOME (Bartholome *et al.*, 2014). Smoother Marey map profiles for chromosomes 1, 4 and 8 were observed when we aligned our *E. grandis* map to this new assembly version (Fig. S2). The high marker density provided by our much higher density linkage maps could further help to improve the reference genome in future assembly efforts. Our mapping work also demonstrates that the EUChip60K provides a powerful SNP genotyping platform for any future linkage mapping project. Because the EUChip60K contains a robust genome-wide set of SNPs transferable across *Eucalyptus* species (Silva-Junior *et al.*, 2015), it will allow detailed interspecific studies of genome structure beyond those reported to date (Hudson *et al.*, 2012). Furthermore, with > 80% of the SNPs at < 10 kb from 30 444 annotated genes in the *Eucalyptus* genome, gene discovery from high-resolution quantitative trait locus mapping becomes a true possibility. The *E. grandis* linkage map reported provided a genome-wide average recombination rate of *c.* 3.18 cM Mb^{-1},

Table 4 Molecular dating of the time to the most recent common ancestor (TMRCA) of *Eucalyptus grandis*

		N_0 = 138 177 (0–2 kb)		N_c = 9658 (0–50 kb)	
Genomic segment position (Mb)	Gene density	Mean TMRCA (ka)	95% credible intervals (ka)	Mean TMRCA (ka)	95% credible intervals (ka)
6.09–9.06	0.005	2917.79	972.07–4863.52	480.94	442.49–519.38
9.00–11.09	0.055	3422.91	1431.92–5413.91	504.99	466.03–543.96
12.65–14.85	0.014	4174.02	2932.34–5415.71	521.41	485.95–556.87
14.71–16.58	0.011	1121.79	994.85–1248.73	488.20	457.93–518.47
33.72–36.59	0.068	1704.83	1242.83–2166.83	463.23	433.44–493.02
39.47–41.65	0.049	2819.61	164.99–5474.22	546.30	498.06–594.55
43.69–46.20	0.070	1904.34	800.95–3007.72	504.47	467.73–541.21
53.08–56.20	0.002	3271.17	1739.18–4803.17	469.46	434.70–504.23
57.69–59.52	0.059	1396.30	590.69–2201.90	523.20	486.59–559.80
59.47–62.70	0.033	1104.92	271.66–1938.19	483.85	452.05–515.66
62.47–64.45	0.010	3260.15	3174.91–3345.39	502.15	468.53–535.78

Estimates were obtained from ancestral recombination graphs (ARGs) along chromosome 8. Gene density as the proportion of the window containing coding sequence. The estimated TMRCAs in thousands of yr ago (ka) were obtained assuming a 10 yr generation time. N_0, ancestral effective population size; N_c, current effective population size.

Table 5 Pearson's correlations among genome-wide recombination rate and genomic features based on 6122 100 kb windows in *Eucalyptus grandis*

	Θ_w	Gene density	GC content	Distance from centromere to the tip of chromosome arm
ρ	0.101 (0.0000)	0.064 (0.0000)	0.037 (0.005)	0.040 (0.0000)
Θ_w	–	−0.443 (0.0000)	−0.245 (0.0000)	0.009 (1.0000)
Gene density		–	0.378 (0.0000)	0.116 (0.0000)
GC content			–	0.126 (0.0000)
Distance to centromere				–

Significance values (*P*-values in parentheses) were obtained using a permutation test to account for the spatial autocorrelation among the 100 kb windows. ρ, population recombination rate; Θ_w, Watterson nucleotide diversity.

somewhat higher than earlier estimates from DArT mapping (Petroli *et al.*, 2012), and a fairly similar recombination pattern across chromosomes with no major recombination deserts as typically evidenced by extensive plateaus of recombination in other plant genomes (Sim *et al.*, 2012; Bauer *et al.*, 2013; Lee *et al.*, 2013).

Genome-wide analysis in *Eucalyptus* supports a fresh look at the extent of LD in forest trees

Recent genome-wide analyses in *Populus* have shown the average LD declining to $r^2 < 0.2$ within *c.* 3–6 kb, substantially more slowly than previous estimates from candidate gene studies (Slavov *et al.*, 2012). LD up to 110 kb was also reported when longer sequence regions of *Cryptomeria japonica* (Moritsuka *et al.*, 2012) were surveyed, and earlier reports in *Populus* (Olson *et al.*, 2010), *Fagus* (Lalague *et al.*, 2014) and flagship conifers (Heuertz *et al.*, 2006; Eckert *et al.*, 2010; Larsson *et al.*, 2013) have also indicated somewhat variable extents of LD when SNPs at variable distances were surveyed in longer candidate gene stretches. Our analysis in *E. grandis* provides solid evidence at a truly genome-wide scale to support a new picture on the extent of LD in an outcrossed undomesticated forest tree genome. By genotyping several thousand SNPs positioned at fairly regularly spaced intervals along the entire genome, across coding and noncoding regions, we were able to capture both moderate and long-range LD. At such a genome-wide scale, r^2 between pairs of SNPs fell to half its initial value within *c.* 3.7–5.7 kb (Figs 3, S3). LD was, however, quite variable across the genome and pairwise estimates of r^2 spanned the entire range of values, from absence to complete LD even up to 50 kb distances (Fig. 3). No difference was seen in the genome-wide extent of LD decay whether using SNPs filtered at MAF > 0.05 or MAF > 0.01 (Fig. S3), although rare SNPs are known to tend to have lower pairwise r^2-values (Pritchard & Przeworski, 2001). This result suggests, however, that our genome-wide estimates of LD based on a relatively large sample of 72 genomes and over 21 000 SNPs are less sensitive to MAF, at least down to 0.01, than LD estimates in short sequence stretches (Lalague *et al.*, 2014), although the true properties of LD around very rare SNPs (MAF < 0.01) still represent an unsettled issue (Pe'er *et al.*, 2006).

Population recombination rates depend on the genomic scale of observation

The extensive genome-wide decay of LD in *Eucalyptus* was further supported by the estimates of genome-wide population-scaled recombination rate (Table 2), one order of magnitude

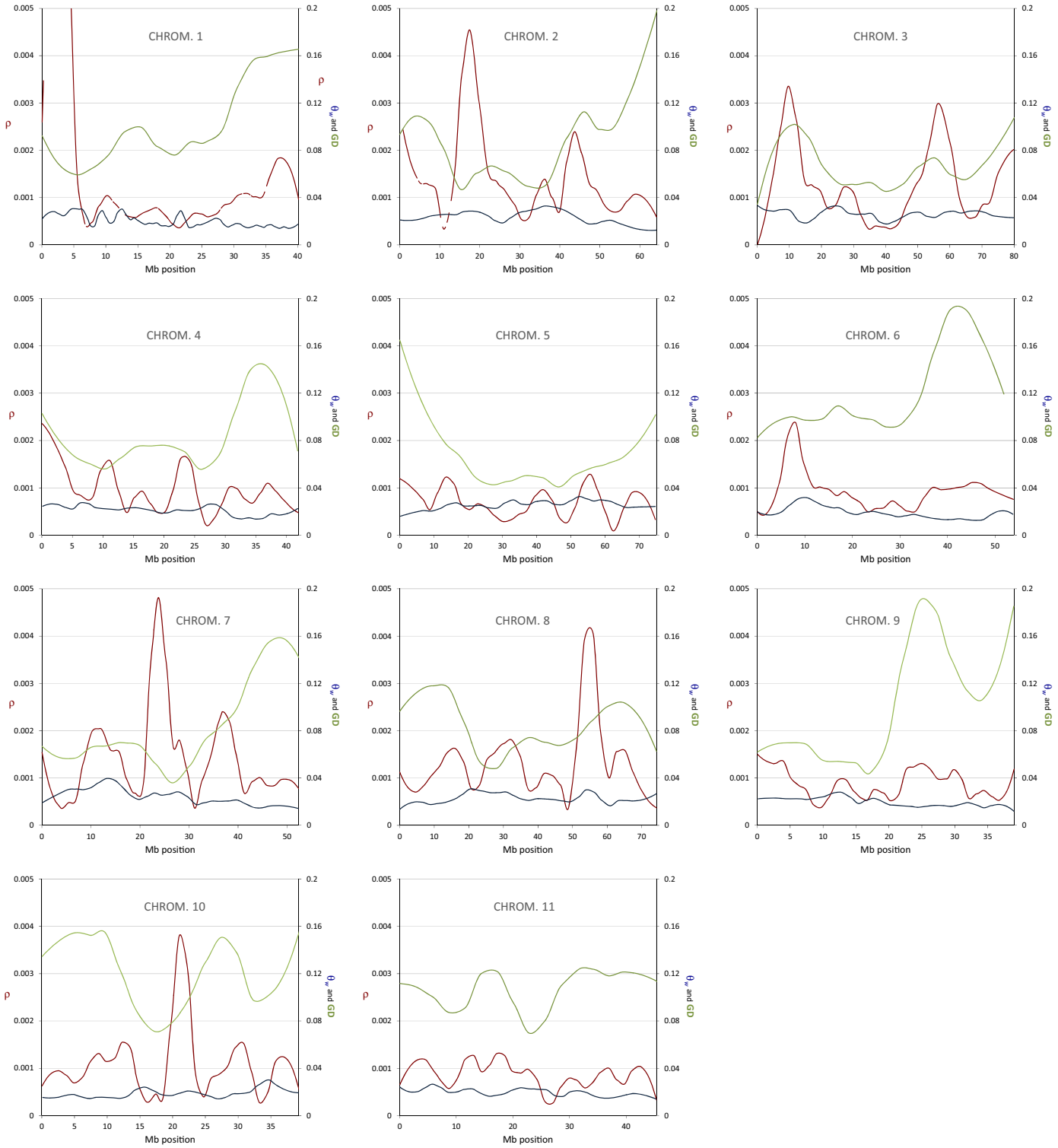

Fig. 5 Genome-wide landscape of population recombination rate (ρ) (red curve), Watterson nucleotide diversity Θ_w (blue curve) and gene density (GD, green curve) across the 11 *Eucalyptus grandis* chromosomes. Note the different scales of the parameters on the *y*-axes. Curves were fitted to the data using locally weighted scatterplot smoothing (LOESS) polynomial regression with a smoothing parameters of 0.2 using R.

lower than the rates previously estimated in candidate genes in angiosperm trees (Ingvarsson, 2008; Lalague *et al.*, 2014) and conifers (Brown *et al.*, 2004; Namroud *et al.*, 2010; Pavy *et al.*, 2012). The genome-wide decay of ρ with pairwise distance found for *Eucalyptus* (Fig. 4) is unexpectedly similar to what has been reported for *Arabidopsis*, where ρ displayed a sharp decrease from 10^{-3} to 2.5×10^{-4} within *c.* 7 kb (Kim *et al.*, 2007). In *Medicago*, another inbred species, a genome-wide

Table 6 Functional categories of the 237 over- and underrepresented annotated genes in the 179 hotspots of recombination detected in the *Eucalyptus grandis* genome

Overrepresented in hotspots			Underrepresented in hotspots		
GO term ID	Description	*P*-value	GO term ID	Description	*P*-value
GO:0055085	Transmembrane transport	0.001	GO:0007165	Signal transduction	0.000
GO:0044238	Primary metabolic process	0.001	GO:0023052	Signaling	0.000
GO:0019538	Protein metabolic process	0.002	GO:0016265	Death	0.001
GO:0009066	Aspartate family amino acid metabolic process	0.006	GO:0007154	Cell communication	0.002
GO:1902578	Single-organism localization	0.006	GO:0006915	Apoptotic process	0.002
GO:0006508	Proteolysis	0.006	GO:0050794	Regulation of cellular process	0.003
GO:0043170	Macromolecule metabolic process	0.007	GO:0002376	Immune system process	0.007
GO:0071704	Organic substance metabolic process	0.008	GO:0006955	Immune response	0.007
GO:0051179	Localization	0.015	GO:0065007	Biological regulation	0.012
GO:0006259	DNA metabolic process	0.022	GO:0050896	Response to stimulus	0.020
GO:0006413	Translational initiation	0.025	GO:0006952	Defense response	0.048
GO:1901135	Carbohydrate derivative metabolic process	0.028			
GO:0044092	Negative regulation of molecular function	0.030			
GO:0055086	Nucleobase-containing small molecule metabolic process	0.032			
GO:0006555	Methionine metabolic process	0.032			
GO:0006184	GTP catabolic process	0.033			
GO:0015671	Oxygen transport	0.033			
GO:0015669	Gas transport	0.033			
GO:0018279	Protein N-linked glycosylation via asparagine	0.033			
GO:0018196	Peptidyl-asparagine modification	0.033			

GO, gene ontology.

estimate of $\rho = 1.8 \times 10^{-3}\,\mathrm{bp}^{-1}$ was reported and LD also decayed within *c.* 5 kb (Branca *et al.*, 2011). Such an unforeseen outcome was also pointed out by Slavov *et al.* (2012), and among the several potential reasons to explain this result, the one that remained is the same one we also propose: published LD estimates based on limited stretches of sequence in a few genes cannot provide accurate expectations of the genome-wide extent of LD. In our study, we nevertheless provided further evidence in support of this proposition: orders of magnitude higher estimates of ρ were seen at the 0–100 bp range when compared with longer ranges. This result corroborates early studies in *Arabidopsis*, humans and *Drosophila*, where LD based on closely spaced SNPs was deemed unsuitable for predicting long-range LD (Frisse *et al.*, 2001; Pritchard & Przeworski, 2001; Andolfatto & Wall, 2003; Nordborg *et al.*, 2005; Kim *et al.*, 2007; Tenesa *et al.*, 2007). Our study in *E. grandis* therefore emphasizes that, more important than estimating the average extent of LD, is understanding its variance as a function of recombination over moderate to large genomic distances (Goldstein & Weale, 2001). Among other things, variation in recombination rates might result from the effect of gene conversion-like-processes unaccompanied by crossovers. While it is challenging to estimate the relative role of gene conversion from genotyping or sequence data (Morrell *et al.*, 2006), the few attempts to assess its influence on the estimates of recombination in *Arabidopsis* have shown a clear impact. Contributions at least equivalent to crossing over (Lu *et al.*, 2012) up to a 130-fold higher (Yang *et al.*, 2012) were reported, although unquestionably occurring along short stretches of DNA (Wijnker *et al.*, 2013), corroborating initial propositions that gene conversion may interrupt LD at short distances, while leaving LD largely unaffected at long distances (Andolfatto & Nordborg, 1998).

The evolutionary history of *E. grandis*

The variable rates of recombination observed at different pairwise SNP distances are known to reflect fluctuations in the effective population size (N_e) over different timescales. In other words, short-range LD is expected to reflect more ancestral N_e, while longer-range LD is expected to reflect more recent N_e (Hayes *et al.*, 2003; Tenesa *et al.*, 2007; Rogers, 2014). Given the high variance observed for ρ at the short distance range, we argue that the resulting estimate of $N_e = 1116\,352$ is probably not very credible. Short-range recombination and LD data may, in fact, lead to invalid inferences about population histories if gene conversion is not taken into account (Frisse *et al.*, 2001). Additionally, estimates of ρ obtained at distances < 250 bp were shown to violate the assumptions of the neutral model (Lynch *et al.*, 2014). We therefore considered that the pairwise SNP distances at moderate and long ranges were the ones providing estimates most likely consistent with the past demographic history of *E. grandis*, with $N_e = 112\,421$ reflecting the more ancient past, and $N_e = 11\,557$ informing demographic processes in the more recent past. To the best of our knowledge, these are the first estimates of N_e reported for eucalypts. Using the estimates of effective population size for chromosome 8, we built ARGs and dated the two corresponding events in the past evolutionary history of *E. grandis*. The TMRCA estimated from N_0 at the 0–2 kb range provided an estimate of 3.2 million yr ago (Ma) with a credible interval of 1.7–4.8 Ma. This date probably marks the lineage split of *E. grandis*, consistent with fossil records that

place species radiation within subgenus *Symphomyrtus* at 2–5 Ma, therefore strongly supporting the hypothesis of a more recent divergence time of the eucalypts (Ladiges *et al.*, 2003), also coherent with the high degrees of SNP sharing with other species of the subgenus (Silva-Junior *et al.*, 2015). The TMRCA based on N_c at the genome-wide level was dated at 469.5 ka (434.7–504.2 ka), a timing consistent with the period of the first ice age in the Southern Hemisphere (390–730 ka) (Fitzsimons & Colhoun, 1991). Besides providing TMRCAs for *E. grandis*, the ARG-based estimates allowed us to infer that the ancestral decline in population size from N_0 to the current N_c probably occurred in a progressive fashion, instead of abruptly, as a consequence of population bottlenecks. This is supported by the fact that the estimated TMRCA at N_c (434.70–504.23 ka) nearly matches the mean coalescence time for a neutral mutation in a population of constant size given by $4N_c$ generations ago ($4 \times 9658 \times 10$ yr per generation = 386 ka) (Lynch *et al.*, 2014). Such a proposition is also consistent with the fact that, while additional glaciations and strong climatic fluctuations occurred since the first ice age in other continents, notably in Europe, Australia was largely unaffected by them, with minimal impact on the flora (White, 1998), precluding the likelihood of major bottleneck events.

Mutation is the main driver of genetic diversity in *Eucalyptus*

Our genome-wide estimate of nucleotide diversity in *E. grandis* is higher than earlier estimates from EST data (0.005) (Novaes *et al.*, 2008) and from a sample of seven lignification genes (0.0065) (Mandrou *et al.*, 2014), while consistent with estimates from resequencing genomic segments of genes for range-wide samples of other *Eucalyptus* species (0.16–0.03) (Kulheim *et al.*, 2009). Our results further confirm that *E. grandis* is positioned at the high end of diversity among plants in general, slightly higher than *Populus trichocarpa* and an order of magnitude higher than most conifers (González-Martínez Dillon *et al.*, 2011). Using the measures of recombination and mutation estimated at the moderate pairwise distance range (0–2 kb), we estimated a genome-wide ratio of $\mu/c = 1.55$ or $\rho/\Theta_w = 0.645$, indicating that mutation has a greater impact than recombination in shaping the genomic diversity of *E. grandis* (Table 3). This result contrasts with the much higher estimates of ρ/Θ_w reported from candidate gene studies in *Populus* (4.47) (Ingvarsson *et al.*, 2008) and *Fagus* (2.85) (Lalague *et al.*, 2014), but in the same range of recent genome-wide estimates for *Mimulus* (0.8) (Hellsten *et al.*, 2013) and *Medicago* (0.29) (Branca *et al.*, 2011).

Recombination as a driver of diversity in genes for normal metabolic processes in *Eucalyptus*

Using sliding windows of 100 kb, recombination was found to be significantly correlated with all genomic features analyzed, although the magnitudes were generally small (Table 5). The positive gradient of recombination with increasing gene density and distance from the centromere is consistent with reports in other plant genomes (Mezard, 2006; Gaut *et al.*, 2007; Slavov *et al.*, 2012). However, the low magnitudes of correlations in *E. grandis* suggest that other factors unaccounted for in our study might impact the distribution of recombination, granting analyses of additional genomic features at finer scales. A relatively conservative analysis of recombination hotspots was carried out (Methods S5), resulting in a considerably smaller number of regions (179) in *E. grandis*, when compared with estimates in other plant genomes, 3235 in *Mimulus* (Hellsten *et al.*, 2013) and 606 in *Populus* (Slavov *et al.*, 2012). The investigation of recombination hotspots at finer genomic scales and their relationship with additional sequence features also await future studies. Nevertheless, a first look at the hotspots detected showed a statistically significant enrichment for GO terms associated with chemical reactions and pathways that are part of the normal metabolic processes, while terms associated with cell communication, signaling and response to stimuli were underrepresented (Table 6). We speculate that a lower recombination rate in genes involved in responding to stimuli suggests that recombination does not play a major role in the evolution of *E. grandis* in facing widespread and rapid environmental changes. Mutation possibly occurs at a rate rapid enough to prevent abrupt population decline, consistent with our finding of $\rho/\Theta_w = 0.645$. On the other hand, the amplification of sequence variation in genes contained in hotspots is consistent with the positive genome-wide correlation between ρ and Θ_w, and fits well with biochemical and metabolic adaptations of the species that may facilitate its survival under suboptimal environmental conditions. This variation can be thought to provide additional flexibility of plant metabolism that may help trees to cope with unavoidable environmental changes imposed upon them. This result is in striking contrast to recent findings in rice (Si *et al.*, 2015), probably reflecting key differences in adaptation strategies between annual and perennial plants.

Conclusions and perspectives

Besides the fundamental aspects related to the evolutionary history of the eucalypts, our genome-wide estimates of recombination and nucleotide diversity and the extensive LD found have considerable practical importance to molecular breeding and the overall study of complex trait variation. SNP density needed for association studies and genomic prediction is dependent upon the degree to which extrapolations from larger-scale estimates of LD predict the degree of association between genetic markers and causative mutations. With a genome-wide usable LD within *c.* 4–6 kb, *c.* 10^5 tag SNPs should provide satisfactory coverage for such applications in *E. grandis*. Genotyping at such density has become technically accessible, warranting a more positive outlook on the prospects of GWAS in eucalypts. We acknowledge, however, that the slower LD decay will complicate the precise pinpointing of the causative quantitative trait nucleotide polymorphism, if such a feat will ever be possible (Rockman, 2012). The EuCHIP60K used in this study, supplying on average one polymorphic SNP every 20 kb, with 80% of the SNPs at ≤ 10 kb of a gene (Silva-Junior *et al.*, 2015), represents a good start to empower upcoming GWAS experiments. Moreover, this

genotyping density corresponds to *c.* 20 markers cM^{-1}, leading to expected accuracies of genomic selection above 0.80 even in breeding populations with relatively large effective sizes ($N_e \approx 100$) (Grattapaglia & Resende, 2011). In elite breeding populations of smaller effective size ($N_e \approx 20$–30), where LD is extending even further, the current EuCHIP60K provides abundant power to capture the majority of linked effects in predictive models. For such elite breeding populations, lower-density genotyping platforms on the order of 3000–5000 with well distributed SNPs should provide suitable marker density for routine GS at considerably lower genotyping costs. In addition to expanding our knowledge on key genomic features of the *E. grandis* genome, this study also demonstrates the exceptional perspectives that lie ahead in extending similar genome-wide studies to the several other *Eucalyptus* species using pooled whole-genome sequencing. Despite the data analysis challenges involved, this cost-effective alternative to sequencing individuals separately (Schlotterer *et al.*, 2014), coupled to the availability of a high-quality genome and the extensive genomic collinearity among the eucalypts, should prove very valuable to advance a broad range of research questions in genome evolution, domestication and ecological genomics in species of the genus.

Acknowledgements

This work was supported by PRONEX-FAP-DF grant 2009/00106-8 'NEXTREE – Nucleus of Excellence in Applied Forest Tree Genomics' and EMBRAPA grant 03.11.01.007.00.00 to D.G. O.B.S-J. was sponsored by an EMBRAPA doctoral fellowship and D.G. by a CNPq research fellowship. We thank the technical assistance of Danielle A. Faria with DNA preparation. We gratefully acknowledge the comments of the three reviewers that helped us to improve our manuscript.

References

Andolfatto P, Nordborg M. 1998. The effect of gene conversion on intralocus associations. *Genetics* **148**: 1397–1399.

Andolfatto P, Wall JD. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* **165**: 1289–1305.

Auton A, Myers S, McVean G. 2014. Identifying recombination hotspots using population genetic data. *arXiv*: 1403.4264.

Bartholome J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion JM. 2014. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* **206**: 1283–1296.

Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W *et al.* 2013. Intraspecific variation of recombination rate in maize. *Genome Biology* **14**: R103.

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 Sciences, USA* **108**: E864–E870.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences, USA* **101**: 15255–15260.

Chakravarti A. 1991. A graphical representation of genetic and physical maps – the Marey map. *Genomics* **11**: 219–222.

Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics* **8**: e1003090.

Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. 2009. Genetic recombination and molecular evolution. *Cold Spring Harbor Symposia on Quantitative Biology* **74**: 177–186.

Denis M, Favreau B, Ueno S, Camus-Kulandaivelu L, Chaix G, Gion JM, Nourrisier-Mountou S, Polidori J, Bouvet JM. 2013. Genetic variation of wood chemical traits and association with underlying genes in *Eucalyptus urophylla*. *Tree Genetics & Genomes* **9**: 927–942.

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

Eldridge K, Davidson J, Harwood C, van Wyk G. 1993. *Eucalypt domestication and breeding*. Oxford, UK: Clarendon Press.

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.

Fitzsimons SJ, Colhoun EA. 1991. Pleistocene glaciation of the King Valley, Western Tasmania, Australia. *Quaternary Research* **36**: 135–156.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics* **69**: 831–843.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics* **8**: 77–84.

Goddard ME, Hayes BJ. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**: 381–391.

Goldstein DB, Weale ME. 2001. Population genomics: linkage disequilibrium holds the key. *Current Biology* **11**: R576–R579.

González-Martínez SC, Dillon S, Garnier-Géré P, Krutovsky K, Alía R, Burgarella C, Eckert A, García-Gil M, Grivet D, Heuertz M *et al.* 2011. Patterns of nucleotide diversity and association mapping. In: Plomion C, Bousquet J, Kole C, eds. *Genetics genomics and breeding of conifers*. Enfield, NH, USA: CRC Press, 239–275.

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

Grattapaglia D, Resende MDV. 2011. Genomic selection in forest tree breeding. *Tree Genetics & Genomes* **7**: 241–255.

Grattapaglia D, Sederoff R. 1994. Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross – mapping strategy and RAPD markers. *Genetics* **137**: 1121–1137.

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.

Haubold B, Pfaffelhuber P, Lynch M. 2010. mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology* **19**: 277–284.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13**: 635–643.

Hellenthal G, Stephens M. 2006. Insights into recombination from population genetic variation. *Current Opinion in Genetics & Development* **16**: 565–572.

Hellsten U, Wright KM, Jenkins J, Shu SQ, Yuan YW, Wessler SR, Schmutz J, Willis JH, Rokhsar DS. 2013. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences, USA* **110**: 19478–19482.

Henderson IR. 2012. Control of meiotic recombination frequency in plant genomes. *Current Opinion in Plant Biology* **15**: 556–561.

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**: 54–78.

Hudson C, Kullan A, Freeman J, Faria D, Grattapaglia D, Kilian A, Myburg A, Potts B, Vaillancourt R. 2012. High synteny and colinearity among *Eucalyptus* genomes revealed by high-density comparative genetic mapping. *Tree Genetics & Genomes* **8**: 339–352.

Hudson RR. 1987. Estimating the recombination parameter of a finite population-model without selection. *Genetical Research* **50**: 245–250.

Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180**: 329–340.

Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S. 2008. Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (*Populus tremula*, Salicaceae). *Genetics* **178**: 2217–2226.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **39**: 1151–1155.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlotterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**: e15925.

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

Ladiges PY, Urdovicic F. 2005. Comment on the molecular dating of the age of eucalypts. *Australian Systematic Botany* **18**: 291–293.

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.

Lalague H, Csillery K, Oddou-Muratorio S, Safrana J, de Quattro C, Fady B, Gonzalez-Martinez SC, Vendramin GG. 2014. Nucleotide diversity and linkage disequilibrium at 58 stress response and phenology candidate genes in a European beech (*Fagus sylvatica* L.) population from southeastern France. *Tree Genetics & Genomes* **10**: 15–26.

Larsson H, Kallman T, Gyllenstrand N, Lascoux M. 2013. Distribution of long-range linkage disequilibrium and Tajima's D values in Scandinavian populations of Norway spruce (*Picea abies*). *G3-Genes Genomes. Genetics* **3**: 795–806.

Lee WK, Kim N, Kim J, Moon JK, Jeong N, Choi IY, Kim SC, Chung WH, Kim HS, Lee SH *et al.* 2013. Dynamic genetic features of chromosomes revealed by comparison of soybean genetic and sequence-based physical maps. *Theoretical and Applied Genetics* **126**: 1103–1119.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.

Lu PL, Han XW, Qi J, Yang JG, Wijeratne AJ, Li T, Ma H. 2012. Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by resequencing *Landsberg erecta* and all four products of a single meiosis. *Genome Research* **22**: 508–518.

Lynch M, Xu S, Maruki T, Jiang X, Pfaffelhuber P, Haubold B. 2014. Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics* **198**: 269–281.

Mandrou E, Denis M, Plomion C, Salin F, Mortier F, Gion JM. 2014. Nucleotide diversity in lignification genes and QTNs for lignin quality in a multi-parental population of *Eucalyptus urophylla*. *Tree Genetics & Genomes* **10**: 1281–1290.

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **108**: 285–291.

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

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581–584.

Mezard C. 2006. Meiotic recombination hotspots in plants. *Biochemical Society Transactions* **34**: 531–534.

Moritsuka E, Hisataka Y, Tamura M, Uchiyama K, Watanabe A, Tsumura Y, Tachida H. 2012. Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica*. *Genetics* **190**: 1145–1148.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. *Genetics* **173**: 1705–1723.

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.

Namroud MC, Guillet-Claude C, Mackay J, Isabel N, Bousquet J. 2010. Molecular evolution of regulatory genes in spruces from different species and continents: heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. *Journal of Molecular Evolution* **70**: 371–386.

Neale DB, Ingvarsson PK. 2008. Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**: 149–155.

Neale DB, Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* **9**: 325–330.

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.

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.

Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186**: 526–536.

Paape T, Zhou P, Branca A, Briskine R, Young N, Tiffin P. 2012. Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome. *Genome Biology and Evolution* **4**: 726–737.

Pavy N, Namroud MC, Gagnon F, Isabel N, Bousquet J. 2012. The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity* **108**: 273–284.

Pe'er I, Chretien YR, de Bakker PIW, Barrett JC, Daly MJ, Altshuler DM. 2006. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *American Journal of Human Genetics* **78**: 588–603.

Petes TD. 2001. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics* **2**: 360–369.

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.

Poke FS, Vaillancourt RE, Elliott RC, Reid JB. 2003. Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (*CCR*) and cinnamyl alcohol dehydrogenase 2 (*CAD2*). *Molecular Breeding* **12**: 107–118.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**: 1–14.

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.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics* **10**: e1004342.

Rockman MV. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**: 1–17.

Rogers AR. 2014. How population growth affects linkage disequilibrium. *Genetics* **197**: 1329–1341.

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.

Si W, Yuan Y, Huang J, Zhang X, Zhang Y, Zhang Y, Tian D, Wang C, Yang Y, Yang S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F_2 plants. *New Phytologist* **206**: 1491–1502.

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

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.

Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA *et al.* 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist* **196**: 713–725.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**: 978–989.

Stumpf MPH, McVean GAT. 2003. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics* 4: 959–968.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**: 520–526.

Thavamanikumar S, McManus LJ, Tibbits JFG, Bossinger G. 2011. The significance of single nucleotide polymorphisms (SNPs) in *Eucalyptus globulus* breeding programs. *Australian Forestry* **74**: 23–29.

Thumma BR, Nolan MR, Evans R, Moran GF. 2005. Polymorphisms in cinnamoyl CoA reductase (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. *Genetics* **171**: 1257–1265.

Van Ooijen JW, Voorrips RE. 2001. *JoinMap 3.0 software for the calculation of genetic linkage maps*. Wageningen, the Netherlands: Plant Research International.

Van Os H, Stam P, Visser RGF, Van Eck HJ. 2005. RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics* **112**: 30–40.

Voorrips RE. 2002. MapChart: software for the graphical presentation of linkage maps and QTLs. *Journal of Heredity* **93**: 77–78.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7: 256–276.

White ME. 1998. *The greening of Gondwana, 3rd edn.* Kenthurst, NSW, Australia: Rosenberg Publishing.

Wijnker E, James GV, Ding J, Becker F, Klasen JR, Rawat V, Rowan BA, de Jong DF, de Snoo CB, Zapata L *et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* **2**: e01426.

Wimmer V, Albrecht T, Auinger HJ, Schon CC. 2012. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* **28**: 2086–2087.

Yang SH, Yuan Y, Wang L, Li J, Wang W, Liu HX, Chen JQ, Hurst LD, Tian DC. 2012. Great majority of recombination events in *Arabidopsis* are gene conversion events. *Proceedings of the National Academy of Sciences, USA* **109**: 20992–20997.

Supporting Information

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

Fig. S1 Marey maps of the 11 *Eucalyptus grandis* chromosomes based on the high-density linkage maps.

Fig. S2 Genome-wide decay of average r^2 within 1 cM bins of the *Eucalyptus grandis* linkage map.

Fig. S3 Decay of LD on chromosome 1 based on the *Eucalyptus grandis* genome versions and the effect of minimum allele frequency on the decay of LD.

Fig. S4 Comparative landscapes of population recombination rate estimated using the *Eucalyptus grandis* genome versions 1.1 and 2.0.

Fig. S5 Boxplots of the chromosome-specific population recombination rates using the *Eucalyptus grandis* genome versions 1.1 and 2.0.

Table S1 Mapping statistics of the *Eucalyptus grandis* and *E. urophylla* linkage maps

Table S2 Pearson's correlations among recombination rate and genomic features for the 11 *Eucalyptus grandis* chromosomes

Methods S1 SNP segregation data processing for linkage mapping.

Methods S2 Estimating population-scaled recombination rates from genome-wide EUChip60K SNP data.

Methods S3 SNP calling from pools for population recombination and nucleotide diversity analyses using mlRho.

Methods S4 Ancestral recombination graph analyses with ARGWEAVER.

Methods S5 Recombination hotspot detection with LDHOT.

Methods S6 Recombination hotspot genomic annotation.

Notes S1 Mapchart datafile with the SNP cM and base pair positions on the high-density linkage maps of *Eucalyptus grandis* and *E. urophylla* aligned to the *Eucalyptus* reference genome 1.1.

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.