

Genomic consequences of artificial selection during early domestication of a wood fibre crop

Marja M. Mostert-O'Neill¹ , Hannah Tate¹ , S. Melissa Reynolds¹ , Makobatjatji M. Mphahlele^{1,2} , Gert van den Berg³, Steve D. Verry⁴ , Juan J. Acosta⁵ , Justin O. Borevitz⁶  and Alexander A. Myburg¹ 

¹Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria 0028, South Africa; ²Mondi Forests, Tree Improvement Technology Programme, Trahar Technology Centre – TTC, Mountain Home Estate, Off Dennis Shepstone Dr., Hilton 3245, South Africa; ³Sappi Forests Research, Shaw Research Centre, PO Box 473, Howick 3290, South Africa; ⁴Creation Breeding Innovations, 75 Kafue St., Lynnwood Glen 0081, South Africa; ⁵Camcore, Department of Forestry and Environmental Resources, North Carolina State University, PO Box 7626, Raleigh, NC 27695, USA; ⁶Research School of Biology and Centre for Biodiversity Analysis, ARC Centre of Excellence in Plant Energy Biology, Australian National University, Canberra, ACT 0200, Australia

Summary

Author for correspondence:
Alexander A. Myburg
Email: zander.myburg@fabi.up.ac.za

Received: 13 October 2021
Accepted: 20 May 2022

New Phytologist (2022) 235: 1944–1956
doi: 10.1111/nph.18297

Key words: artificial selection, domestication, eucalypt, forestry, population genomics, selection signatures.

- From its origins in Australia, *Eucalyptus grandis* has spread to every continent, except Antarctica, as a wood crop. It has been cultivated and bred for over 100 yr in places such as South Africa. Unlike most annual crops and fruit trees, domestication of *E. grandis* is still in its infancy, representing a unique opportunity to interrogate the genomic consequences of artificial selection early in the domestication process.
- To determine how a century of artificial selection has changed the genome of *E. grandis*, we generated single nucleotide polymorphism genotypes for 1080 individuals from three advanced South African breeding programmes using the EUChip60K chip, and investigated population structure and genome-wide differentiation patterns relative to wild progenitors.
- Breeding and wild populations appeared genetically distinct. We found genomic evidence of evolutionary processes known to have occurred in other plant domesticates, including interspecific introgression and intraspecific infusion from wild material. Furthermore, we found genomic regions with increased linkage disequilibrium and genetic differentiation, putatively representing early soft sweeps of selection.
- This is, to our knowledge, the first study of genomic signatures of domestication in a timber species looking beyond the first few generations of cultivation. Our findings highlight the importance of intra- and interspecific hybridization during early domestication.

Introduction

Understanding changes in genomic architectures underlying the domestication of plants aids in the discovery of genetic targets for crop improvement and enhances our knowledge of the evolutionary forces involved in species adaptation (Ross-Ibarra *et al.*, 2007; Purugganan & Fuller, 2009; Olsen & Wendel, 2013). For most domesticates, the genotypes intermediate between wild and domesticated are missing. In some cases, even the wild progenitors remain disputed (Cornille *et al.*, 2012; Wu *et al.*, 2014), complicating efforts to untangle the evolutionary forces that shaped the genomes of domesticates and to detect genomic signatures of artificial selection. Current breeding practices in plantation forestry (Isik *et al.*, 2015) mimic that of early fruit and annual crop domestication, including exploitation of interspecific hybrids (Wu *et al.*, 2014), genetic infusions (intentional introduction of unrelated genetic diversity from the same species) from wild, unimproved genotypes (Cornille *et al.*, 2012; Hufford *et al.*, 2012), and vegetative propagation of favourable genetic combinations (Myles *et al.*, 2011;

Cornille *et al.*, 2014). In addition to traits associated with general plant domestication syndrome such as determinate growth with reduced branching and reallocation of resources to the harvested parts of the plant (Ross-Ibarra *et al.*, 2007), forest tree domestication could also include changes in wood properties, such as wood density and wood chemistry, where breeders directly selected for such traits (Tuskan, 2007; Thomas *et al.*, 2018).

Most cultivated forestry species are fewer than three generations removed from their wild progenitors. As such, genetic investigations have focused on early responses to cultivation (Jones *et al.*, 2006; Bouffier *et al.*, 2008; Varghese *et al.*, 2009; De La Torre *et al.*, 2014; Skråppa & Steffenrem, 2016) or genomic responses to natural selection (Prunier *et al.*, 2011; Evans *et al.*, 2014; Acosta *et al.*, 2019; Collevatti *et al.*, 2019; Wang *et al.*, 2020). An exception is the domestication of *Eucalyptus grandis*, a forestry species that has been grown and bred *ex situ* for over a century, representing a unique opportunity to observe the genomic consequences of ongoing formal and informal artificial selection early in the domestication process.

Cultivation of *E. grandis* as a timber and wood fibre crop has been ongoing for over 100 yr in various exotic environments around the world (Bennett, 2011). From its origins in Australia, the species has been transplanted to every continent except Antarctica (Marco, 1991; Rockwood & Meskimen, 1991; Huoran *et al.*, 1992; Chaix *et al.*, 2003; Hunde *et al.*, 2003; Dos Santos *et al.*, 2004; Verryin *et al.*, 2009; Luo *et al.*, 2010; Boulay *et al.*, 2012; Santos *et al.*, 2017). Its fast growth has been further improved in exotic breeding programmes where artificial selection resulted in trees reaching harvestable age 10–15% earlier (Verryin, 2002), and produced increases of 16% in stem volume per generation of breeding (Meskimen, 1983). These improvements in growth resulted, in part, from the selection of genotypes better adapted to the exotic environment (Rockwood & Meskimen, 1991) and an expanded range of genotypes produced by intraspecific hybridization resulting from crosses between individuals from different provenances. Other economically important traits such as stem form and wood properties were also improved by artificial selection (Verryin *et al.*, 2009). Quantitative genetics studies of early *E. grandis* breeding trials have therefore indicated substantial genetic gains for production phenotypes, but it is not clear how these genetic gains have manifested in the genomes of these trees.

As reviewed by Ross-Ibarra *et al.* (2007), Purugganan & Fuller (2009) and Olsen & Wendel (2013), most domestication studies use one of two broad strategies to identify candidate genetic variants that can subsequently be used for functional verification of their role in domestication traits and/or targeted for crop improvement. The first involves quantitative trait locus (QTL) mapping or genome-wide association studies to identify genomic regions associated with a trait of economic importance. Published examples of these so-called top-down studies (starting with the trait to identify underlying genes) in *E. grandis* and its hybrids include the detection of QTLs associated with vegetative propagation (Grattapaglia *et al.*, 1995; Marques *et al.*, 2002), growth and wood properties (Grattapaglia *et al.*, 1996; Rocha *et al.*, 2007; Kullán *et al.*, 2012), and resistance to pests and pathogens (Alves *et al.*, 2012; Mhoswa *et al.*, 2020). These strategies generally detect large-effect loci segregating in a particular family. To aggregate the genetic variation of genome-wide small effects underlying quantitative phenotypes, genomic selection has also been used in the species (Mphahlele *et al.*, 2020).

Complementing this, the second strategy starts by comparing genome-wide patterns of genetic diversity and differentiation among and between domesticated and wild progenitor populations to identify regions of the genome that show signatures of selection (Ross-Ibarra *et al.*, 2007; Purugganan & Fuller, 2009; Olsen & Wendel, 2013). Gene Ontology (GO) terms (Ashburner *et al.*, 2000) associated with genes within these regions can subsequently reveal the biological processes under artificial selection. Since this bottom-up strategy is phenotype-naïve, it could also reveal traits that have been selected unintentionally. As reviewed by Cutter & Payseur (2013), the number of genes underlying the selected traits, strength of selection on individual loci, recombination rates and number of generations determine

our ability to decipher the genomic footprints left by recurrent selection. Furthermore, this strategy requires extensive genomics resources including genome-wide genotyping tools and an annotated reference genome for the identification of genes in linkage with genomic loci under selection. The economic importance of *E. grandis* as a wood fibre crop, as a pure species or as a hybrid partner, has led to the development of numerous transcriptomic (Mangwanda *et al.*, 2015; Oates *et al.*, 2015; Vining *et al.*, 2015) and genomic resources, including annotated nuclear and organellar genome sequences (Myburg *et al.*, 2014; Bartholomé *et al.*, 2015; Pinard *et al.*, 2019b) and a high-throughput EUChip60K single nucleotide polymorphism (SNP) array (Silva-Junior *et al.*, 2015). This raises the possibility of combining trait-based gene discovery efforts with a bottom-up approach to uncover genomic regions under artificial selection in the genomes of *E. grandis* individuals in early domestication.

South Africa has some of the most advanced *E. grandis* breeding programmes globally. In this study, we investigate the genomic consequences of a century of cultivation in three such programmes. The three populations share a common gene pool, which originated from multiple seed imports from Australia starting as early as 1896 (Bennett, 2011). Formal breeding programmes, where growth and sawn timber quality were under selection, commenced in the 1960s with dedicated provenance trials using newly imported seed lots from across most of the latitudinal range of the species (Poynton, 1979), and material that has been selected and advanced informally in South Africa for up to five generations previously. These programmes followed tree improvement methodologies later described by Zobel & Talbert (1984), with an average breeding age of 8 yr (although *E. grandis* can flower from as young as 3 yr depending on field conditions). The breeding objective was to make genetic gains whilst maintaining genetic diversity (since forest trees have high genetic load and suffer inbreeding depression), and as such, the top-performing individual(s) for each family were advanced and on average 300 families were maintained and selected based on family means within and across sites. In the process, some families would not go forward in the breeding programme. Since the 1990s, this germplasm was advanced for three to five generations in separate private breeding programmes by forestry companies Hans Merensky, Mondi and Sappi (Fig. 1a).

First, we aim to test the hypothesis that a century of domestication has resulted in *E. grandis* genotypes that are genetically distinct from their wild progenitors. We also investigate the possibility that interspecific hybridization and recent infusions from unimproved, wild material have contributed to the genetic diversity in South African breeding populations, as is suggested to have occurred in the domestication of other crops (He *et al.*, 2011; Myles *et al.*, 2011; Cornille *et al.*, 2012; Wu *et al.*, 2014; Baute *et al.*, 2015). This is done by elucidating the population structure and genetic differentiation of breeding populations relative to wild *E. grandis* populations (Mostert-O'Neill *et al.*, 2021) and species with which *E. grandis* could have hybridized *ex situ* (Silva-Junior *et al.*, 2015). Next, we define the core *E. grandis* breeding germplasm, representative

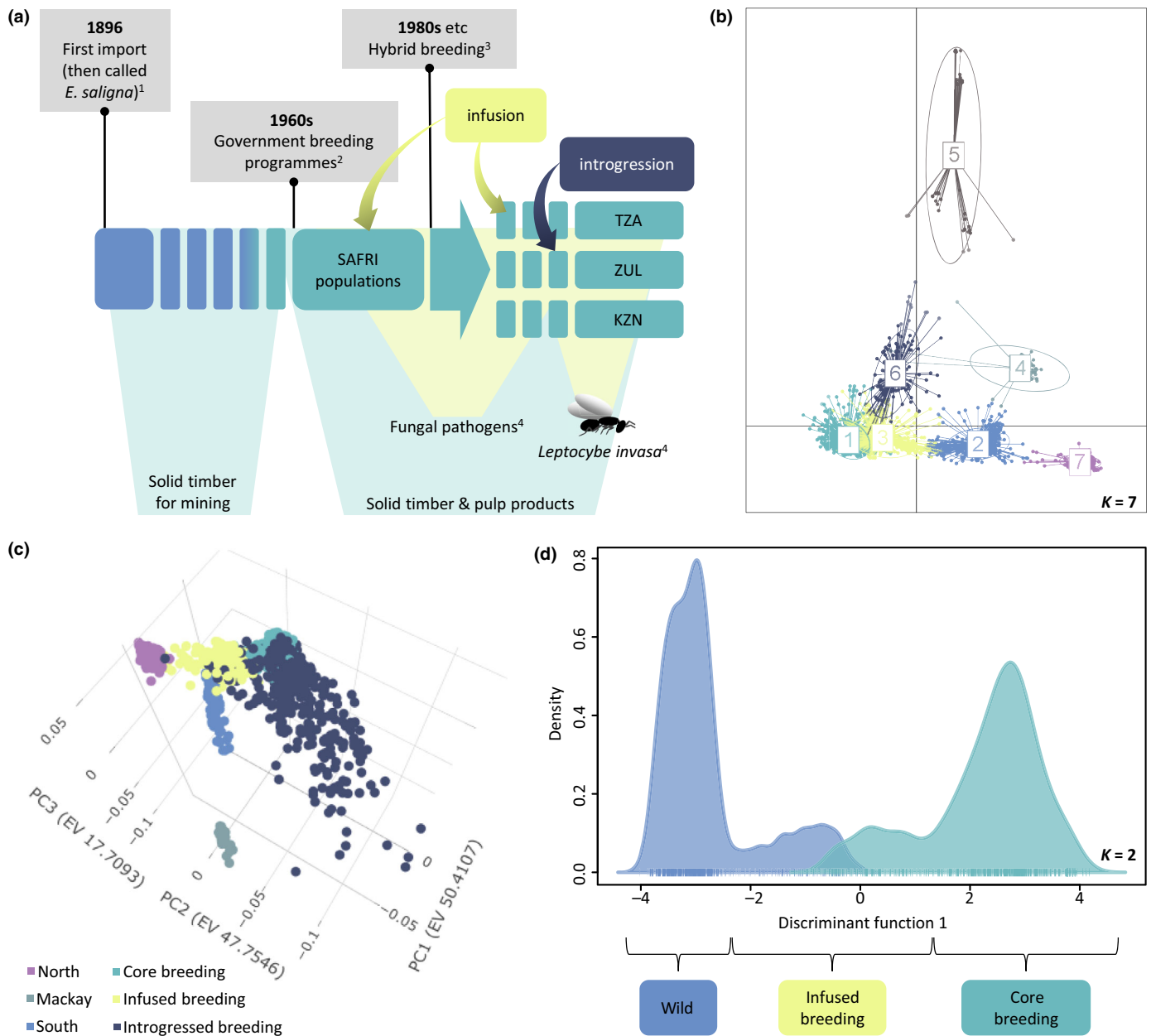


Fig. 1 Breeding *Eucalyptus grandis* genetic differentiation and population structure relative to wild progenitors and potential introgressing species. (a) Diagram of the plantation and breeding history of the three South African *E. grandis* populations, TZA, ZUL and KZN, with main end-product (turquoise shade) and known biotic challenges (pale yellow shade) given below, and sources of genetic change (breeding practices, intentional genetic infusions and unintentional introgression) given above the main timeline. ¹Bennett (2011); ²Van Wyk & Roeder (1978); ³Denison & Kietzka (1993); ⁴Wingfield *et al.* (2008). (b) Discriminant analysis of principal components (DAPC; see Supporting Information Fig. S1 for supporting BIC plot) at $K = 7$ with two dimensions shown (24 306 informative single nucleotide polymorphisms (SNPs) used) of core (cluster 1), infused (cluster 2) and introgressed (cluster 6) breeding *E. grandis*, and North (cluster 7), South (cluster 2) and Mackay (cluster 4) wild subpopulations. Cluster 5 contained other species that could potentially introgress with breeding *E. grandis*, including *E. urophylla*, *E. saligna* and *E. grandis* × *E. urophylla* (GU) hybrids as obtained from Silva-Junior *et al.* (2015). (c) Population structure principal components analysis plot for the first three principal components (eigenvalues given in parentheses) of all breeding *E. grandis* and wild progenitor subpopulations (23 661 informative SNPs used, see Fig. S2 for supporting scree plot and <https://chart-studio.plotly.com/~Marja/125/#/> for an interactive version), excluding species that could potentially introgress in breeding populations. (d) DAPC analysis at $K = 2$ of all breeding *E. grandis* (excluding introgressed individuals), and Northern and Southern wild subpopulations, used for identification of infused breeding individuals (23 661 informative SNPs used).

of the advanced-generation population that has been under selection for a century. In these trees, we detect genomic regions that show potential signatures of selection. Variants that exhibit localized differentiation patterns between breeding

and wild populations are identified. We also compare genome-wide patterns of heterozygosity and linkage disequilibrium (LD) in breeding material to that in wild progenitors as support for potential signatures of selection.

Materials and Methods

Study population, SNP genotyping and population structure

Individuals were sampled from three core, open-pollination breeding programmes (Table 1). The first population, TZA, consisted of 285 fifth-generation (since privatization in the 1990s) individuals, representing 282 families, which were bred in the temperate Tzaneen area of the Mpumalanga province. The second population, ZUL, represented by 43 families with 248 third-generation individuals, was bred for subtropical climates in Zululand, in northern KwaZulu-Natal province. The KZN population consisted of a core breeding population of 547 third- and fourth-generation individuals (62 families) established from trees bred in temperate and subtropical sites in KwaZulu-Natal. DNA was isolated from leaf or cambial tissues using the Nucleospin DNA extraction kit (Machery-Nagel, Düren, Germany) and used for SNP genotyping with the EUChip60K chip (Silva-Junior *et al.*, 2015). Genotypic classes were redefined as described by Silva-Junior *et al.* (2015) and informative SNPs (unique map position on v.2 reference genome assembly, minor allele frequency (MAF) > 0.02 and genotyped in at least 90% of individuals) were extracted using the SNP & Variation Suite™ v.8.x (SVS8; Golden Helix Inc., Bozeman, MT, USA). Samples were also interrogated to ensure that at least 90% of informative markers were successfully genotyped in all individuals. Identity by descent analysis in SVS8 (Identity by Descent Estimation, SNP & Variation Suite Manual v.8.x; Golden Helix) was used to confirm half-sib relationships and to remove full-sib individuals from over-represented families. Only one such family from KZN was identified with nine putative full-sibling individuals, of which only one was retained for subsequent analysis (results not shown).

Population differentiation patterns were investigated and compared among breeding populations, and between breeding populations, wild *E. grandis* (including 362 individuals from three subpopulations; Mostert-O'Neill *et al.*, 2021) and other *Latoangulatae* species as published by Silva-Junior *et al.* (2015) using four approaches: principal component analysis (PCA) with normalization to each marker's standard deviation in SVS8; sparse nonnegative matrix factorization (sNMF) using the LEA R package (Frichot *et al.*, 2014; Frichot & François, 2015) – the values for K tested were $K = 2$ to $K = 10$ with five repetitions of each value and the minimum cross-entropy (CE) was determined for each value of K

and visualized; discriminant analysis of principal components (DAPC) using the ADEGENET R package (Jombart, 2008; Jombart *et al.*, 2010) with Bayesian information criterion (BIC) used to determine the most probable cluster number in the data set with $K = 1$ to $K = 15$ tested; and the extent of differentiation among breeding populations, and between breeding and wild *E. grandis* populations was quantified as F -statistics, F_{ST} , as described by Weir & Cockerham (1984), with 95% confidence intervals in SVS8.

Recent introgression, as a consequence of interspecific hybridization, can confound the detection of genomic segments under selection. To detect introgression in South African breeding programmes, population structure was investigated using PCA, sNMF and DAPC with the inclusion of published SNP genotypic data (Silva-Junior *et al.*, 2015) for other species within the section *Latoangulatae* (10 *E. saligna*, 19 *E. urophylla* and 16 *E. grandis* × *E. urophylla* (GU) hybrids). Suspected introgression was further tested by interrogating individual genotypes for the presence of genomic segments not originating from *E. grandis* by ancestry mapping using the Efficient Inference of Local Ancestry (EILA) R package (Yang *et al.*, 2013) as described by Mostert-O'Neill *et al.* (2021) with the breakpoint penalty $\lambda = 30$. Briefly, probable ancestry was calculated for each SNP using the same three reference populations as described by Mostert-O'Neill *et al.* (*E. grandis*, non-*E. grandis Latoangulatae* or *Maidenaria*-like). Next, the cumulative probability estimates of all SNPs on a chromosomal segment were used to assign each segment to one of the three ancestral populations. The penalty for allowing breakpoints (λ), where ancestry switches from one ancestral population to another along a chromosomal segment, was previously optimized based on ancestry mapping conducted in pure species and hybrid individuals for which the ancestry was known (Mostert-O'Neill *et al.*, 2021). This approach allowed for the identification of even small introgressed genomic segments, which could confound the detection of genomic regions selected during early domestication. For ancestry mapping, SNPs with zero missing data were used. Individuals with evidence of introgression (presence of genomic segments assigning to non-*E. grandis* ancestry) were excluded from subsequent analyses.

Differentiation between wild and breeding populations

To detect genomic regions selected during early domestication, a core breeding population, representing individuals that were

Table 1 Study populations and collection sites.

Breeding population	Number of families	Number of individuals	Site name	Latitude	Longitude	Elevation (m) [†]	MAP (mm)	MAT (°C)
TZA	284	287	Rooikoppies	−23.80	30.10	826	965	20
ZUL	43	285	Palm Ridge	−28.32	32.26	60	900	22
KZN	62	208	Siya Qubeka	−28.65	32.15	76	1196	21
		167	Nyalazi	−28.21	32.35	52	999	21
		185	Mtunzini	−29.03	31.66	84	1220	21

TZA, ZUL and KZN, South African *Eucalyptus grandis* populations.

MAT, mean annual temperature; MAP, mean annual precipitation.

[†]Elevation was determined based on GPS coordinates using the online resource, MAPS.ie (<https://www.maps.ie/coordinates.html>).

differentiated from the wild subpopulation, was identified using DAPC. This also allowed for the detection of individuals that probably shared a more recent ancestry with wild progenitors because of genetic infusions (introduction of wild, unimproved germplasm). Since there was no genetic evidence or historical records indicating that Mackay provenances, previously shown to be genetically distinct with evidence of natural interspecific introgression (see fig. 1 in Mostert-O'Neill *et al.*, 2021), were ever introduced to South Africa, the Mackay subpopulation was not included as wild progenitors. Based on the BIC results, DAPC was repeated for $K = 2$ to $K = 4$, and $K = 2$ was used to distinguish samples with recent genetic infusion from wild and breeding material. Group membership probabilities were used to detect breeding individuals that had more than 0.05 probabilistic assignment to the wild *E. grandis* cluster. To compare population structure resulting from the removal of introgressed and infused individuals, analyses using PCA, sNMF, DAPC and F_{ST} estimates were repeated on three data sets. The first was all *E. grandis* (using 23 661 informative SNPs), and the second was all *E. grandis* excluding introgressed (using 23 661 informative SNPs), in which introgressed breeding individuals were excluded. The third data set (using 21 991 informative SNPs) contained the North and South wild subpopulations (Mostert-O'Neill *et al.*, 2021) and core breeding *E. grandis* with recently infused breeding individuals excluded. The last data set was also used for outlier detection. Genetic diversity statistics, including average heterozygosity and inbreeding coefficients, were calculated for retained core breeding *E. grandis* using HIERFSTAT v.0.04-22 (Goudet, 2005).

Chloroplast haplotype diversity in wild and breeding populations

A subset of 361 individuals, representing 175 wild and 186 breeding families (representing introgressed, recently infused and core breeding individuals), were also genotyped using the Axiom™ Euc72K SNP chip through the genomics service provider, Thermo Fisher Scientific (Santa Clara, CA, USA), which allowed genotyping with chloroplast (cp) targeting assays. Of the 175 wild individuals, 14 were not previously genotyped by Mostert-O'Neill *et al.* (2021) using the EUChip60K SNP chip (Silva-Junior *et al.*, 2015) but were instead siblings of previously genotyped individuals. The SNP data were processed using the Axiom™ Analysis Suite (v.3.1 User Guide) and SVS8 to retain cp SNPs that were informative ($MAF \geq 0.05$) in at least 95% of the individuals. The informative cp SNP calls were concatenated for each individual to extract the cp haplotypes. Haplotype sequences (concatenated alleles) were exported as FASTA files using MEGA X (Kumar *et al.*, 2018) and haplotype networks were analysed following the guidelines of Toparslan *et al.* (2020) using the PEGAS R package (Paradis, 2010).

Identification and functional dissection of genomic outliers

Genome-wide patterns of LD, measured as the squared correlation (R^2) between allelic values at two loci, were determined in

SVS8 (LD Pairwise Analysis. SNP & Variation Suite Manual v.8.x. © 2017 Golden Helix) and visualized using LDHEATMAP (Shin *et al.*, 2006) and SVS8 LD plots for each of the 11 chromosomes, individually. To compare genome-wide patterns of heterozygosity between breeding and wild populations, Hardy–Weinberg equilibrium (HWE) signed R values, indicative of whether a marker is more homozygous (positive values) or heterozygous (negative values) in the population, were calculated in SVS8 (Signed HWE Correlation R . SNP & Variation Suite Manual v.8.x; Golden Helix) across the breeding and wild populations, and for each population separately.

Genomic loci differentiated between wild and breeding *E. grandis* were identified by comparing allele frequencies of 21 991 SNPs using two approaches: DAPC, to score SNP contributions in differentiating wild and breeding material into $K = 2$ clusters for each chromosome, separately (Jombart *et al.*, 2010); and marker-specific F_{ST} estimates as calculated using SVS8 based on the algorithm by Weir & Cockerham (1984). Loci were considered high-confidence outliers if they were within the 99th percentile of both outlier detection methods. A Wilcoxon signed-rank test was performed in R (v.3.5.1; R Development Core Team, 2018) to determine whether the mean of the outliers differed significantly from the mean of the remaining SNPs for DAPC SNP contribution scores, marker-specific F_{ST} and HWE signed R values because loci under directional selection are expected to be more homozygous. Outlier detection results were visualized using TABLEAU DESKTOP (Professional Edition ©2020). The breeding population consisted of 514 individuals (after removal of introgressed and recently infused individuals), and the wild progenitors were represented by 317 individuals from the Northern and Southern subpopulations.

Next, genes up- and downstream of high-confidence outliers were interrogated for GO functional enrichment against the full SNP-captured gene set as described by Pinar *et al.* (2019a) and Mostert-O'Neill *et al.* (2021). Two sets of genes, within 2 and 6 kb, were analysed based on the lower and upper estimates of LD decay as determined by Silva-Junior & Grattapaglia (2015), to account for large variations in genome-wide LD patterns in the breeding populations. Detailed interrogation of allele and genotype frequencies of outlier SNPs in LD with genes that showed functional enrichment for photosynthesis led us to question whether some SNP probes on the EUChip60K SNP chip had targeted organellar genome sequences in addition to nuclear genome targets. Basic Local Alignment Search Tool for nucleotides (BLASTN) analysis (Altschul *et al.*, 1990) was performed for all 57 567 SNP probe sequences that had unique mapping locations in the reference nuclear genome (Myburg *et al.*, 2014; Bartholomé *et al.*, 2015) against the *E. grandis* plastid and mitochondrial genome sequences (Pinar *et al.*, 2019b). Thereafter, outlier detection and GO enrichment analyses were repeated for 21 938 SNPs, excluding those with potential organellar genome targets. Population structure and differentiation analyses were also repeated with 53 organellar genome-targeting SNPs excluded, with no noticeable change to the results.

Loci putatively under selection were also detected by a multivariate approach using the PCADAPT R package (Luu *et al.*, 2017). This approach did not require predefined grouping of individuals, as in the case of DAPC SNP contribution scores and F_{ST} estimates. Instead, outliers were identified, for each chromosome separately, based on the Mahalanobis distance test statistic as differentiated from allele frequencies correlated with the first two principal components ($K = 2$) in a population structure PCA. Control for false discovery rate was done using the QVALUE R package (Dabney *et al.*, 2010) and loci with q -values < 0.05 were considered outliers. To determine the effect that different subpopulations had on the outliers detected, PCADAPT scans were repeated with sequential exclusion of each of the breeding and wild subpopulations. A PCADAPT scan was also repeated using wild subpopulations only, to detect outliers differentiated between the Northern and Southern wild subpopulations.

Results

SNP genotyping and population structure

Of the 64 639 SNPs assayed, 24 306 were informative (MAF > 0.02 , unique mapping position in the reference genome, called in at least 90% of individuals), and 2631 had zero missing data across the three breeding populations, the wild *E. grandis* and other *Latoangulatae* species. Most of the *E. grandis* breeding material appeared to be genetically distinct from wild *E. grandis* subpopulations in the sNMF analysis at $K = 3$ (Supporting Information Figs S1, S2). Some breeding individuals appeared to group away from the main *E. grandis* breeding cluster (Fig 1b,c) towards the *E. urophylla* and GU hybrid clusters in the population structure PCA plot (Fig. S1a) and DAPC analysis at $K = 7$, suggesting interspecific introgression. In particular, 163 of the 248 ZUL individuals had genomic assignment to *E. urophylla* and GU hybrids according to sNMF analyses from $K = 2$ (Fig. S1c). Ancestry mapping confirmed that these individuals had genomic segments assigned to non-*E. grandis* ancestry (segments assigned with non-*E. grandis* ancestry are visible as nonzero values in Table S1). Of the 1080 individuals from the three breeding programmes, only 685 had no introgressed genomic regions detected by ancestry mapping.

To introduce potentially adaptive genetic variation and reduce inbreeding, genetic infusions of wild, unimproved germplasm is common practice in forestry breeding. Since recent infusions can conceal genomic regions that are differentiated between breeding and wild populations in response to artificial selection, the next aim was to identify individuals that appeared to have recently introduced wild ancestry. Joint interrogation of DAPC and sNMF (Fig. S2) analyses (excluding introgressed individuals) indicated a separation between the majority of the breeding germplasm and wild progenitor populations (Fig. 1d) with a set of breeding individuals, predominately from TZA, appearing to share breeding and wild ancestry. Furthermore, of the three breeding populations, TZA appeared to be the least differentiated from the Southern wild subpopulation ($F_{ST} = 0.02$, Fig. S3). The putatively infused breeding individuals also grouped

between the main breeding cluster and the wild progenitor subpopulations in PCA plots (Figs 1c, S1a, S2a).

Since genetic infusions aim to introduce adaptive genetic variation into breeding populations, we also wanted to determine the origin of the infused germplasm. We were able to distinguish breeding samples that had wild ancestry derived from provenances in the Southern (light blue shade) vs the Northern (purple shade) subpopulations at $K = 3$ and $K = 4$ in the sNMF analysis (Fig. S2b) and confirmed these results by DAPC at $K = 3$ (Table S2b). Specifically, 98, 24 and two TZA, ZUL and KZN individuals, respectively, grouped with the Southern wild progenitor population cluster, while 16 TZA and ZUL individuals were assigned to the Northern wild subpopulation cluster. At $K = 2$ in DAPC analysis (excluding introgressed breeding and Mackay individuals), a separation between wild and the main breeding clusters was observed, with the suspected infused individuals either completely or partially assigned to the wild cluster (Fig. 1d; Table S2).

A total of 514 individuals (92 from TZA, 49 from ZUL and 373 from KZN) were retained as the core breeding germplasm (referred to as the *E. grandis* retained or core breeding population) for further analyses. We considered this a single group because there was no observable genetic differentiation among the three breeding populations once infused and introgressed individuals were removed (results not shown). To quantify the extent of genetic differentiation between the core breeding germplasm, the wild progenitor populations and the other species within section *Latoangulatae*, F_{ST} estimates were calculated for all of these comparisons (shown in Fig. S3). The core breeding population was as differentiated from the wild progenitors as the wild subpopulations were from each other (Fig. S3c). The breeding populations had negative average inbreeding coefficients (\hat{F}_{IS} ; higher observed heterozygosity than expected), which could be explained by novel genetic diversity from intraspecific (interprovenance) hybrids being advanced in the breeding programmes (Table S3).

Chloroplast haplotype diversity analysis was conducted to confirm that the breeding populations originate from a wide sampling of the natural populations, as suggested from breeding records. The analysis revealed 15 unique cp haplotypes based on 24 informative SNPs (Fig. S4). Of these, two (H8 and H14) were only detected in one core breeding family, each, and the only cp haplotype present in the Mackay wild subpopulation (H10) was absent from all analysed breeding germplasm. The presence of Northern and Southern wild subpopulation-derived haplotypes was observed in introgressed, infused and core breeding material.

Genomic regions differentiated between wild and core breeding *E. grandis*

Genomic regions under artificial selection were expected to be differentiated between the core breeding and wild progenitor populations, leading to changes in SNP marker heterozygosity and LD. The wild population generally had slightly more positive (homozygous) genome-wide HWE signed R values compared to

the breeding population (Fig. S5); however, outlier loci in the breeding population had significantly higher HWE signed R scores (i.e. were more often homozygous) compared to the rest of the SNPs as determined using a one-tailed Wilcoxon signed-rank test ($P = 6.22e^{-39}$, Table S4). High-confidence differentiated loci were distributed across the genome (Fig. 2a), although it should be noted that peaks of multiple differentiated loci appeared to overlap regions of increased LD in the breeding population on chromosomes 4 and 10 (Figs 2b, S6). Other large regions of increased LD in the core breeding population compared to LD in the wild were observed on chromosomes 2 and 11. Genome-wide patterns of LD varied noticeably in the breeding population among and within chromosomes (Figs S6, S7), with genome-wide average decay ($R^2 < 0.2$) at 1.8 kb.

Initial outlier detection in 21 991 SNPs revealed 85 loci that were in the 99th percentile of DAPC SNP contributions and marker-specific F_{ST} values. Photosynthesis-related GO terms were enriched among genes within 2 and 6 kb up- and downstream of outlier SNPs compared against the full SNP-captured gene space (Table S5). Detailed interrogation of these outlier SNPs revealed that heterozygous individuals were completely absent from the wild and breeding populations and that the SNP probes had high to complete sequence similarity with the plastid and/or mitochondrial genomes (Table S6). No GO enrichment was found for genes in LD with outliers detected after exclusion of the 53 informative markers that could target the organellar genomes in addition to the nuclear genome.

The large, differentiated region on chromosome 4 was also detected using PCADAPT. Outliers in this region were correlated with PC1, along which breeding and wild germplasm grouped separately (Fig. S2a). This 4 Mbp genomic region (from position 36 406 226 to 40 449 556) appears to be differentiated in all three breeding subpopulations as it was still detected when any of the three subpopulations were excluded from the PCADAPT scans. This region was not detected when the scan was conducted on the wild germplasm only (Fig. S8); therefore, it is not differentiated between the Northern and Southern wild subpopulations. The region contained 310 genes with no significant GO term enrichment. Other large outlier peaks correlated with PC2 were also identified. For example, the large peak on chromosome 2 appeared to be outliers differentiated in ZUL and KZN subpopulations, since this peak was not observed when either of these subpopulations was excluded from the PCADAPT scan. Also, the PC2-correlated peak on chromosome 10 appeared to be differentiated in KZN, specifically, as this peak was not observed when this subpopulation was excluded (Fig. S8).

Discussion

The aim of this study was to investigate the genomic consequences of artificial selection of exotic *E. grandis* populations that have been cultivated and bred *ex situ* for over 100 yr, representing a woody perennial in the early stages of domestication. This is, to our knowledge, the first study of plantation forestry domestication looking beyond five generations of selective breeding. Although the SNP markers used in our study

were sufficient for population structure and differentiation analyses, denser genome-wide genotyping, such as that achieved by whole-genome resequencing, will have to be performed to conclusively detect and discern signatures of selection. Still, our genome-wide investigation suggests that selection footprints would be discernible at this stage of the domestication process.

Domestication studies in other crops and in fruit trees, in particular, suggest that intra- and interspecific hybridization have contributed important genetic variation to cultivated populations (He *et al.*, 2011; Myles *et al.*, 2011; Cornille *et al.*, 2012; Wu *et al.*, 2014). Congruent with this, we found evidence of introgression from unintended hybridization, particularly in the ZUL breeding population. Since the 1980s, *E. grandis* plantations around the world have been challenged by fungal pathogens including *Chrysosporthe austroafricana* and *Coniothyrium* cankers (Wingfield *et al.*, 2008). This has led to widespread breeding and deployment of *E. grandis* × *E. urophylla* (GU) hybrids, which harnessed disease tolerance from *E. urophylla* while maintaining the favourable growth characteristics of *E. grandis* (see Potts & Dungey, 2004, for a comprehensive review of eucalypt hybrid breeding). The ZUL population was specifically bred in a subtropical region where biotic stress caused by these pathogens probably resulted in the selection of *E. grandis* × GU cryptic hybrids. Consequently, this breeding population is now enriched with introgressed genotypes from *E. urophylla*. Maintaining pure *E. grandis* breeding populations may become even harder as more pests and pathogens begin to thrive in subtropical zones, giving cryptic *E. grandis*-hybrids an adaptive advantage over pure species genotypes.

Individuals that appeared to be hybrids based on PCA, DAPC and sNMF analyses had extensive non-*E. grandis* genomic segments detected by ancestry mapping. Some individuals excluded as potentially introgressed had only small genomic segments assigned as non-*E. grandis* in origin and grouped within the core *E. grandis* PCA and DAPC clusters. The small non-*E. grandis* genomic segments in these individuals could have originated from interspecific hybridization, or be the result of incomplete lineage sorting. Extensive gene sequence data would be required to differentiate between these possible sources (Joly *et al.*, 2009; Meng & Kubatko, 2009; Yu *et al.*, 2013). Even where whole-genome sequence data are available, distinguishing between incomplete lineage sorting and hybridization can be problematic in closely related taxa (e.g. Meleshko *et al.*, 2021), and therefore is beyond the scope of this study. Another consideration is that the SNP chip used in the study, being a multispecies array, was enriched for SNP markers shared by two or more related species (Silva-Junior *et al.*, 2015). Even though SNP allele frequencies can differ very much between species, the preferential inclusion of such shared SNPs may have contributed to background levels of shared polymorphism. However, these are unlikely to account for the large genomic segments identified as non-*E. grandis* (i.e. hybrid in origin).

Although population differentiation patterns related to potential genetic infusions could also be explained by incomplete lineage sorting, recent genetic infusions of unimproved wild

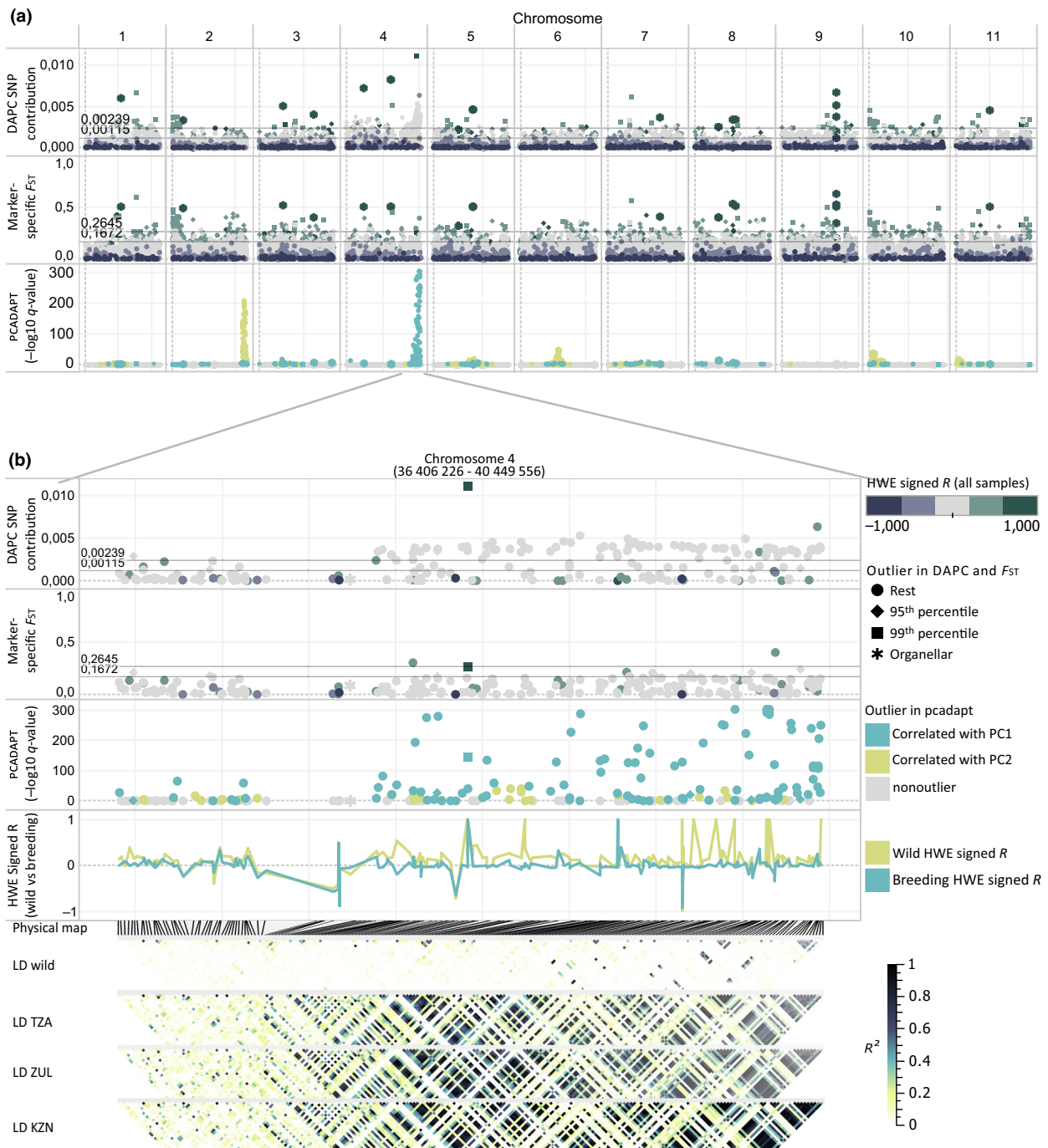


Fig. 2 Genomic regions differentiated between the core breeding and wild populations. (a) Discriminant analysis of principal components (DAPC) single nucleotide polymorphism (SNP) contributions, indicative of a marker's informativeness in separating breeding and wild samples into $K = 2$ clusters (Jombart *et al.*, 2010), marker-specific F_{ST} values as calculated for breeding (excluding introgressed and infused individuals) vs wild progenitors (Northern and Southern subpopulations), are given for each of the 21 991 SNPs with genomic positions given on the x-axis. In each panel, the 95th and 99th percentile values for outlier detection excluding SNPs with organellar genome targets) for each of the outlier detection methods are indicated as horizontal lines. Markers identified as differentiated in the 95th and 99th percentile in both analyses are indicated as squares and diamonds, respectively, and markers that had potential organellar genome targets are indicated as asterisks (these are included for illustration purposes only and were not considered for population structure and functional enrichment analysis). The colour scale is based on the Hardy–Weinberg equilibrium (HWE) signed R values of each SNP, indicative of whether a marker is more homozygous (green) or heterozygous (blue) across the breeding and wild populations. The third panel provides PCADAPT $-\log_{10} q$ -values for 21 991 SNPs, detected per chromosome. Outliers correlated with PC1 and PC2 are indicated in turquoise and yellow, respectively. (b) The same DAPC SNP loadings and F_{ST} estimates and PCADAPT outliers as shown in (a) for the outlier region on chromosome 4 (position 36 406 226 to 40 449 556). The fourth panel shows HWE signed R values for each marker as calculated in the wild (yellow) and breeding (turquoise) populations to illustrate changes in marker-specific heterozygosity. Beneath this plot is a physical map of all SNPs and linkage disequilibrium (LD) calculated as the squared correlation (R^2) between alleles at two loci in the wild progenitors and three breeding populations, TZA, ZUL and KZN.

material from Coffs Harbour and Atherton provenances in the Southern and Northern wild subpopulations, respectively, were expected based on breeding records. For example, wild germplasm and unrelated families from other breeding trials were known to have been introduced into the TZA breeding programme, formerly managed by the South African Council for Scientific and Industrial Research (CSIR; Verryin *et al.*, 2009), and germplasm from the Northern wild subpopulation is known to have been introduced into the KZN programme in the 1990s. We observed evidence supporting the presence of genetic infusions (Table S2b; Fig. S2b) in TZA, particularly from the Southern wild subpopulation. The TZA population was selected for solid wood products and bred for temperate climates (Table 1), while ZUL and KZN were, in recent years, mostly bred for pulp-derived products in subtropical and warm- to cool-temperate climates, respectively. This supports the preferential retention of genotypes originating from the temperate South in TZA, despite breeding records and the cp SNP haplotype network analysis also pointing to recent introductions from Atherton in the Northern wild subpopulation; this is congruent with the notion proposed by Bennett (2011) that one of the first steps in domestication involves capturing existing adaptive genetic variation that matches seed source and *ex situ* climates. Interspecific hybridization and continued infusions from wild populations are prevalent in domestication; however, since these events occurred very recently in our study populations, the retention of introgressed and infused individuals could confound and mask genomic signatures resulting from artificial selection over a period of 100 yr. Therefore, we excluded potentially introgressed and recently infused genotypes from subsequent analyses, although these genotypes represent important genetic variation for future selective breeding.

Cocultivation of genotypes originating from different provenances would have resulted in intraspecific (interprovenance) hybrids in subsequent generations. We saw evidence of this as genome-wide heterozygosity was higher in the core South African breeding germplasm compared to the wild (Table S3; Figs S5, S6), possibly counteracting genetic bottlenecks that could have occurred at the start of domestication relative to each of the wild source populations. Similarly, Jones *et al.* (2006) observed increased heterozygosity in first-generation selections of *E. globulus*, suggesting that intraspecific hybrids were advanced early in domestication. Still, hybridization among individuals from different provenances alone does not explain the clear genetic differentiation of breeding populations from the wild progenitors (Figs 1, S2).

Genetic drift and selection could have contributed to the differentiation observed between the breeding and wild populations. Empirical support to differentiate the contributions of these evolutionary forces could come from analysing older breeding material from over 50 yr ago. Sadly, no such material remains in current breeding archives. Assessing the effect and impact of genetic drift might also be difficult considering the repeated introductions of genotypes from diverse, wild populations as is reported to have occurred since the 1960s (Poynton, 1979). When looking at genomic changes over 10 generations of

adaptive domestication in maize, Wisser *et al.* (2019) described two phases: early fixation of a small number of large-effect variants followed by gradual allele frequency changes at many loci due to selection of quantitative traits. The latter phase could explain shifts in allele frequencies that resulted in the genome-wide genetic differentiation between breeding and wild *E. grandis* material, which probably represent the genomic changes underlying rapid genetic gains achieved early on for highly complex traits (Verryin, 2002; Verryin *et al.*, 2009). Even when selection pressure is high and large genetic gains are observed phenotypically, selection on complex traits typically does not translate into classic selection signatures, known as hard and soft sweeps (Cutter & Payseur, 2013).

Selection sweeps arise in the genome when a novel mutation (Smith & Haigh, 1974) or standing genetic variation (Innan & Kim, 2004; Hermisson & Pennings, 2005) confers a strong selective advantage and becomes fixed in a population (Pritchard *et al.*, 2010). They appear as stretches of elevated homozygosity, increased differentiation (e.g. higher localized F_{ST} estimates) and increased LD (Cutter & Payseur, 2013), since genetic variants surrounding the locus under selection also become fixed due to genetic hitchhiking (Smith & Haigh, 1974). We uncovered one such region on chromosome 4, with several SNPs differentiated between breeding and wild populations, and elevated LD in all three breeding programmes (Figs 2, S6, S8). We postulate that this region contains variants that were under either negative or neutral selection in the wild but were preferentially (positively) advanced in South Africa, thereby reducing the genetic variation and increasing LD surrounding the selected locus in breeding populations; that is, this may represent an early soft sweep. Because domestication had occurred for approximately five generations of formal breeding preceded by up to as many generations of informal selections, allowing only a limited number of recombination events, this genomic region remains large, limiting our ability to identify candidate genes and biological processes.

Enrichment for photosynthesis-related GO terms observed in an initial screen of genes in LD with outlier SNPs suggested that several of the EUChip60K SNP probe sets must have additional target sequences in the plastid and/or mitochondrial genomes (Table S6). Gene transfers among different genomes in a cell is well documented for *E. grandis* (Pinar *et al.*, 2019b). Even though these SNP probes could detect nuclear and organellar sequences, for 30 SNPs, which were polymorphic in the breeding and wild populations, we observed a complete deficiency of heterozygotes in the wild and breeding populations (Table S7). It is likely that for these SNPs, the genotypes were dominated by organellar genome template, which is in vast excess in genomic DNA samples. It is possible that SNPs targeting organellar and nuclear genome sequences were detected as outliers as they would reflect founder effects if only some provenances were introduced to South Africa. The source provenances that constituted the original seed imports from the first half of the 20th century remain unknown. Since imports and subsequent exchange of genetic material occurred mostly via seed, maternally inherited

cp SNPs were used to inform which provenances were introduced to South Africa. The cp haplotype network (Fig. S4) supported that some wild haplotypes (H4, H10, H12 and H15) were not detected in breeding populations while two haplotypes present in the breeding germplasm were not present in the wild material, possibly representing unsampled wild provenances. Although we excluded all of these putative cp-targeting SNPs from further analyses, we chose to include their outlier detection values in Figs 2 and S6 for illustration purposes.

To conclude, by interrogating genome-wide SNP allele frequencies in *E. grandis* breeding and wild populations, we have uncovered genomic evidence of evolutionary processes similar to those that have shaped the genomes of other domesticates. In addition to the genome-wide genetic differentiation between breeding and wild populations, probably caused by early artificial selection of polygenic traits, we observed localized allele frequency shifts with increased differentiation and LD. A lack of recombination events required to uncouple loci under selection from the neutral genomic background meant that these regions were still too broad for candidate gene identification. Although we used SNPs to tag genomic regions under artificial selection, we know from published reports that the causative variants could have been single nucleotide, presence/absence and copy number variants, as well as other structural variants (see review by Olsen & Wendel, 2013). Additionally, the use of SNP arrays results in the exclusion of rare variants that may be more informative in terms of recent differentiation events (Dokan *et al.*, 2021). Therefore, our future aim is to use sequenced-based genotyping to elucidate structural variants and haplotypes in *E. grandis* breeding and wild progenitor populations that may be associated with adaptation to *ex situ* environments and early domestication.

Acknowledgements









We thank D. Pinard (University of Pretoria) for assistance in BLASTn against organellar genomes. Also, Hans Merensky, Sappi Forests and Mondi South Africa for biological specimens used in this study. This work was funded in part by the Department of Science and Innovation and Technology Innovation Agency (DSI/TIA, Strategic Grant-*Eucalyptus* Genomics Platform), the Forestry Sector Innovation Fund (FSIF *Eucalyptus* Genome Diversity Atlas grant), National Research Foundation (NRF) of South Africa – Bioinformatics and Functional Genomics Programme (BFG Grant UID 97911), the Technology and Human Resources for Industry Programme (THRIP Grant UID 96413), and by the Forest Molecular Genetics (FMG) Industry Consortium at the University of Pretoria. MMM-O acknowledges PhD scholarship support from the NRF.

Author contributions

MMM-O and AAM developed the idea. MMM-O, SMR, MMM, GvdB, SDV, JJA, JOB and AAM contributed to the design of the study. MMM-O, SMR, MMM, GvdB, SDV and

AAM performed sample collection and generated the data. HT analysed chloroplast marker data and generated chloroplast haplotype networks. MMM-O analysed the data and wrote the first draft of the manuscript. All authors read, edited and approved the final manuscript.

ORCID

Juan J. Acosta  <https://orcid.org/0000-0002-9429-5166>
 Justin O. Borevitz  <https://orcid.org/0000-0001-8408-3699>
 Marja M. Mostert-O'Neill  <https://orcid.org/0000-0002-6318-3508>
 Makobatjatji M. Mphahlele  <https://orcid.org/0000-0003-2563-7764>
 Alexander A. Myburg  <https://orcid.org/0000-0003-0644-5003>
 S. Melissa Reynolds  <https://orcid.org/0000-0002-8995-5836>
 Hannah Tate  <https://orcid.org/0000-0002-8698-5183>
 Steve D. Verry  <https://orcid.org/0000-0001-8026-0707>

Data availability

The genomic data generated and analysed in this study are available online via the Dryad archives under accession <https://doi.org/10.5061/dryad.h18931zj6>.

References

- Acosta JJ, Fahrenkrog AM, Neves LG, Resende MFR, Dervinis C, Davis JM, Holliday JA, Kirst M. 2019. Exome resequencing reveals evolutionary history, genomic diversity, and targets of selection in the conifers *Pinus taeda* and *Pinus elliottii*. *Genome Biology and Evolution* 11: 508–520.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
- Alves AA, Rosado CCG, Faria DA, Guimarães LMS, Lau D, Brommonschenkel SH, Grattapaglia D, Alfenas AC. 2012. Genetic mapping provides evidence for the role of additive and non-additive QTLs in the response of inter-specific hybrids of *Eucalyptus* to *Puccinia psidii* rust infection. *Euphytica* 183: 27–38.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.* 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.
- Bartholomé J, Mandrou E, Mabilia A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion JM. 2015. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* 206: 1283–1296.
- Baute GJ, Kane NC, Grassa CJ, Lai Z, Rieseberg LH. 2015. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytologist* 206: 830–838.
- Bennett BM. 2011. Naturalising Australian trees in South Africa: climate, exotics and experimentation. *Journal of Southern African Studies* 37: 265–280.
- Bouffier L, Raffin A, Kremer A. 2008. Evolution of genetic variation for selected traits in successive breeding populations of maritime pine. *Heredity* 101: 156–165.
- Boulay A, Tacconi L, Kanowski P. 2012. Drivers of adoption of eucalypt tree farming by small holders in Thailand. *Agroforestry Systems* 84: 179–189.
- Chaix G, Gerber S, Razafimaharo V, Vigneron P, Verhaegen D, Hamon S. 2003. Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. *Theoretical and Applied Genetics* 107: 705–712.

- Collevatti RG, Novaes E, Silva-Junior OB, Vieira LD, Lima-Ribeiro MS, Grattapaglia D. 2019. A genome-wide scan shows evidence for local adaptation in a widespread keystone neotropical forest tree. *Heredity* 123: 117–137.
- Cornille A, Giraud T, Smulders MJM, Roldán-Ruiz I, Gladieux P. 2014. The domestication and evolutionary ecology of apples. *Trends in Genetics* 30: 57–65.
- Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel J, Olonova M, Feugey L *et al.* 2012. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* 8: e1002703.
- Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* 14: 262–274.
- Dabney A, Storey JD, Warnes G. 2010. *QVALUE: Q-value estimation for false discovery rate control; R package version 1.24.20*. [WWW document] URL <https://github.com/StoreyLab/qvalue> [accessed 11 February 2020].
- De La Torre AR, Wang T, Jaquish B, Aitken SN. 2014. Adaptation and exogenous selection in a *Picea glauca* × *Picea engelmannii* hybrid zone: implications for forest management under climate change. *New Phytologist* 201: 687–699.
- Denison N, Kietzka J. 1993. The use and importance of hybrid intensive forestry in South Africa. *South African Forestry Journal* 165: 55–60.
- Dokan K, Kawamura S, Teshima KM. 2021. Effects of single nucleotide polymorphism ascertainment on population structure inferences. *G3* 11: jkab128.
- Dos Santos PET, Geraldi IO, Garcia JN. 2004. Estimates of genetic parameters of wood traits for sawn timber production in *Eucalyptus grandis*. *Genetics and Molecular Biology* 27: 567–573.
- Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G. 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics* 46: 1089–1096.
- Frichot E, François O. 2015. LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution* 6: 925–929.
- Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196: 973–983.
- Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* 5: 184–186.
- Grattapaglia D, Bertolucci F, Sederoff R. 1995. Genetic mapping of QTLs controlling vegetative propagation in *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. *Theoretical and Applied Genetics* 90: 933–947.
- Grattapaglia D, Bertolucci FLG, Penchel R, Sederoff RR. 1996. Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. *Genetics* 144: 1205–1214.
- He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu C-I, Shi S. 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genetics* 7: e1002100.
- Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335–2352.
- Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaepler SM *et al.* 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* 44: 808–811.
- Hunde T, Duguma D, Gizachew B, Mamushet D, Teketay D. 2003. Growth and form of *Eucalyptus grandis* provenances at Wondo genet, Southern Ethiopia. *Australian Forestry* 66: 170–175.
- Huoran W, Yongqi Z, Daoqun Z, Xiuwu C. 1992. Provenance variation in growth and wood properties of *Eucalyptus grandis* in China. In: Brown AG, ed. *Australian tree species research in China. Proceedings of an international workshop, Zhangzhou, P.R. China, 2–5 November 1992*. Canberra, Australian Centre for International Agricultural Research, Proceedings No. 48, 105–107.
- Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences, USA* 101: 10667–10672.
- Isik F, Kumar S, Martínez-García PJ, Iwata H, Yamamoto T. 2015. Chapter three - acceleration of forest and fruit tree domestication by genomic selection. In: Plomion C, Adam-Blondon A-F, eds. *Advances in botanical research, vol. 74*. Oxford, UK: Elsevier, 93–124.
- Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist* 174: E54–E70.
- Jombart T. 2008. ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405.
- Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94.
- Jones TH, Steane DA, Jones RC, Pilbeam D, Vaillancourt RE, Potts BM. 2006. Effects of domestication on genetic diversity in *Eucalyptus globulus*. *Forest Ecology and Management* 234: 78–84.
- Kullan ARK, van Dyk MM, Hefer CA, Jones N, Kanzler A, Myburg AA. 2012. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* 13: 1–12.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* 35: 1547–1549.
- Luo J, Zhou G, Wu B, Chen D, Cao J, Lu W, Pegg RE, Arnold RJ. 2010. Genetic variation and age-age correlations of *Eucalyptus grandis* at Dongmen Forest farm in southern China. *Australian Forestry* 73: 67–80.
- Luu K, Bazin E, Blum MGB. 2017. PCADAPT: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* 17: 67–77.
- Mangwanda R, Myburg AA, Naidoo S. 2015. Transcriptome and hormone profiling reveals *Eucalyptus grandis* defence responses against *Chrysoperthe austroafricana*. *BMC Genomics* 16: 319.
- Marco MA. 1991. Seed source trials of *Eucalyptus grandis* in Argentina. *Investigación Agraria. Sistemas y Recursos Forestales* 1: 111–119.
- Marques C, Brondani R, Grattapaglia D, Sederoff R. 2002. Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species. *Theoretical and Applied Genetics* 105: 474–478.
- Meleshko O, Martin MD, Korneliusen TS, Schröck C, Lamkowski P, Schmutz J, Healey A, Piatkowski BT, Shaw AJ, Weston DJ *et al.* 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Molecular Biology and Evolution* 38: 2750–2766.
- Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75: 35–45.
- Meskimen G. 1983. Realized gain from breeding *Eucalyptus grandis* in Florida. In: Standiford RB, Ledig FT, eds. technical coordinators. *Proceedings of a workshop on eucalyptus in California, June 14–16, 1983, Sacramento, California. Gen. Tech. Rep. PSW 69, vol. 69*. Berkeley, CA, USA: Pacific southwest Forest and range Experiment Station, Forest Service, US Department of Agriculture, 121–128.
- Mhoswa L, O'Neill MM, Mphahlele MM, Oates CN, Payn KG, Slippers B, Myburg AA, Naidoo S. 2020. A genome-wide association study for resistance to the insect pest *Leptocybe invasa* in *Eucalyptus grandis* reveals genomic regions and positional candidate defence genes. *Plant and Cell Physiology* 61: 1285–1296.
- Mostert-O'Neill MM, Reynolds SM, Acosta JJ, Lee DJ, Borevitz JO, Myburg AA. 2021. Genomic evidence of introgression and adaptation in a model subtropical tree species, *Eucalyptus grandis*. *Molecular Ecology* 30: 625–638.
- Mphahlele MM, Isik F, Mostert-O'Neill MM, Reynolds SM, Hodge GR, Myburg AA. 2020. Expected benefits of genomic selection for growth and wood quality traits in *Eucalyptus grandis*. *Tree Genetics & Genomes* 16: 49.
- 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.
- Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D *et al.* 2011. Genetic structure and

- domestication history of the grape. *Proceedings of the National Academy of Sciences, USA* 108: 3530–3535.
- Oates CN, Külheim C, Myburg AA, Slippers B, Naidoo S. 2015. The transcriptome and terpene profile of *Eucalyptus grandis* reveals mechanisms of defense against the insect pest, *Leptocybe invasa*. *Plant and Cell Physiology* 56: 1418–1428.
- Olsen KM, Wendel JF. 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annual Review of Plant Biology* 64: 47–70.
- Paradis E. 2010. PEGAS: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26: 419–420.
- Pinard D, Fierro AC, Marchal K, Myburg AA, Mizrachi E. 2019a. Organellar carbon metabolism is coordinated with distinct developmental phases of secondary xylem. *New Phytologist* 222: 1832–1845.
- Pinard D, Myburg AA, Mizrachi E. 2019b. The plastid and mitochondrial genomes of *Eucalyptus grandis*. *BMC Genomics* 20: 132.
- Potts BM, Dungey HS. 2004. Interspecific hybridization of *eucalyptus*: key issues for breeders and geneticists. *New Forests* 27: 115–138.
- Poynton RJ. 1979. *Tree planting in southern Africa: the eucalypts*. Pretoria, South Africa: Department of Forestry.
- Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* 20: R208–R215.
- Prunier J, Laroche J, Beaulieu J, Bousquet J. 2011. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology* 20: 1702–1716.
- Purugganan MD, Fuller DQ. 2009. The nature of selection during plant domestication. *Nature* 457: 843–848.
- R Development Core Team. 2018. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rocha RB, Barros EG, Cruz CD, Rosado AM, EFD A. 2007. Mapping of QTLs related with wood quality and developmental characteristics in hybrids (*Eucalyptus grandis* × *Eucalyptus urophylla*). *Revista Árvore* 31: 13–24.
- Rockwood DL, Meskimen GF. 1991. Comparison of *Eucalyptus grandis* provenances and seed orchards in a frost frequent environment. *South African Forestry Journal* 159: 51–59.
- Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences, USA* 104: 8641–8648.
- Santos SAO, Vilela C, Domingues RMA, Oliveira CSD, Villaverde JJ, Freire CSR, Neto CP, Silvestre AJD. 2017. Secondary metabolites from *Eucalyptus grandis* wood cultivated in Portugal, Brazil and South Africa. *Industrial Crops and Products* 95: 357–364.
- Shin J-H, Blay S, McNeney B, Graham J. 2006. LDHEATMAP: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software* 16: 1–10.
- Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *eucalyptus* tree genomes across 12 species. *New Phytologist* 206: 1527–1540.
- Silva-Junior OB, Grattapaglia D. 2015. 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*. *New Phytologist* 208: 830–845.
- Skroppa T, Steffenrem A. 2016. Selection in a provenance trial of Norway spruce (*Picea abies* L. karst) produced a land race with desirable properties. *Scandinavian Journal of Forest Research* 31: 439–449.
- Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* 23: 23–35.
- Thomas B, Raj MC, Joy J, Moores A, Drisko GL, Cm S. 2018. Nanocellulose, a versatile green platform: from biosources to materials and their applications. *Chemical Reviews* 118: 11575–11625.
- Toparlan E, Karabag K, Bilge U. 2020. A workflow with R: phylogenetic analyses and visualizations using mitochondrial cytochrome *b* gene sequences. *PLoS ONE* 15: e0243927.
- Tuskan G. 2007. Bioenergy, genomics, and accelerated domestication: a US example. *FAO, Papers and Presentations from The Role of Agricultural Biotechnologies for Production of Bioenergy in Developing Countries*. [WWW document] URL <http://www.fao.org/biotech/seminaroct2007.htm> [accessed 21 August 2020].
- Van Wyk G, Roeder KR. 1978. The status of tree breeding in South Africa. *South African Forestry Journal* 107: 54–59.
- Varghese M, Kamalakannan R, Harwood C, Lindgren D, McDonald M. 2009. Changes in growth performance and fecundity of *Eucalyptus camaldulensis* and *E. tereticornis* during domestication in southern India. *Tree Genetics & Genomes* 5: 629–640.
- Verryn SD. 2002. Harvesting genetics for productive plantations. *Southern African Forestry Journal* 195: 83–87.
- Verryn SD, Snedden CL, Eatwell KA. 2009. A comparison of deterministically predicted genetic gains with those realised in a south African *Eucalyptus grandis* breeding program. *Southern Forests: A Journal of Forest Science* 71: 141–146.
- Vining KJ, Romanel E, Jones RC, Klocko A, Alves-Ferreira M, Hefer CA, Amarasinghe V, Dharmawardhana P, Naithani S, Ranik M *et al.* 2015. The floral transcriptome of *Eucalyptus grandis*. *New Phytologist* 206: 1406–1422.
- Wang J, Street NR, Park E-J, Liu J, Ingvarsson PK. 2020. Evidence for widespread selection in shaping the genomic landscape during speciation of *Populus*. *Molecular Ecology* 29: 1120–1136.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.
- Wingfield MJ, Slippers B, Hurley BP, Coutinho TA, Wingfield BD, Roux J. 2008. Eucalypt pests and diseases: growing threats to plantation productivity. *Southern Forests: A Journal of Forest Science* 70: 139–144.
- Wisser RJ, Fang Z, Holland JB, Teixeira JEC, Dougherty J, Weldekidan T, de Leon N, Flint-Garcia S, Lauter N, Murray SC *et al.* 2019. The genomic basis for short-term evolution of environmental adaptation in maize. *Genetics* 213: 1479–1494.
- Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J *et al.* 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology* 32: 656–662.
- Yang JJ, Li J, Buu A, Williams LK. 2013. Efficient inference of local ancestry. *Bioinformatics* 29: 2750–2756.
- Yu Y, Barnett RM, Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology* 62: 738–751.
- Zobel B, Talbert J. 1984. *Applied forest tree improvement*. New York, NY, USA: John Wiley & Sons.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Population structure in relation to wild *Eucalyptus grandis* and other species in section *Latoangulatae* based on principal component analysis, discriminant analysis of principal components and sparse nonnegative matrix factorization.

Fig. S2 Breeding *Eucalyptus grandis* population structure for all breeding samples, those excluding introgressed, and those excluding infused individuals in relation to the wild progenitor populations based on principal component analysis, sparse nonnegative matrix factorization and discriminant analysis of principal components analyses.

Fig. S3 Population differentiation F_{ST} estimates among breeding *Eucalyptus grandis*, wild *E. grandis* and other species in section *Latoangulatae*.

Fig. S4 Chloroplast (cp) haplotype network based on 24 cp single nucleotide polymorphisms.

Fig. S5 Marker-specific Hardy–Weinberg equilibrium signed R values of wild vs breeding populations.

Fig. S6 Genomic outliers and linkage disequilibrium plots per chromosome.

Fig. S7 Breeding population linkage disequilibrium decay over genomic distance in kb.

Fig. S8 Outlier detection by PCADAPT scan.

Table S1 Ancestry assignment of chromosomal segments.

Table S2 Cluster assignment of samples using discriminant analysis of principal components to identify genetically infused breeding individuals.

Table S3 Summary statistics of genetic diversity using HIERFSTAT v.0.04-22.

Table S4 Wilcoxon signed rank test P -values supporting the alternative hypothesis that the mean of the outliers was greater than the mean of the rest of the single nucleotide polymorphisms.

Table S5 Gene Ontology enrichment analysis for genes in linkage disequilibrium with outlier single nucleotide polymorphisms (SNPs) before excluding organellar-targeting SNPs.

Table S6 BLASTN against the organellar genomes.

Table S7 Marker statistics of single nucleotide polymorphisms with multigenome targets.

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.