https://doi.org/10.1186/s12864-024-10810-4

BMC Genomics

RESEARCH

Open Access

# A telomere-to-telomere *Eucalyptus regnans* genome: unveiling haplotype variance in structure and genes within one of the world's tallest trees

Scott Ferguson[1*], Yoav D Bar-Ness[2], Justin Borevitz[1] and Ashley Jones[1*]

## Abstract

**Background** *Eucalyptus regnans* (Mountain Ash) is an Australian native giant tree species which form forests that are among the highest known carbon-dense biomasses in the world. To enhance genomic studies in this ecologically important species, we assembled a high-quality, mostly telomere-to-telomere complete, chromosome-level, haplotype-resolved reference genome. We sampled a single tree, the Centurion, which is currently a contender for the world's tallest flowering plant.

**Results** Using long-read sequencing data (PacBio HiFi, Oxford Nanopore ultra-long reads) and chromosome conformation capture data (Hi-C), we assembled the most contiguous and complete *Eucalyptus* reference genome to date. For each haplotype, we observed contig N50s exceeding 36 Mbp, scaffold N50s exceeding 43 Mbp, and genome BUSCO completeness exceeding 99%. The assembled genome revealed extensive structural variations between the two haplotypes, consisting mostly of insertions, deletions, duplications and translocations. Analysis of gene content revealed haplotype-specific genes, which were enriched in functional categories related to transcription, energy production and conservation. Additionally, many genes reside within structurally rearranged regions, particularly duplications, suggesting that haplotype-specific variation may contribute to environmental adaptation in the species.

**Conclusions** Our study provides a foundation for future research into *E. regnans* environmental adaptation, and the high-quality genome will be a powerful resource for conservation of carbon-dense giant tree forests.

**Keywords** *Eucalyptus regnans*, Centurion, Diploid genome assembly, Haplotype-resolved, Telomere-to-telomere

*Correspondence:
Scott Ferguson
scott.ferguson.papers@gmail.com
Ashley Jones
ashley.jones@anu.edu.au
[1]Research School of Biology, Australian National University, Canberra, ACT, Australia
[2]Giant Tree Expeditions, Hobart, TAS, Australia

## Background

*Eucalyptus* forests are widespread across Australia and extend north into tropical islands. They provide habitat to a rich biodiversity of marsupials, birds and insects, being key foundation species in natural ecosystems [1]. *Eucalyptus* trees are highly diverse and adaptable, exhibiting resistance to extreme droughts, fires and floods. The *Eucalyptus* genus contains over 900 species that have variable genome sizes of approximately 400–700 Mbp [2], high heterozygosity [3] and high frequency of structural variants [4]. With different phenotypes and adaptive traits to varying environments, there is an increasing need for representative genomes.

*Eucalyptus regnans* (Mountain Ash, also known as Swamp Gum and Stringy Gum) is part of the diverged subgenera *Eucalyptus*, formerly known as *Monocalyptus* with several sections and 100 species, including the alpine specialist snow gum, *E. pauciflora*, that is facing dieback [5]. Representing a grove of giant trees in Tasmania, *E. regnans* forests are among the highest known carbon-dense biomasses in the world [6]. They annually sequester and store large amounts of carbon, with the wet temperate forests having the highest above and below ground carbon densities of 1,000 tC/h [7] to 1,312 tC/h [6]. This highlights their significance in mitigating climate change. However, *E. regnans* forests are under threat from climate change, particularly widespread bushfires that can occur in Australia [8, 9]. Increased logging and deforestation is also becoming a widespread concern. Therefore, there is an increasing need for conservation, management, and restoration of these forests.

Among these forests, the *E. regnans* tree known as "Centurion" (Fig. 1), is currently a pre-eminent candidate for the world's tallest known flowering plant [10]. Captivating researchers and enthusiasts alike, it has been measured at 99.6 m by an aerial laser scanning LIDAR forest inventory [11] and at 99.8 m by tape drop techniques [12]. In 2018, the tree was remeasured at 100.5 m using ground-based observations with a Laser Technologies TruPulse 360 [13]. Still alive and growing, despite being partially burnt in a bushfire, it is among the tallest known angiosperms [14, 15]. Currently, the tallest known trees

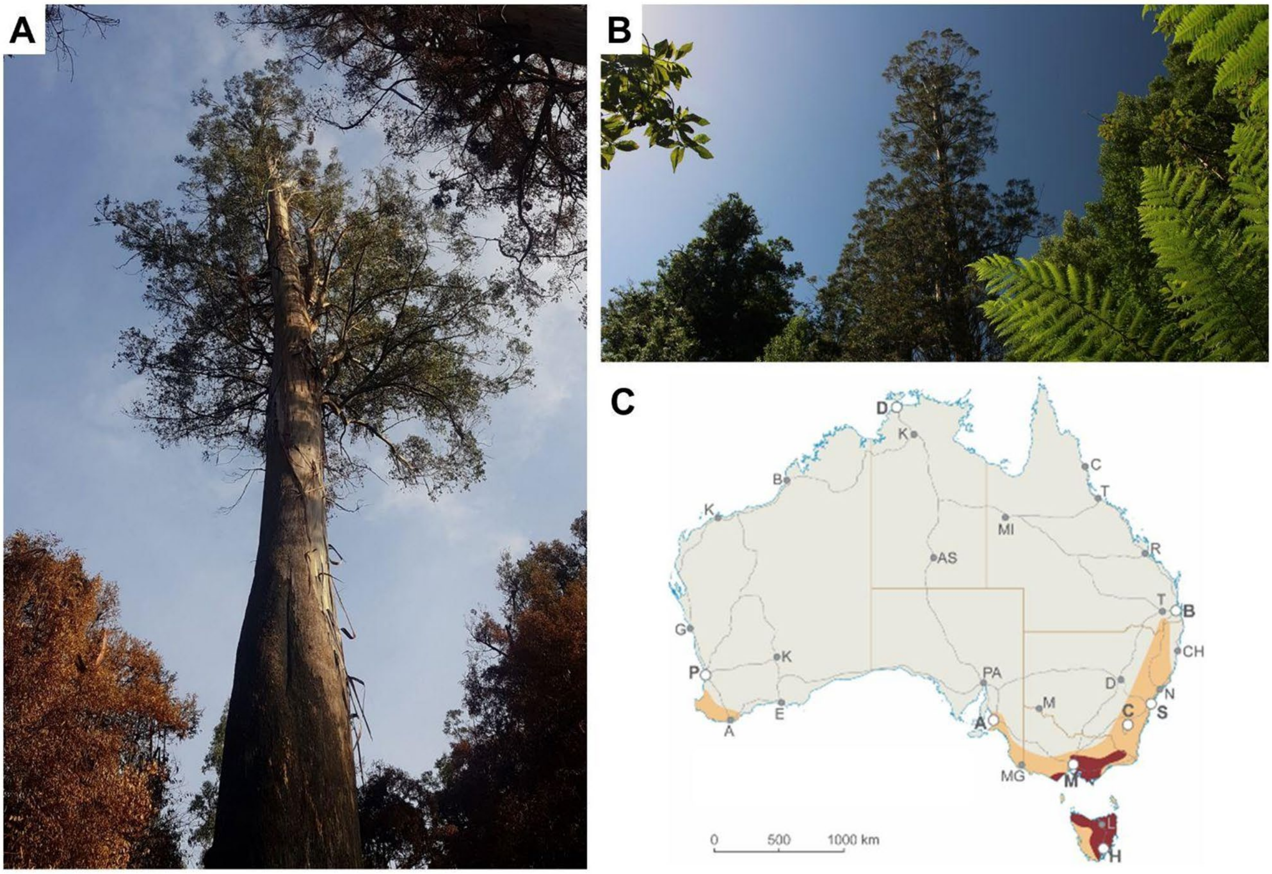


**Fig. 1** *Eucalyptus regnans* the Centurion, located in Tasmania, Australia. (**A**) Picture is of the Centurion after national bushfires in Australia. (**B**) The Centurion before the bushfires. Photographs by Yoav Daniel Bar-Ness (Giant Tree Expeditions). (**C**) Species occurrence distribution map of *E.regnans* in Australia. Red represents natural distribution, orange represents low frequency plantations. Alphabetical letters represent major cities in Australia, bold is capital cities. Key areas of natural distribution are Hobart, H, Launceston, L, and Melbourne, M. Map sourced from [21], being provided by the author Dean Nicolle

**Table 1** Sequencing data summary

| | | ONT | |
|---|---|---|---|
| | HiFi | Raw | Filtered |
| Size (Gbp) | 92.07 | 23.53 | 21.34 |
| Coverage | 167.39 | 42.78 | 38.81 |
| N50 (Kbp) | 15.80 | 46.73 | 48.35 |
| Number (1,000) | 6,482.95 | 718.32 | 467.74 |
| Longest (Kbp) | 47.93 | 219.14 | 219.14 |
| Shortest (bp) | 105 | 36 | 20,000 |

are the non-flowering *Sequoia sempervirens* (coast redwoods, such as the "Hyperion"), which can achieve over 112 m in height [16]. Further research into the world's tallest trees provides valuable opportunities to understand tree growth, carbon sequestration and wood production.

To enable further studies into *E. regnans* and carbon-dense giant trees, we assembled a haplotype-resolved chromosome-level genome of the Centurion, using Pacific Biosciences (PacBio) HiFi reads, Oxford Nanopore Technologies (ONT) ultra-long reads and Hi-C chromosome conformation capture, in the hybrid assembler Hifiasm ultra-long (UL) [17]. Long-read sequencing technologies now enable complete, telomere-to-telomere (T2T) assemblies of complex genomes, such as the human genome [18], kiwifruit [19] and maize genome [20]. Using this approach, we assemble the most complete *Eucalyptus* genome to date, being the first chromosome-level, haplotype phased, diploid assembly, approaching complete T2T quality. Our genome provides insights into the structural variation between haplotypes and enables further studies into genome evolution in *Eucalyptus*.

## Results

### Long-read native DNA sequencing

To assemble the genome of *E. regnans* the Centurion, we extracted high-molecular weight DNA from leaves for long-read sequencing with PacBio for HiFi reads and ONT for ultra-long reads≥40 kb. A portion of leaf tissue was crosslinked for Hi-C chromosome conformation capture followed by short-read sequencing with Illumina. Sequencing generated of 92.07 Gbp HiFi (N50 15.80 Kbp, ~176x coverage), 23.53 Gbp ONT (N50 46.73 Kbp, ~49x coverage), and 8.03 Gbp Hi-C (150 bp paired end) (Table 1; Fig. 2.A, 2.B). ONT reads were filtered, removing all reads<20 Kbp and <Q7, leaving 21.34 Gbp (N50 28.35 Gbp, ~41x coverage). Hi-C sequences were contained in 26.77 million read pairs.

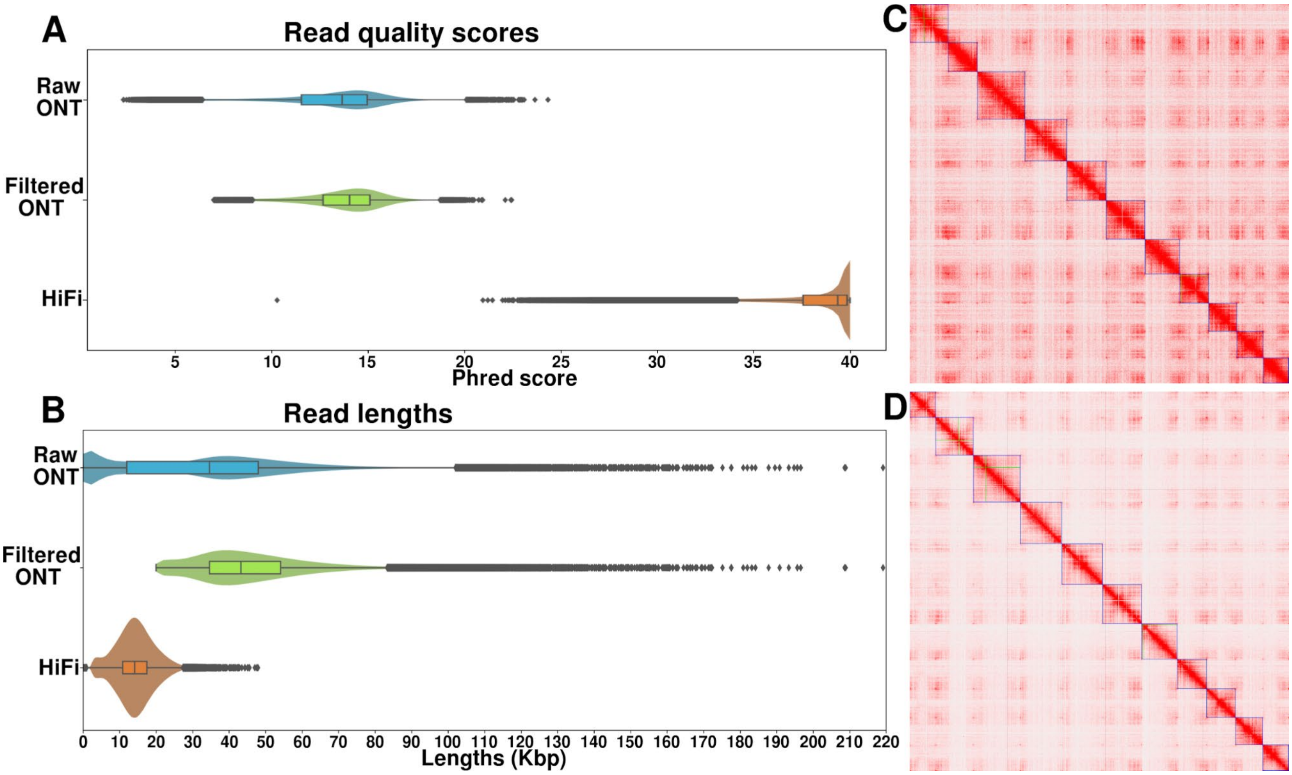


**Fig. 2** Long-read sequencing statistics and Hi-C contact map. (**A**) Summary statistics for all assembly sequencing data (HiFi, raw ONT, and filtered ONT). (**B**) Violin plot of read quality scores. (**C**) Violin plot of sequencing read lengths. (**D**) Hi-C contact map of scaffolded contigs in *E. regnans* haplotype (1) E) Hi-C contact map of scaffolded contigs in *E. regnans* haplotype (2) Hi-C contact heatmaps were visualised with Juicebox [22]

**Table 2** *E. regnans* genome assembly statistics

| | Haplotype 1 | Haplotype 2 |
|---|---|---|
| Scaffolded genome Size (bp) | 523,250,160 | 504,553,124 |
| Identified telomeres | 21/22 | 22/22 |
| % of genome in scaffolds | 91.60% | 96.07% |
| Scaffolded N50 (bp) | 43,985,167 | 48,117,852 |
| Scaffold count | 11 | 11 |
| Contig N50 (bp) | 36,825,038 | 37,748,426 |
| Contig L50 | 7 | 7 |
| Number of contigs | 795 | 269 |
| Genome BUSCO complete and (Duplicated) | 99.27% (2.24%) | 99.14% (1.98%) |
| Repetitive % (TE %) | 39.27% (38.16%) | 41.77% (40.67%) |
| Predicted gene candidates | 71,726 | 64,961 |
| Proportion of genome in predicted genes | 14.09% | 13.67% |
| Gene BUSCO complete and (Duplicated) | 97.25% (13.93%) | 95.23% (12.51%) |

**Table 3** Hi-C chromosome conformation capture linkage statistics

| | Haplotype 1 | Haplotype 2 |
|---|---|---|
| Hi-C Read Pairs | 26,768,465 | |
| Inter-chromosomal | 2,387,828 (8.92%) | 2,456,369 (9.18%) |
| Intra-chromosomal | 3,408,313 (12.73%) | 3,384,634 (12.64%) |
| Uninformative Read Pairs | 20,972,324 (78.35%) | 20,927,462 (78.18%) |

## Assembly, scaffolding, and telomere identification

As eucalypt genomes have a highly conserved karyotype (11 diploid chromosomes) with minimal differences in total genome size [2, 4, 23], a successful assembly was expected to generate a diploid genome with 11 chromosomes per haplotype, approximately 500–550 Mbp in size. Using all data types as input into HiFiasm (UL) (ONT, HiFi, and Hi-C reads), the *E. regnans* genome was assembled into two haplotypes of the expected size. Haplotype 1 contained 795 contigs with a size of 523 Mbp, and haplotype 2 contained 269 contigs with a size of 505 Mbp (Table 2). The Contig N50s for haplotypes 1 and 2 were 36.83 Mbp and 37.75 Mbp, respectively. After assembly, we investigated our contigs for contamination sequences, finding and removing 18.1 Mbp in haplotype 1 and 17.6 Mbp in haplotype 2. Hi-C reads were independently aligned to both *E. regnans* haplotypes, followed by removal of low-quality aligned reads (MAPQ < 30), chimeric reads, and PCR duplicates. This revealed approximately 2.42 million (9.05%) read pairs contained inter-chromosomal linkage information, and approximately 3.40 million (12.69%) contained intra-chromosomal linkage information (Table 3). Using the linkage information from Hi-C read pairs, both haplotypes were scaffolded into 11 pseudo-chromosomes, representing the correct number of chromosomes (Fig. 2.D, 2.E). Further analysis revealed a total of 5 joins within the two haplotypes. Haplotype 1 had a single join in Chromosomes 2 and 11, while haplotype 2 had joins in Chromosomes 2, 7, and 9.

BUSCO analysis indicated high completeness for both haplotypes (Tables 2 and Supplementary Table S1). Scaffolding was assessed by aligning both haplotypes to *E. grandis* [24] and against each other, Supplementary Figures S1, S2, and S3. Scaffolding was confirmed and all scaffolds were named according to the *E. grandis* chromosome names, which is the custom for *Eucalyptus*. All telomeres were identified for haplotype 2, and 21 of 22 telomeres were identified for haplotype 1. Haplotype 1 on Chromosome 4 was missing the 5' telomere (Supplementary Table S2). The identified telomere sequence was AA ACCCT. This telomere sequence has also been observed in the majority (282 of 332) of Dicotyledons (Magnoliopsida) listed in the telomeric repeat database (https://github.com/tolkit/a-telomeric-repeat-database*).*

## Genome annotation and gene orthogrouping

Both haplotypes were *de novo* annotated for transposable elements (TE), simple repeats, and genes (Table 2). Repeat annotation resulted in the identification of ~40.52% of both haplotypes as repetitive, of which ~39.42% was TE. After soft masking, both haplotypes were annotated for genes. Haplotype-specific HMM models were trained on all available NCBI [25] gene transcripts for *A. thaliana* (Taxonomy ID: 3702) and *Myrtaceae* (Taxonomy ID: 3931). Subsequently, 71,726 and 64,961 genes were predicted for haplotype 1 and 2 respectively.

To examine how similar (or dissimilar) the gene content of the two *E. regnans* haplotypes are, all primary transcripts (longest) were orthogrouped. Orthogrouping places highly similar genes into groups, within and between haplotypes. Genes within each orthogroup are identical genes, gene duplicates or members of the same gene family, and have identical or highly similar function. Of a predicted 125,904 primary transcripts, 97,314 (77.3%) were placed into an orthgroup, the remaining 28,590 (22.7%) were found to be too dissimilar to all other transcripts and not placed within an orthogroup. A total of 86,149 primary transcripts were found to be shared between both haplotypes (haplotype 1: 41,598 haplotype 2: 44,551). The remaining transcripts (haplotype 1: 24,492 haplotype 2: 15,263) were unique to each haplotype (Table 4). Orthogrouping created 39,959 groups, of which 36,882 (95.4%) were shared.

## Gene functional annotation

To explore the potential functions of haplotype-specific and shared genes, all transcripts were functionally annotated. After choosing the best functionally annotated

**Table 4** *E. regnans* genes and orthogroups

| | | Haplotype 1 | Haplotype 2 |
|---|---|---|---|
| Genes | Predicted gene candidates | 66,090 | 59,814 |
| | Genes in orthogroups | 48,880 (74.0%) | 48,434 (81.0%) |
| | Number haplotype-specific genes | 24,492 (37.1%) | 15,263 (25.5%) |
| | Number of shared genes | 41,598 (62.9%) | 44,551 (74.5%) |
| Orthogroups | Total number of orthogroups | 39,959 | |
| | Number of shared orthogroups | 36,882 (95.4%) | |
| | Haplotype-specific orthogroups | 1,765 (4.6%) | 1,312 (3.4%) |
| | Number of orthogroups containing haplotype | 38,647 (96.7%) | 38,194 (95.6%) |

transcript for each gene based on the lowest e-value and highest score, we examined their COG (Clusters of Orthologous Groups) categories (Fig. 3). Functional annotation successfully annotated 67.79% and 69.26% of genes for haplotype 1 and haplotype 2 respectively. Only 48.01% of all genes unplaced within an orthogroup were successfully functionally annotated, contributing the largest proportion to all non-functionally annotated genes. Genes not placed within an orthogroup and not functionally annotated may be false positives. Comparing COG categories of shared and non-shared genes revealed several categories containing different proportions of shared and haplotype-specific genes. These COG categories were found within genes associated with metabolism and information storage and processing, and also poorly characterised genes.

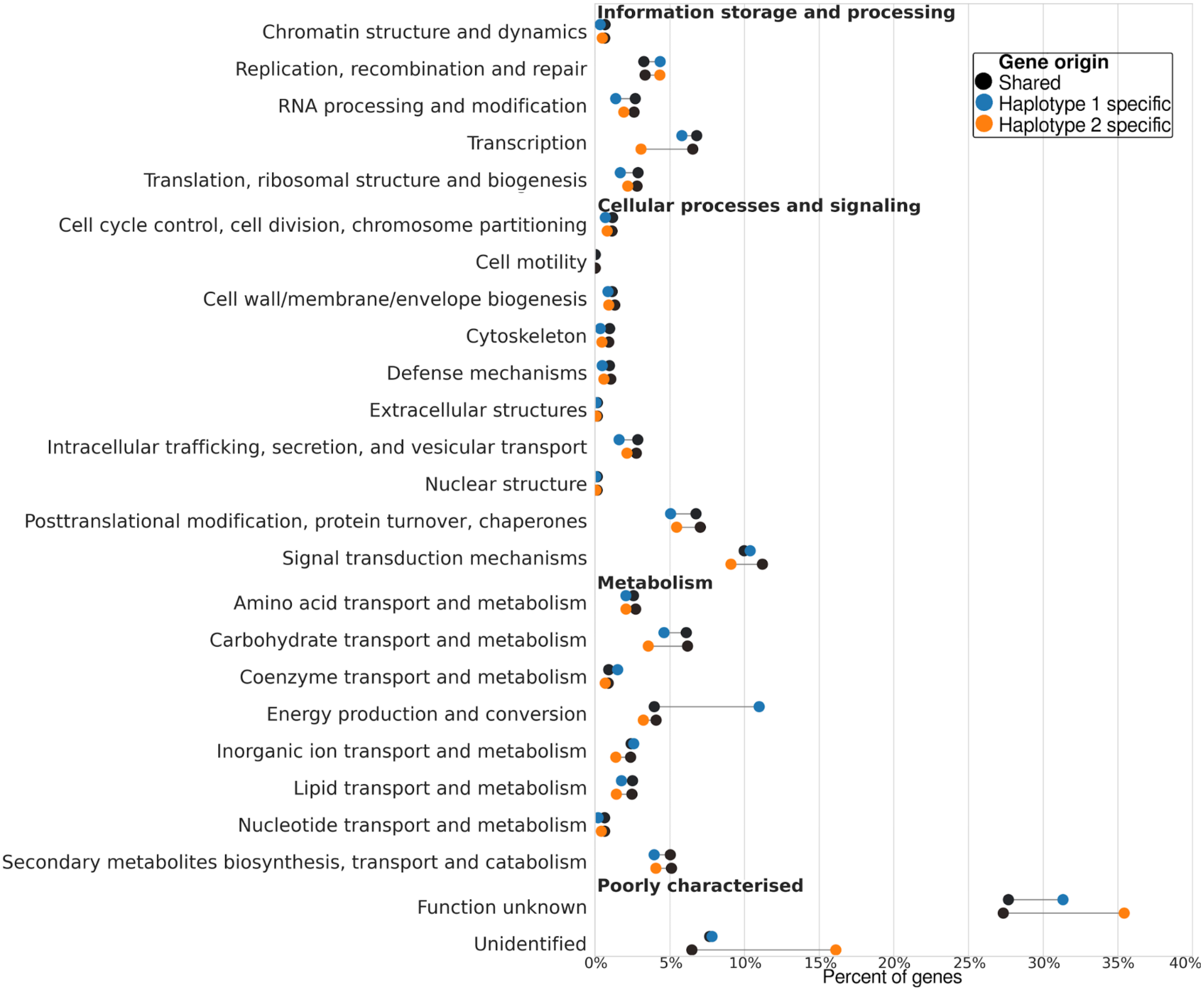


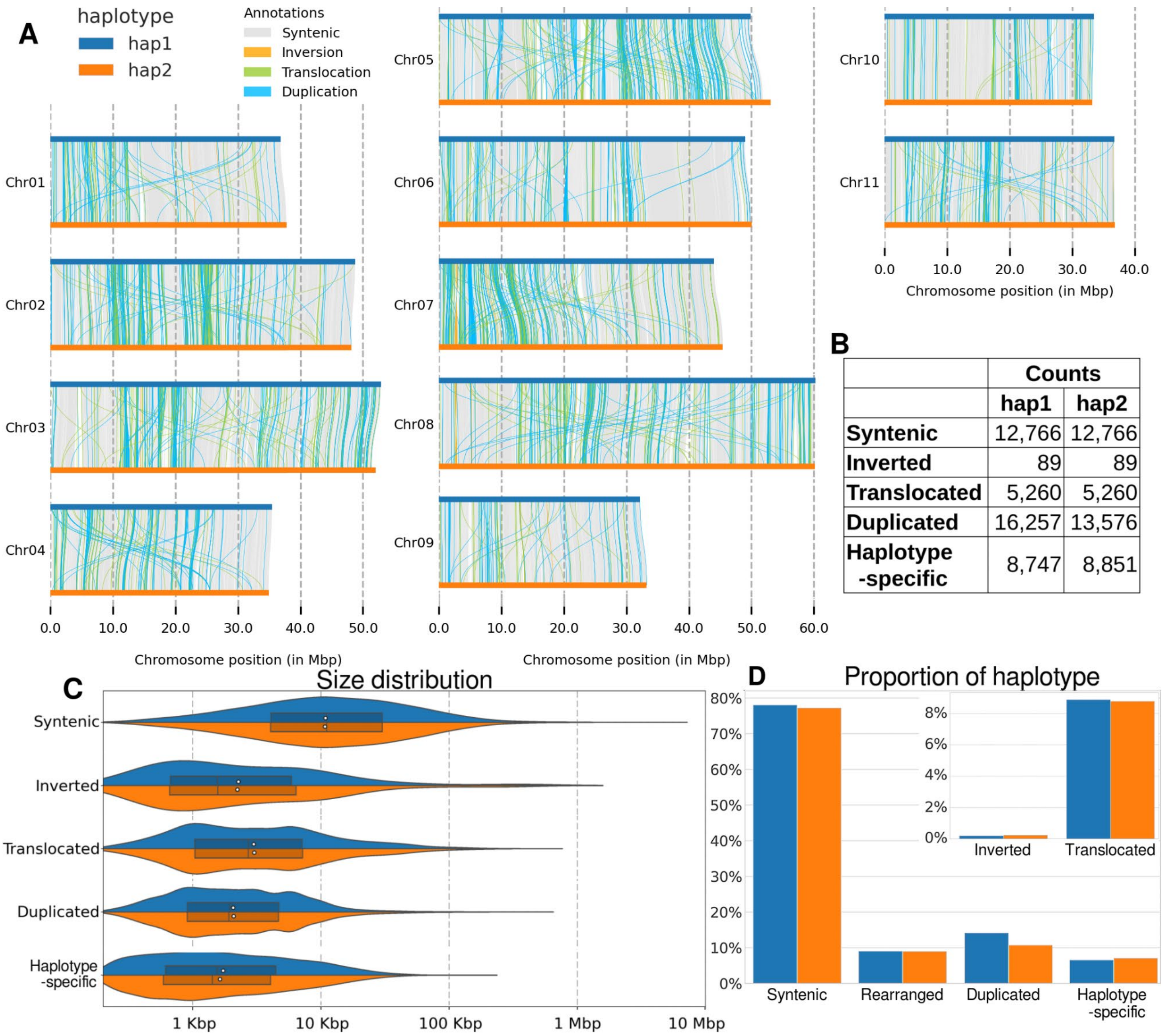
**Fig. 3** Investigating gene functions in both haplotypes. After functionally annotating all transcripts, the COG category of the best scoring transcript was chosen. All transcripts for both haplotypes were further categorised as shared, or haplotype-specific. Note: No genes were annotated as "General function prediction only"

## Genome synteny and structural variation between haplotypes

In addition to examining the gene differences between the two *E. regnans* haplotypes, we also examined the conservation of genome structure. After aligning the two haplotypes to each other, synteny, inversions, translocations, duplications, and haplotype-specific regions were annotated (Fig. 4.A). Haplotype 1 was found to be 78.11% syntenic to haplotype, conversely haplotype 2 was 77.28% syntenic to haplotype 1. The remaining proportion contained numerous structural variations, including inversions (haplotype 1: 0.20%; haplotype 2: 0.24%), translocations (haplotype 1: 8.89%; haplotype 2: 8.79%), duplications (haplotype 1: 6.21%; haplotype 2: 6.54%), or haplotype-specific regions (haplotype 1: 6.59%; haplotype 2: 7.14%), representing insertions/deletions (Fig. 4.D). Further examination of all genome regions revealed that syntenic regions are very large and very common, inversions are rare, translocations are moderately common, duplications are very common, and haplotype-specific regions are moderately common (Fig. 4.C, 4.B).

### Distribution of genes and TEs within across haplotypes

To examine the impact of structural variations on the genic content of each haplotype, the location of all genes was analysed. This analysis classified genes as originating



| | Counts | |
|---|---|---|
| | hap1 | hap2 |
| Syntenic | 12,766 | 12,766 |
| Inverted | 89 | 89 |
| Translocated | 5,260 | 5,260 |
| Duplicated | 16,257 | 13,576 |
| Haplotype -specific | 8,747 | 8,851 |

**Fig. 4** *E. regnans*genome synteny and structural variation. Haplotype 1 and 2 were aligned and all genome regions classified as syntenic, inverted, translocated, duplicated, or haplotype-specific. (**A**) Karyotype plot shows locations of all genome regions, except inter-chromosomal translocations. (**B**) The total number of each annotation type present within each haplotype. (**C**) Size distribution of genome regions in both haplotypes. (**D**) The proportion of each haplotype annotated as syntenic, haplotype-specific, inverted, translocated, and duplicated

in part of the genome that was syntenic, inverted, translocated, duplicated or haplotype specific. Similarly, the location of TEs were analysed, to determine if the inverted, translocated, duplicated or haplotype specific regions resulted from the movement, insertion, or deletion of TEs.

Analysis of genes revealed that the majority resided in syntenic regions, ~87.34% of shared genes and ~49.25% of haplotype-specific genes (Fig. 5). Notably, the proportion of haplotype-specific genes within syntenic regions was significantly lower than that of shared genes. Shared genes outside of syntenic regions were predominantly found within duplications (~8.21%), with a few found in translocations (~4.23%). Inversions (~0.11%) and haplotype-specific (~0.11%) regions contained very few shared genes. Haplotype-specific genes were predominantly found within duplications (~34.04%) and, to a lesser extent, translocations (~16.34%), outside syntenic regions. This distribution significantly differed from shared genes, which were rarely found in these regions. Inversions (~0.19%) and haplotype-specific (~0.19%) regions again contained minimal shared genes. TE location analysis showed a similar trend, with the majority found within syntenic regions (~70.62%). The remaining TEs were found predominantly in duplications (~14.33%), haplotype-specific (~7.98%) and translocated regions (~6.86%). Inversions contained very few TEs.

## Discussion

### A chromosome-level, haplotype phased genome resource forEucalyptus regnans

In this study, we generated high-coverage, long-read sequencing data, consisting of PacBio HiFi and ONT ultra-long, and chromosome conformation capture (Hi-C), to assemble a high-quality, haplotype-resolved genome for *E. regnans*, the Centurion. Achieving contig N50s>36 Mbp per haplotype and 11 chromosome-scale scaffolds of N50s>43 Mbp per haplotype, it is the most contiguous diploid *Eucalyptus* genome to date (Supplementary Table S3) [4, 24, 26]. The assembly was highly complete, achieving BUSCO completeness scores>99% per haplotype. We identified 43 out of 44 expected telomeres, with haplotype 1 lacking only a single telomere, indicating a near T2T complete genome assembly. Long-read sequencing combined with advancements in *de novo* assembly algorithms, such as Hifiasm [17] and Verkko [27] has provided an unprecedented level of insight into genomes, by retaining haplotype sequence data that would typically be collapsed and/or removed in previous genome assembly pipelines. A critical advancement has been the recent integration of both long-read sequencing platforms (PacBio and ONT), and Hi-C data, in a single de novo assembler HiFiasm (UL) [17], utilised here in our study. This hybrid approach creates better genome assemblies by utilising both HiFi and ultra-long ONT reads to create a genome graph, and Hi-C data to improve haplotype phasing, resulting in highly contiguous, phased genomes.

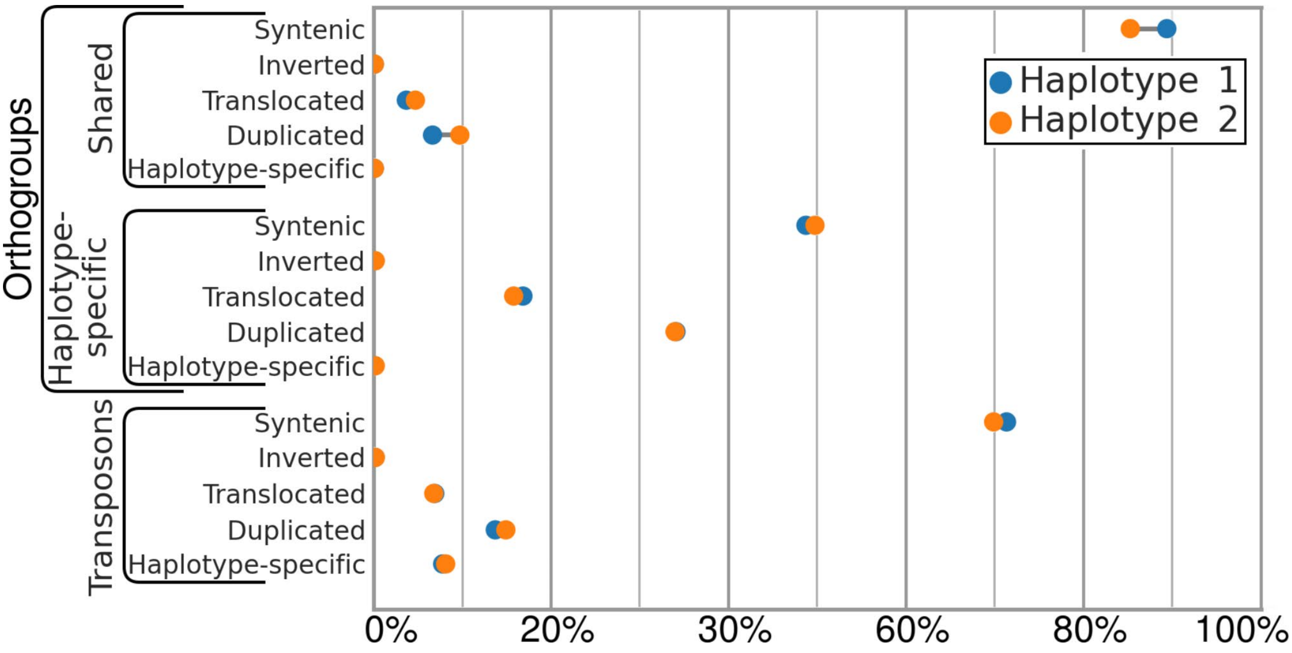


**Fig. 5** Comparison of the distribution of genes and TEs. All genes and TEs were classified based on their location within the genome (syntenic, inverted, haplotype-specific, duplicated, or translocated). Additionally, genes were categorised as shared (belonging to an orthogroup shared by both haplotypes) or haplotype-specific (belonging to a haplotype-specific orthogroup or unplaced)

### *Eucalyptus* genome architecture is shaped by structural variations

Investigation of the two haplotypes of *E. regnans* revealed ~ 93.13% sequences were shared, and the remaining ~ 6.87% was found to be haplotype-specific, likely originating from insertion and deletion polymorphisms. The shared sequences were highly syntenic (~ 77.7%), but also had high levels of structural variations, including inversions (0.2%), translocations (8.8%), and duplications (~ 12.5%). These structural variations were dramatically higher in abundance, size and distribution, in contrast to the highly syntenic haplotypes observed in agricultural crop genomes, such as rice [28], mango [29], grapes [30] and strawberry [31, 32]. Our findings align with a recent study across 33 *Eucalyptus* genomes that revealed an interplay of stable genome structure and accumulation of structural variations that drive genome divergence over time [4]. Similar results are being observed in other wild Myrtaceae family plants, such as Melaleuca [33]. However, the abundance of structural variations between the two haplotypes of a single *E. regnans* tree suggests a much greater degree of genetic variation within this species than previously assumed. This highlights the value of haplotype phased reference genomes and underscores the need for pan-genome approaches to capture the full spectrum of genetic diversity [34].

Given the very high coverage of HiFi and ONT data, and Hi-C used in this study, the assembled genome is of very high quality. This high quality genome enables a reliable analysis of inter-haplotype structural variants (SVs). However, it is important to note that some of the identified SVs may be false positives. Therefore, the validity of any particular SV should be confirmed using population data or validated through PCR or cytogenetic techniques.

Comparative analysis of annotated genes revealed a high degree of conservation between haplotypes, with most of the orthogroups and genes found in both. However, a substantial number of genes were haplotype-specific. To investigate the origin of these unique genes, their locations were compared to syntenic, inverted, translocated, duplicated, and haplotype-specific regions. This analysis revealed a clear distinction: half of the haplotype-specific genes resided within structurally rearranged regions, primarily duplications, with a smaller proportion found in translocations. Conversely, shared genes were overwhelmingly located within syntenic regions. Similar to the analysis of genes, TE locations were examined to determine if structural variations between the haplotypes originated from TE movement. The majority of TEs resided in regions unaffected by inversions, translocations, duplications, insertions, or deletions. A small fraction was found within duplications, with an even smaller proportion residing in translocated and insertion/deletion regions. These findings suggest that while the core genome of *E. regnans* exhibits a high degree of synteny, extensive structural variations, particularly duplications, play a significant role in shaping the unique features of each haplotype. This dynamic genome structure, potentially fueled by TE induced recombination errors, may be a key factor underlying the remarkable adaptability of *Eucalyptus* to diverse environmental conditions [35, 36].

Placing all transcripts within COG categories indicated that the most significant difference in haplotype gene complements were in energy production and conservation, and transcription. This suggests that haplotype-specific variation contributes to environmental adaptation in *E. regnans*. Genes related to energy production and conservation could allow different individuals to thrive in different environmental niches [37, 38]. Similarly, variation in genes associated with transcription might enable them to respond to specific environmental signals by differentially regulating gene expression [39, 40]. These findings highlight the potential role of haplotype- or SV-specific genes in driving environmental adaptation.

While the genes annotated in this study used state of the art methods, this approach relied solely on gene homology, which may introduce false positives and false negatives. Future transcriptomic studies will shed light on the functional implications of the structural variations identified in our research.

### An increasing need to conserve Australia's giant *Eucalyptus* tree forests

Climate change is rapidly altering the environment worldwide, with increased intensity of drought, fire and floods [41–43]. This is having a significant impact on Australia's iconic eucalypts, with extreme weather causing high tree mortality rates and dieback of forests [44]. Furthermore, unprecedented, mega-bushfires that occurred 2019–2020 caused widespread destruction of the natural landscape, especially eucalypt forests [8, 9]. This included *E. regnans* forests, and indeed the potentially record-breaking Centurion tree was partially burnt (Fig. 1A), highlighting vulnerability of even the giants. Fires are particularly concerning for typically wet *E. regnans* forests, a keystone species with limited fire tolerance [45]. Unlike other eucalypts that can regrow vegetatively after fire (epicormic resprouters), *E. regnans* relies solely on seeds for regeneration (obligate seeder) [46], which take decades to become mature. Furthermore, tall trees such as *E. regnans* are huge stores of above and below-ground carbon (stem and root mass). Their loss would have catastrophic consequences for the forest ecosystem and loss of large carbon stores [47]. Loss of giant trees would greatly diminish the forest's ability to annually sequester more atmospheric carbon, as these trees are still growing and gaining mass. The unique habitat provided by these forests would be substantially lost, impacting dependent wildlife populations and overall biodiversity [1]. For instance, tall, old growth *E. regnans* forests provide critical nesting sites and cavities (hollows) needed for a high biodiversity of birds and arboreal marsupials [48]. Therefore, protecting *E. regnans* forests and their towering giants becomes increasingly critical [49]. This study's contribution lies in providing a high-quality genome of *E. regnans*, using samples from the Centurion itself. This resource will be instrumental in future research efforts aimed at understanding *E. regnans* population genomic diversity, complex growth traits including carbon capture and storage, and informing tall forest ecosystem conservation strategies.

## Conclusions

In this study, we assembled a high-quality, near T2T complete, haplotype-resolved diploid genome reference for *E. regnans*, the Centurion, a leading contender for the world's tallest known flowering plant. This resource represents the most contiguous and complete reference genome for a *Eucalyptus* species, offering a foundation for future research into population genomics, functional genomics, and conservation of this ecologically significant tree species. Analysis revealed extensive structural variations and gene content differences between the two Centurion haplotypes, highlighting the remarkable genomic variation within *E. regnans*. Among the numerous SVs observed between haplotypes, gene-containing duplications were particularly abundant. These duplications may have contributed to the development of *de novo* genes [50], potentially driving novel functions and the divergence of *E. regnans* [51–53]. Further exploration of this variation through pan-genomic or genome-graph approaches could provide deeper insights into the extent of the species' genomic variation [34]. This is becoming tractable, given highly accurate long-reads from sequence consensus [54], specific base caller models [55], and deep learning error correction methods [56].

Sampling and sequencing additional trees across populations and performing genotype-environment associations can help uncover the molecular mechanisms of how *E. regnans* navigates variable environments, climates and potential threats like drought or fire [57]. This will lead to a better understanding of the genetic basis of environmental adaptation, carbon capture, and other key biological processes in *E. regnans*. Such knowledge is crucial for informing sustainable management practices and conservation efforts to protect carbon-dense giant tree forests and the unique ecosystems they support.

Methods.

### Sample collection and DNA extraction

The *Eucalyptus regnans* tree known as Centurion is located in the Huon Valley of Southern Tasmania, approximately 50 km SW of Hobart within the forestry estate (GPS −43.07708, 146.76859). Standing in an isolated small terrace of intact forest, surrounded by post-clearfell regeneration of *Eucalyptus*, it is thirty km south of the world's tallest known flowering forest grove, the Tall Trees Reserve of the Styx Valley [58]. Tree identification and climbing to sample leaf tissue was performed by Yoav D Bar-Ness (Giant Tree Expeditions). This leaf tissue was sent by local postal service with a cool pack to Australian National University, Canberra, where it was cryogenically stored in a -80 °C until DNA extraction and Hi-C preparation. A voucher specimen of *E. regnans* the Centurion is publicly available at the Tasmanian Herbarium, Hobart, Australia (accession number: HO598012, project code: GESA 003).

High-molecular weight DNA was extracted following our previously described magnetic bead-based protocol [59]. After homogenising the leaf material with a mortar and pestle, the optional sorbitol wash was performed to help further remove polysaccharides and secondary metabolites, notably oils and phenolic compounds. After extraction, the DNA was size selected for fragments ≥ 20 kb using a BluPippin (Sage Science) for PacBio HiFi sequencing and ≥ 40 kb using a PippinHT (Sage Science) for ONT sequencing.

### Long-read sequencing of native DNA

For PacBio HiFi sequencing, the HMW DNA was sheared to approximately 18 kb fragments with a Megaruptor 3 (Diagenode), using 1 cycle 31x speed and 1 cycle at 32x speed. A PacBio SMRTbell library was prepared according to the manufacturer's SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences). Sequencing was performed on a PacBio Revio 25 M SMRT cell, using circular consensus sequencing (CCS) to generate high-accuracy HiFi reads. DeepConsensus was automatically performed on the PacBio Revio, which increased sequencing accuracy [54].

The ONT ultra-long reads were generated as part of our previous study [4]. In brief, ONT native DNA sequencing libraries (1D Genomic DNA by Ligation SQK-LSK109) were sequenced on MinION Mk1B devices, using two FLO-MIN106D R9.4.1 flow cells. Flow cells were washed and re-loaded, twice per flow cell (Flow Cell Wash Kit EXP-WSH004).

### Chromosome conformation capture with Hi-C

A proximity ligation library for chromosome conformation capture was created with a Phase Genomics Proximo Hi-C (Plant) Kit (version 4), according to the manufacturer's instructions (document KT3040B). This kit utilised

DpnII, HinFI, MseI, DdeI to digest the genome, sticky ends were then filled with biotin labelled nucleotides and the subsequent blunt ends were re-ligated to neighbouring molecules. The library was multiplexed with other projects and sequencing was performed on a NovaSeq 6000 (Illumina), using an S4 flow cell with a 300 cycle kit (150 bp paired-end sequencing).

### Assembly and scaffolding

All PacBio HiFi reads from the Revio were used in the assembly (≥Q20). For ONT reads, both read ends were trimmed of 200 bp, followed by filtered to length of ≥1 kb and Q7, with NanoFilt (version: 2.8.0) [60]. For Hi-C reads, Illumina adapter sequences were removed (--nextera) and read pairs validated, using Trim Galore! (version: 0.6.10) [61].The *de novo* genome assembly was performed with Hifiasm ultra-long (UL) (version: 0.19.6-r595) [17], incorporating the PacBio HiFi, ONT ultra-long (--ul) and Hi-C reads (--h1 --h2). After assembly, both haplotypes were screened for contaminant contigs using BlobTools [62], which utilised minimap2 (version: 2.24) [63] and blast (version: 2.11.0) [64]. Subsequently both haplotypes were independently scaffolded using YaHS (version: 1.2a.2) [65] following the Arima Genomics mapping pipeline [66]. Briefly, bwa mem (0.7.17) [67] independently aligns both R1 and R2 Hi-C read sets to the current haplotype. Alignments were then filtered to remove chimeric reads and reads with poor MAPQ scores, subsequently the R1 and R2 alignment files were merged. Next, PCR duplicates were removed using Picard Tools (version: 2.26) [68]. Processed and combined pair-end alignments were then analysed with YaHS, generating a Hi-C contact map. As our ONT reads are very long and had high coverage, YaHS was run without the assembly error correction step. YaHS's Hi-C contact map was checked, and when necessary, manually edited using Juicebox (version: 2.17) [69]. After manual curation, scaffolds were finalised with Juicer tools (version: 1.6) [70], producing chromosome-scale de novo genomes. Using tidk (version: 0.2.41) [71] candidate telomere sequences were generated and each candidate subsequently tested using seqtk telo (version 1.4) [72]. BUSCO completeness analysis was performed using compleasm (version: 0.2.2) [73]. All commands used during assembly and scaffolding, including their parameters, are available at https://github.com/fergsc/Eucalyptus-regnans-genome.

### Annotation

De novo repeat libraries were generated for each haplotype using EDTA (version: 1.9.6) [74], including both simple repeats and transposable elements (TEs), and annotated with RepeatMasker (version: 4.0.9) [75]. The repeat-masked genomes were annotated for genes using BRAKER3 (version: 3.0.6) [76]. BRAKER3 aligned training proteins to our haplotypes using DIAMOND (version: 0.9.24) [77], and subsequently the ProtHint (version: 2.6.0) [78] pipeline generated the training data for AUGUSTUS (version: 3.5.0) [79]. Training protein sequences were obtained from the National Center for Biotechnology Information (NCBI) [25], including all available transcripts for Myrtaceae (Taxonomy ID: 3931) and Arabidopsis thaliana (Taxonomy ID: 3702).

Predicted gene candidates were subsequently organised into orthogroups using Orthofinder (version: 2.5.5) [80]. All candidate genes were functionally annotated for eggNOG orthogroup, COG category, GO term, KEGG term, and PFAM using eggNOG-mapper (version: 2.1.12; parameters: -m diamond --itype CDS --tax_scope Viridiplantae) [81].

### Genome alignments

We identified all shared sequences between our two *E. regnans* haplotypes through alignment using the MUMmer tool (version: 3.23) [82] with NUCmer (--maxmatch -l 40 -b 500 -c 200). NUCmer identified all shared 40-mers between genomes and merged adjacent 40-mers into a single alignment. Alignments were filtered to remove those < 200 bp and with an identity < 80% using MUMmer's delta-filter tool. We chose a conservative 80% sequence identity threshold considering the high heterozygosity of *Eucalyptus* [3], and a higher score may incorrectly filter out real alignments. The filtered NUCmer alignments were then analysed for syntenic, inverted, translocated, and duplicated regions using SyRI (version: 1.6.3) [83]. All unaligned regions were annotated as haplotype-specific. A karyotype plot showing synteny and structural variations between haplotypes was created with Plotsr [84]. All commands to align and structurally annotate the two haplotypes, including their parameters, are available at https://github.com/fergsc/Eucalyptus-regnans-genome.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10810-4.

Supplementary Material 1

Supplementary Material 2

**Acknowledgements**

We would like to thank the ACRF Biomolecular Resource Facility at the John Curtin School of Medical Research, ANU in Canberra, Australia, where PacBio Revio sequencing was conducted. This research acknowledges the support provided by NCRIS-enabled Bioplatforms Australia infrastructure. Computational resources were provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme. Additional resources were provided by the Australian National University Research School of Biology Dayhoff server. Collection of E. regnans (the Centurion) was performed by Yoav D Bar-Ness as part of Giant Tree Expeditions. Collection and associated fieldwork was supported by the Tasmanian Museum and Art Gallery Jayne Wilson Bequest. We also thank Dean

Nicolle for supplying the species occurrence distribution map of E. regnans in Australia and his ongoing dedications to Eucalyptus trees.We would like to acknowledge and thank Australia's First Nations Peoples, the Traditional Custodians of the land on which E. regnans grows. We acknowledge and thank Elders past, present and emerging.

### Author contributions

S.F. performed genome scaffolding, bioinformatic analysis and managed the data. A.J. performed DNA extractions, sequencing, de novo assembly. Y.D.B-N. performed fieldwork and sampling. All authors contributed to writing and review of the manuscript.

### Funding

This project was funded by a PacBio Revio 25 M SMRT cell trial at the ACRF Biomolecular Resource Facility at the John Curtin School of Medical Research (ANU in Canberra, Australia), and internal funds from ANU Research School of Biology (Canberra, Australia) for contributions to teaching, both granted to Ashley Jones. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an National Collaborative Research Infrastructure Strategy (NCRIS)-enabled capability supported by the Australian government. Scott Ferguson was supported by an Australian Government Research Training Program (RTP) Scholarship.

### Data availability

All raw sequencing data is available on the NCBI under the biosample accession number SAMN14929765. The genome generated for haplotype 1 is available at NCBI under bioproject accession number PRJNA1062543. Haplotype 2 is available at NCBI under bioproject accession number PRJNA1062542. Annotations (genes, repeats, and syntey/rearrangement) are available as figshare at https://figshare.com/projects/A_diploid_chromosome-level_genome_of_Eucalyptus_regnans/211543 Scripts used for assembling, contamination filtering, scaffolding, annotating, and performing alignments can be found on our GitHub at https://github.com/fergsc/Eucalyptus-regnans-genome

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 30 June 2024 / Accepted: 17 September 2024
Published online: 30 September 2024

### References

1. Bennett AF. Eucalypts, wildlife and nature conservation: from individual trees to landscape patterns. Proc R Soc Vic. 2016;128:71–86.
2. Carvalho GMA, Carvalho CR, Soares FAF. Flow cytometry and cytogenetic tools in eucalypts: genome size variation × karyotype stability. Tree Genet Genomes. 2017;13:106.
3. Murray KD, Janes JK, Jones A, Bothwell HM, Andrew RL, Borevitz JO. Landscape drivers of genomic diversity and divergence in Woodland Eucalyptus. Mol Ecol. 2019;28:5232–47.
4. Ferguson S, Jones A, Murray K, Andrew R, Schwessinger B, Borevitz J. Plant genome evolution in the genus Eucalyptus is driven by structural rearrangements that promote sequence divergence. Genome Res. 2024. https://doi.org/10.1101/gr.277999.123.
5. Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, et al. The draft nuclear genome assembly of Eucalyptus pauciflora: a pipeline for comparing de novo assemblies. GigaScience. 2020;9:giz160.
6. Sanger JC, Ferrari A. The Grove of Giants: Tasmania's most carbon-dense forest. Austral Ecol. 2023;48:1245–51.
7. Keith H, Mackey BG, Lindenmayer DB. Re-evaluation of forest biomass carbon stocks and lessons from the world's most carbon-dense forests. Proc Natl Acad Sci U S A. 2009;106:11635–40.
8. Abram NJ, Henley BJ, Sen Gupta A, Lippmann TJR, Clarke H, Dowdy AJ, et al. Connections of climate change and variability to large and extreme forest fires in southeast Australia. Commun Earth Environ. 2021;2:1–17.
9. Collins L, Bradstock RA, Clarke H, Clarke MF, Nolan RH, Penman TD. The 2019/2020 mega-fires exposed Australian ecosystems to an unprecedented extent of high-severity fire. Environ Res Lett. 2021;16:044029.
10. Williams JL, Lindenmayer D, Mifsud B. The largest trees in Australia. Austral Ecol. 2023;48:653–71.
11. Esri ArcWatch. February 2010 - World's Tallest Eucalyptus Tree Found with Lidar and GIS. https://www.esri.com/news/arcwatch/0210/the-centurion.html. Accessed 28 Mar 2024.
12. Sillett SC, Van Pelt R, Kramer RD, Carroll AL, Koch GW. Biomass and growth potential of *Eucalyptus regnans* up to 100 m tall. Ecol Manag. 2015;348:78–91.
13. Burgess G. Tall tree Centurion passes 100-metre mark, creating milestone for Tasmanian wilderness. ABC News. 2018.
14. Mifsud BM, Harris GJ. Victoria's giant trees: a contemporary survey. Vic Nat. 2016;133:36.
15. Shenkin A, Chandler CJ, Boyd DS, Jackson T, Disney M, Majalap N et al. The World's tallest tropical tree in three dimensions. Front Glob Change. 2019;2.
16. Sillett SC, Van Pelt R, Koch GW, Ambrose AR, Carroll AL, Antoine ME, et al. Increasing wood production through old age in tall trees. Ecol Manag. 2010;259:976–94.
17. Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat Methods. 2024;21:967–970.
18. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.
19. Yue J, Chen Q, Wang Y, Zhang L, Ye C, Wang X, et al. Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit Actinidia chinensis. Hortic Res. 2023;10:uhac264.
20. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere assembly of the maize genome. Nat Genet. 2023;55:1221–31.
21. Nicolle D. Taller eucalypts for planting in Australia: their selection, cultivation and management. Dean Nicolle; 2016.
22. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.
23. Ribeiro T, Barrela RM, Bergès H, Marques C, Loureiro J, Morais-Cecílio L et al. Advancing Eucalyptus Genomics: Cytogenomics reveals conservation of Eucalyptus genomes. Front Plant Sci. 2016;7.
24. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of Eucalyptus grandis. Nature. 2014;510:356–62.
25. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;49:D10–7.
26. Shen C, Li L, Ouyang L, Su M, Guo K. E. urophylla × E. Grandis high-quality genome and comparative genomics provide insights on evolution and diversification of eucalyptus. BMC Genomics. 2023;24:223.
27. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;41:1474–82.
28. Abdullah M, Furtado A, Masouleh AK, Okemo P, Henry RJ, Abdullah M et al. An improved haplotype resolved genome reveals more rice genes. Trop Plants. 2024;3.
29. Wijesundara UK, Masouleh AK, Furtado A, Dillon NL, Henry RJ. A chromosome-level genome of mango exclusively from long-read sequence data. Plant Genome. 2024;n/a(n/a):e20441.
30. Zhang K, Du M, Zhang H, Zhang X, Cao S, Wang X, et al. The haplotype-resolved T2T genome of teinturier cultivar Yan73 reveals the genetic basis of anthocyanin biosynthesis in grapes. Hortic Res. 2023;10:uhad205.
31. Mao J, Wang Y, Wang B, Li J, Zhang C, Zhang W, et al. High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. Hortic Res. 2023;10:uhad002.
32. Zhou Y, Xiong J, Shu Z, Dong C, Gu T, Sun P, et al. The telomere-to-telomere genome of Fragaria vesca reveals the genomic evolution of Fragaria and the origin of cultivated octoploid strawberry. Hortic Res. 2023;10:uhad027.
33. Chen SH, Martino AM, Luo Z, Schwessinger B, Jones A, Tolessa T, et al. A high-quality pseudo-phased genome for Melaleuca quinquenervia shows allelic diversity of NLR-type resistance genes. GigaScience. 2023;12:giad102.

34. Schreiber M, Jayakodi M, Stein N, Mascher M. Plant pangenomes for crop improvement, biodiversity and evolution. Nat Rev Genet. 2024;25:563–577.
35. Ferguson S, Jones A, Murray K, Schwessinger B, Borevitz JO. Interspecies genome divergence is predominantly due to frequent small scale rearrangements in Eucalyptus. Mol Ecol. 2022;n/a n/a.
36. Ferguson S, Jones A, Murray K, Andrew RL, Schwessinger B, Bothwell H, et al. Exploring the role of polymorphic interspecies structural variants in reproductive isolation and adaptive divergence in Eucalyptus. GigaScience. 2024;13:giae029.
37. Huot B, Yao J, Montgomery BL, He SY. Growth–defense tradeoffs in plants: a Balancing Act to Optimize Fitness. Mol Plant. 2014;7:1267–87.
38. Monson RK, Trowbridge AM, Lindroth RL, Lerdau MT. Coordinated resource allocation to plant growth–defense tradeoffs. New Phytol. 2022;233:1051–66.
39. Ni F-T, Chu L-Y, Shao H-B, Liu Z-H. Gene expression and regulation of higher plants under Soil Water stress. Curr Genomics. 2009;10:269–80.
40. Burman A, Garcia-Milian R, Whirledge S. Gene X environment: the cellular environment governs the transcriptional response to environmental chemicals. Hum Genomics. 2020;14:19.
41. Bevacqua E, Vousdoukas MI, Zappa G, Hodges K, Shepherd TG, Maraun D, et al. More meteorological events that drive compound coastal flooding are projected under climate change. Commun Earth Environ. 2020;1:1–11.
42. Cook BI, Smerdon JE, Cook ER, Williams AP, Anchukaitis KJ, Mankin JS, et al. Megadroughts in the common era and the Anthropocene. Nat Rev Earth Environ. 2022;3:741–57.
43. Linley GD, Jolly CJ, Doherty TS, Geary WL, Armenteras D, Belcher CM, et al. What do you mean, 'megafire'? Glob Ecol Biogeogr. 2022;31:1906–22.
44. Losso A, Challis A, Gauthey A, Nolan RH, Hislop S, Roff A, et al. Canopy dieback and recovery in Australian native forests following extreme drought. Sci Rep. 2022;12:21608.
45. Waters DA, Burrows GE, Harper JDI. Eucalyptus regnans (Myrtaceae): a fire-sensitive eucalypt with a resprouter epicormic structure. Am J Bot. 2010;97:545–56.
46. Nicolle D. A classification and census of regenerative strategies in the eucalypts (Angophora, Corymbia and Eucalyptus—Myrtaceae), with special reference to the obligate seeders. Aust J Bot. 2006;54:391–407.
47. Collins L, Day-Smith ML, Gordon CE, Nolan RH. Exposure to canopy fire reduces the biomass and stability of carbon stored in fire tolerant eucalypt forests. Ecol Manag. 2023;528:120625.
48. Lindenmayer D, Bowd E. Critical ecological roles, structural attributes and conservation of Old Growth Forest: lessons from a case study of Australian Mountain Ash forests. Front Glob Change. 2022;5.
49. Lindenmayer D, Taylor C, Bowd E, Ashman K. The case for listing Mountain Ash forests in the Central Highlands of Victoria as a threatened Ecological Community. Pac Conserv Biol. 2023;30:NULL–NULL.
50. Pegueroles C, Laurie S, Albà MM. Accelerated evolution after gene duplication: a time-dependent process affecting just one Copy. Mol Biol Evol. 2013;30:1830–42.
51. Adams KL, Wendel JF. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 2005;8:135–41.
52. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 2008;148:993–1003.
53. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725–32.
54. Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat Biotechnol. 2023;41:232–8.
55. Ferguson S, McLay T, Andrew RL, Bruhl JJ, Schwessinger B, Borevitz J, et al. Species-specific basecallers improve actual accuracy of nanopore sequencing in plants. Plant Methods. 2022;18:137.
56. Stanojević D, Lin D, de Sessions PF, Šikić M. Telomere-to-telomere phased genome assembly using error-corrected simplex nanopore reads. bioRxiv 2024;:2024.05.18.594796.
57. Lasky JR, Josephs EB, Morris GP. Genotype–environment associations to reveal the molecular basis of environmental adaptation. Plant Cell. 2023;35:125–38.
58. Rudman T, Balmer J. Giant trees and very tall forest values in the tasmanian wilderness World Heritage Area. Department of Primary Industries, Parks, Water and Environment; 2018.
59. Jones A, Torkel C, Stanley D, Nasim J, Borevitz J, Schwessinger B. High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. PLoS ONE. 2021;16:e0253830.
60. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
61. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B et al. TrimGalore. Zenodo. 2023;:https://doi.org/10.5281/zenodo.7598955
62. Laetsch DR, Blaxter ML, BlobTools. Interrogation of genome assemblies. F1000Research. 2017;6:1287.
63. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
64. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
65. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39:btac808.
66. Arima Genomics' mapping pipeline. 2023.
67. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;:https://arxiv.org/abs/1303.3997
68. Picard toolkit. Broad Institute, GitHub repository. 2019.
69. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing Loop-Resolution Hi-C experiments. Cell Syst. 2016;3:95–8.
70. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
71. Brown M, De la González P, Mark B. A Telomere Identification Toolkit. 2023. https://doi.org/10.5281/zenodo.10091385
72. Li H. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences. 2013.
73. Huang N, Li H. Compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics. 2023;39:btad595.
74. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.
75. Smit A, Hubley R, Green P. RepeatMasker Open–4.0. 2020. http://www.repeatmasker.org. Accessed 11 Feb 2020.
76. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77
77. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77.
78. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinforma. 2020;2:lqaa026.
79. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
80. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
81. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, Orthology assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38:5825–9.
82. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.
83. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:277.
84. Goel M, Schneeberger K. Plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics. 2022;38:2922–6.

## Publisher's note