Haplotype-resolved genome assembly of an F_1 hybrid of *Eucalyptus urophylla* x *E. grandis*

by

Anneri Lötter

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

July 2021

Under the supervision of Professor Alexander A. Myburg

and co-supervision of Doctor Tuan A. Duong, Professor Eshchar Mizrachi

and Professor Jill L. Wegrzyn

Declaration

I, Anneri Lötter declare that this dissertation, which I hereby submit for the degree MSc Genetics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution

Anneri Lötter

30 July 2021

Dissertation Summary

Haplotype-resolved assembly of an F_1 hybrid genome of *Eucalyptus urophylla* x *E. grandis*

Anneri Lötter

Supervised by **Prof. A.A. Myburg**

Co-supervised by **Dr T.A. Duong, Prof. E. Mizrachi** and **Prof. J.L. Wegrzyn**

Submitted in partial fulfilment of the requirements for the degree ***Magister Scientiae.***

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

Eucalyptus interspecific hybrids are used to develop fast-growing, disease resistant clonal varieties in Eucalypt breeding. *Eucalyptus urophylla* x *E. grandis* hybrids are currently the most widely planted eucalypt hybrid combination in subtropical and temperate regions worldwide, as pure species plantations of either *E. urophylla* or *E. grandis* have limited deployment ranges and have lower success. In crop species with multiple high-quality reference genomes, breeding strategies that incorporate haplotype or structural variant information have greater success rates. The availability of single-molecule DNA sequencing technologies, in combination with phased genome assembly strategies, have enabled assembly of multiple genomes for the same plant species to a level where haplotype and structural variants can be assessed.

To determine if phased genome assembly strategies can be used effectively to assemble haplogenomes for *Eucalyptus*, we made use of a trio-binning strategy in combination with Nanopore sequencing

technology (Oxford Nanopore Technologies), to assemble phased parental haplogenomes of an interspecific F_1 hybrid of *E. urophylla* and *E. grandis*. In addition, we performed a whole-genome comparison between the assembled haplogenomes to identify structural variants between the two genomes.

The objectives of this MSc study were to i) assess the validity and success of using a trio-binning based genome assembly approach to assemble the two haplogenomes of an F_1 *E. urophylla* x *E. grandis* interspecific hybrid, to ii) generate high-quality phased reference genomes for both parental species as a first step towards a *Eucalyptus* reference pan-genome of haplotype and structural variation and to iii) identify genomic similarities and differences between *E. urophylla* and *E. grandis* based on a whole-genome comparison.

The highly heterozygous nature of the F_1 eucalypt hybrid enhanced separation of Nanopore sequencing data into parental read groups, and 99.98% of reads could be grouped into either parental haplotype. Separate assembly of the resulting read bins resulted in a 544.51 Mb *E. urophylla* haplogenome and a 566.72 Mb *E. grandis* haplogenome assembly, with a contig N50 of at least 3.9 Mb and a BUSCO completion score of greater than 98.8% before scaffolding. Scaffolding using high density genetic linkage maps of both parents resulted in placement of more than 88% of the assembled haplogenome contigs onto a pseudo-chromosome assembly. Subsequently, a genome-wide comparison between the haplogenomes allowed identification of 48,729 structural rearrangements between *E. urophylla* and *E. grandis*.

The success of the trio-binning haplogenome assembly approach shows that it is a promising method to construct the pan-genome of haplotype- and structural variation in eucalypts. The results of this study shows that this approach can be applied in other *Eucalyptus* hybrids for *de novo* reference genome assembly and haplotype- and structural variant discovery. We further show that SVs are more pervasive

than previously though between the two parental species genomes. Future studies will focus on discovery of genes underlying the identified SVs, including more individuals to create a pan-genome of SVs and to understand how these SVs may influence traits of importance to breeding.

Preface

Eucalyptus is an economically important hardwood tree genus of importance to the forestry industry. *Eucalyptus* hybrids have a greater potential environmental footprint as these hybrids combine favourable characteristics of both parental species. The most widely planted *Eucalyptus* hybrid combination are interspecific hybrids of *Eucalyptus grandis* and *E. urophylla*. These hybrids combine the fast growth and desirable wood properties of the subtropical/temperate species, *E. grandis,* with the superior disease resistance of the tropical species, *E. urophylla*. However, to better understand and exploit hybrid compatibility and performance and further improve the predictive accuracy for tree deployment, we require ever more accurate breeding strategies. Current breeding strategies employ molecular markers to guide breeding and deployment decisions. Unfortunately, such markers have limited capability of describing the causal allelic variants underlying desired characteristics due to sampling little of the genome. Haplotype-based molecular breeding strategies have been shown to be more accurate as it samples more of the genome and takes haplotype and structural variant information into account allowing discovery of causal allelic variants.

Advances made in long-read sequencing technologies, improved genome assembly and structural variant calling programs, have allowed the assembly of multiple phased genomes for *Arabidopsis* (Jiao & Schneeberger, 2020) and tomato (Wang *et al.*, 2020b; Alonge *et al.*, 2020b), allowing discovery of structural variants. In addition to SV discovery, related studies have revealed some of the functional impacts these variants have. In tomato SV have been shown to influence i) fruit flavour through multiple SV haplotypes, ii) fruit size as a result of increased gene expression in duplicated regions and iii) how many fruits are produced per plant due to epistatic interactions between SV (Alonge *et al.*, 2020b), and in *Arabidopsis* they have been shown to influence recombination patterns between chromosomes (Jiao & Schneeberger, 2020). These studies highlight that a single reference genome cannot explain the phenotypic diversity observed within and between populations and species. As such, there is a movement

towards assembly of a pan-reference genome, a concept that incorporates variants from multiple individuals in a species (reviewed by Sherman & Salzberg, 2020; Bayer *et al.*, 2020).

Our research group is focused on providing the South African forestry industry with methods that improve their international breeding competitiveness, with a focus on *Eucalyptus*. As seen above, SV have a direct impact on breeding traits, and incorporating haplotype information in breeding decisions were shown to result in better prediction accuracy of resulting crop performance (Ogawa *et al.*, 2018, 2019). However, to discover haplotype and structural variants requires a high-quality reference genome that is phased. The reference genome that is currently available for *E. grandis* is still quite fragmented and scaffolds are a mosaic representation of chromosomes, in other words, both haplotypes are combined together into the reference genome. As a result, using the current reference genome to discover haplotype and structural variants will be difficult. A reference haplogenome assembled from long-read sequencing data will make variant discovery much easier.

For this reason, the overall aim of this study was to reconstruct the haplogenomes of the *E. grandis* and *E. urophylla* parents that are present within an interspecific hybrid of *E. urophylla* x *E. grandis* and to identify sequence and structural differences between the two species. To evaluate the possibility to achieve this in the context of using a trio-binning strategy, we had the following objectives: 1) generate at least 50X coverage Nanopore sequencing data for the F_1 hybrid and Illumina sequencing data for the parents to enable trio-binning, 2) assemble phased parental haplogenomes present in the F_1 hybrid and 3) identify local and structural variants via genome-wide comparison of the haplogenomes. We were able to apply trio-binning read separation to separately assemble the parental haplogenomes of an F_1 hybrid. In addition, genome-wide comparison of the resulting haplogenomes allowed us to identify structural variants between the parental species.

Chapter 1 provides an overview of plant genome assembly challenges and advances. For the section on *Eucalyptus*, I provide an overview of genomic resources currently available for eucalypt tree improvement as well as background to *E. urophylla* as a species and hybrid with *E. grandis*. I also give an overview of challenges related to plant genome assembly, how these challenges have been overcome in the past and how long-read sequencing technologies have aided in overcoming these challenges. I give a brief overview of the best strategies and programs for genome assembly. Lastly, I discuss how long-read sequencing has advanced understanding of genomic variant functions and helped in molecular breeding.

In **Chapter 2**, I describe the assembly of the *E. urophylla* and *E. grandis* haplogenomes that are present within a F_1 *E. urophylla* x *E. grandis* hybrid using a trio-binning approach to separate the parental haplotypes. In addition, I identify structural variants and annotate repeat elements. Results indicate that this strategy can be applied to other eucalypt hybrid combinations to construct a reference pan-genome for the species.

I report on research undertaken from January 2019 to December 2020 in the Department of Biochemistry, Genetics and Microbiology and the Forestry and Agricultural Institute (FABI) at the University of Pretoria. This study was completed under the supervision of Prof A.A. Myburg and co-supervised by Dr T.A. Duong, Prof. E. Mizrachi and Prof. J.L. Wegrzyn. The first-generation hybrid as well as both parental pure species used in this study was constructed and maintained by Sappi Forest Research (Hilton, KZN, South Africa).

Preliminary results of this MSc have been presented at the following national and international conferences:

International:

Anneri Lötter, Julia Candotti, Tuan A. Duong, Eshchar Mizrachi, Jill L. Wegrzyn and Alexander A. Myburg, Structural variant discovery in haplotype resolved genomes of Eucalyptus grandis and E. urophylla, July 19-23, Plant Biology 2021 Worldwide Summit (Poster presentation).

Anneri Lötter, Tuan A. Duong, Julia Candotti, Eshchar Mizrachi, Jill L. Wegrzyn and Alexander A. Myburg, Phased assembly of an F1 Eucalyptus urophylla x E. grandis hybrid genome using trio-binning approach, July 17-21, Plant Biology 2020 Worldwide Summit (Poster presentation).

Anneri Lötter, Tuan A. Duong, Eshchar Mizrachi, Jill L. Wegrzyn and Alexander A. Myburg, Sequencing and Phased Assembly of an Eucalyptus urophylla x E. grandis F1 Hybrid and Parental Genomes, 11 – 15 January 2020, San Diego, California, USA, Plant and Animal Genome XXVIII Conference (Poster presentation).

National:

Anneri Lötter, Julia Candotti, Tuan A. Duong, Eshchar Mizrachi, Jill L. Wegrzyn and Alexander A. Myburg, Towards haplotype-based genomic breeding: Phased assembly of a *Eucalyptus urophylla* x *E. grandis* F1 hybrid genome, 08 – 11 March 2020, Pretoria, South Africa, 13th Southern African Plant Breeding Symposium 2020 (Speed talk).

Acknowledgements

I would like to express my sincere appreciation and gratitude to the following people and institutes for their involvement and support throughout this MSc:

- Prof Zander Myburg for always being inspiring and going above and beyond to cultivate a research environment that promotes growth as a researcher and achievement of research excellence. I also would like to thank you sincerely for the opportunities you gave me to improve my bioinformatic skills abroad and expand my knowledge through attendance of conferences. It is really much appreciated.
- Dr Tuan Duong for always being supportive and helping me understand genome assembly concepts. Thank you for your patience when explaining concepts and willingness to always help.
- Prof Jill Wegrzyn for her guidance in plant genome assembly and allowing me to visit. The experience, guidance and expertise you provided was invaluable for completion of my project and future studies.
- Prof Eshchar Mizrachi for reminding us of the biological and aspect of the study and those important what do you want to achieve questions during committee meetings.
- Prof Fourie Joubert for teaching me how to navigate Linux and for endless bioinformatic support.
- Dr Stephanie Cornelissen (former Agricultural Research Council - Biotechnology Platform Genomics Application Specialist and current Roche Red Scientist II) and Ms Mary Ranketse for the modified SDS-based DNA extraction protocol
- Ms Frances Lane for her friendship, encouragement, support and tea meetings via zoom during the COVID lockdown.
- Ms Melissa Reynolds for her friendship and support.
- The Computational Biology Core at the University of Connecticut for use of Bioinformatic and Computational resources.
- Members of the Plant Computational Genomics research group

    - Ms Susan McEvoy, Ms Sumaira Zaman, Mr Jeremy Bennett and Mr Alex Trouern-Trend for helping me figuring out where I have issues with my scripts and sharing expertise.
- Sappi Forest Research for providing the plant material for this project.
- Department of Science and Innovation (DSI), Technology Innovation Agency (TIA) and Technology and Human Resources for Industry Programme (THRIP) for funding the project.
- National Research Foundation (NRF) for partial bursary funding.
- Department of Biochemistry, Genetics and Microbiology and the Forestry and Agricultural Biotechnology Institute (FABI) at the University of Pretoria for providing facilities and a excellent research environment.
- UP Postgraduate Studies Abroad Programme for providing funding for a research visit to the Plant Computational Genomics laboratory at the University of Connecticut, Connecticut, USA.
- Oxford Nanopore Technologies Ltd for a travel support bursary for a poster presentation presented at the International Plant and Animal Genome (PAG XXVIII) conference.
- My parents and brother for their support, unconditional love and encouragement. I have great appreciation for all the things you have done for me.

Table of Contents

List of Tables

List of Figures

Chapter 1 Literature review

Plant genome assembly and phasing using a long-read sequencing approach

1.1. Introduction

Predicted human population growth, modelled to be over 9.8 billion by 2050, and anthropogenic climate change promise to place increasing strain on land use, agriculture, energy production and natural resources (Chase *et al.*, 2011). Fast-growing plantation trees, such as widely planted eucalypts, are renewable resources for biomaterial (timber), bioenergy and various medicinal essential oils. Exotic plantations, which allows management of timber production as rotational crops, could also alleviate strain on natural forests (Grattapaglia & Kirst, 2008). As such, good breeding and deployment strategies are needed for plantations that exploit phenotypic plasticity (Rezende *et al.*, 2014) to improve the environmental footprint of plantations, whilst maintaining the ability to provide sustainable end products.

The efficiency of breeding strategies is influenced by our ability to adapt current breeding strategies so that they address the predicted outcomes associated with climate change. Employing molecular markers to assist with breeding promise to improve management of genetic resources in breeding programmes and increase accuracy of matching genotypes with suitable environments for improved production and response to environmental challenges (Grattapaglia and Kirst 2008). Molecular markers like microsatellite markers (short sequences of 2 – 10 nucleotides that are repeated multiple times in the genome), more recently replaced with single nucleotide polymorphisms (SNPs, single base variations with a frequency of greater than 1% in the population, Vignal *et al.*, 2002), are currently being used for routine management and genomic selection of eucalypts. SNP chips are available due to availability of the current reference genome for *E. grandis* and extensive resequencing of other genus members (Silva-Junior *et al.*, 2015). However, recent advances in single-molecule long-read sequencing technologies, may improve the existing resources even more by allowing dissection of the haplotype (blocks of genetic variants found on one homologue of a chromosomal set, Zheng *et al.* 2016) and structural variation responsible for phenotypic variation of multiple eucalypt species.

Plant genomes are particularly challenging to assemble, due to their high repetitive content, high levels of ploidy and heterozygosity, and large genome sizes (Kyriakidou *et al.*, 2018) which often results in highly fragmented genome assemblies. However, long-read sequencing (LRS) technologies has enabled more complete assembly with greater contiguity for many reference genomes, as LRS platforms offer greater read lengths that span repetitive sequences. Long-reads can also span haplotype variants, creating opportunities for studying haplotype variation and its incorporation in molecular breeding strategies. In addition to improving the quality of current reference genomes, LRS is advancing the field of genomics by allowing variant identification in the context of building reference pan-genomes (Jiao & Schneeberger, 2019) and giving users greater insight into transcriptomic and epigenomic landscapes (Sedlazeck *et al.*, 2018a; Alonge *et al.*, 2020b), even for non-model species (Jansen *et al.*, 2017).

This review is focused on challenges associated with plant genome assembly, and how LRS technologies can be used to overcome these challenges with a focus on using haplotype information for haplotype-based molecular breeding in the plantation forestry industry, specifically for *E. urophylla* and *E. grandis*. As there is a lot more information available for *E. grandis*, a brief introduction to *E. urophylla* is given. The reader is referred to the paper by Myburg *et al.*, (2014) if more information is required on *E. grandis*. A few computational approaches and challenges of LRS are discussed, however this review does not provide information on the full scope of programs available for *de novo* genome assembly, or all the short- and long-read platforms available for genome sequencing or haplotype variant identification (for more information, the reader is referred to reviews by Basantani *et al.* (2017), Jiao and Schneeberger (2017), Kyriakidou *et al.* (2018) and Sedlazeck *et al.* (2018)).

1.2. Genomic resources for *Eucalyptus*

Eucalypts are the most widely planted hardwood trees worldwide, comprising more than 20 million hectares (Global *Eucalyptus* map 2009 - Eucalyptologics: GIT Forestry consulting information resources on *Eucalyptus* cultivation worldwide). Fast growth, desirable wood properties, environmental

adaptability, suitability to vegetative propagation and reduction of pressure on native forest species (Iglesias & Wiltermann; Bauhus *et al.*, 2010; Rezende *et al.*, 2014) are major drivers for the use of eucalypts as renewable resources for pulp and paper production, biomaterial and bioenergy as well as various essential oils (Grattapaglia & Kirst, 2008). In the face of climate change, their environmental adaptability and phenotypic plasticity makes them sustainable resources for the growing human population and efficient breeding strategies are needed to effectively reduce the environmental footprint of *Eucalyptus* plantations while increasing production.

Eucalypts are part of the angiosperm family Myrtaceae, which are dicotyledonous woody plants, native to Australia and the islands to its north (Ladiges *et al.*, 2003). More than 700 species are recognised (Brooker, 2000), of which most are outcrossing (Moran *et al.*, 1989; Gaiotto *et al.*, 1997) with hermaphroditic flowers that are pollinated by insects (Byrne, 2008). Eucalypt genomes are highly heterozygous and genome size varies between species, mostly due to non-transposable element derived changes (Myburg *et al.*, 2014). The majority of eucalypts are diploid with n = 11 chromosomes (Grattapaglia *et al.*, 2012).

1.2.1. *Eucalyptus urophylla* and its hybrids

Eucalyptus urophylla is part of the section *Latoangulatae* (Brooker, 2000) and has an estimated genome size of 650 Mb (Grattapaglia & Bradshaw Jr., 1994). It is one of four *Eucalyptus* species that has a natural range outside Australia (Brooker & Kleinig, 1983), being native to the Lesser Sunda Islands of eastern Indonesia. *E. urophylla* occupies areas from almost sea level up to 3000 m elevation (Eldridge *et al.*, 1993) and has the greatest altitudinal range of all eucalypts (Gunn & McDonald, 1991). It has been described as one of the most genetically variable eucalypt species, with some provenances proposed as a separate species (Pryor *et al.*, 1995).

The genetic an morphological diversity in *E. urophylla* may be partly due to introgression from *E. alba* (Dvorak *et al.*, 2008) with which it shares a habitat at lower elevation, allowing natural hybridisation to occur (Martin & Cossalter, 1975). As a result, some *E. urophylla* selections used in breeding are actually hybrids that have *E. alba* genes, with the exception of selections made from Timor island (Dvorak *et al.*, 2008). This shared habitat and occurrence of hybridisation makes it difficult to identify the genetic makeup of either *E. alba* or *E. urophylla* (Dvorak *et al.*, 2008), as the level of introgression is unknown.

Hybrid breeding incorporating multiple species is advantageous in areas where pure species are not suited to the environment. Intra- and inter-specific hybrids of many plant species exhibit heterosis, whereby traits are better in the hybrid offspring compared to that of the parents (Goulet *et al.*, 2017). As a result, hybrid clones make up a large portion of existing commercial plantations and have a positive influence on forestry productivity, product quality and production costs (de Assis, 2000; Grattapaglia & Kirst, 2008). For example, in *Eucalyptus*, *E. grandis* is favoured in the plantation forestry industry due to its fast growth, coppicing ability and suitability to the pulpwood industry. Unfortunately, it is very susceptible to canker development and foliar fungal pathogens, resulting in severe plantation losses (Vigneron *et al.*, 2000). By combining the *E. grandis* genotype with that of *E. urophylla*, the resulting hybrids have good survival rates, greater disease tolerance and higher wood density of *E. urophylla* and the rapid early growth characteristics of *E. grandis* (Retief & Stanger, 2009).

1.2.2. Current genomic resources for *Eucalyptus* breeding

Current breeding strategies in *Eucalyptus* makes use of microsatellite and SNP markers. Clonal identification, estimation of distance between individuals and species distinction is possible using a microsatellite marker panel consisting of 18 markers which was developed for use in *Eucalyptus* breeding (Faria *et al.*, 2011). Following the release of the *E. grandis* reference genome (Myburg *et al.*, 2014), Silva-Junior *et al.* (2015) developed a high-throughput, multi-species *Eucalyptus* SNP Chip (EUChip60K) containing 59,222 highly transferable and polymorphic SNPs for all major eucalypt

species (approximately providing one SNP every 11.8 kb on average). As SNP markers occur more frequently in the genome (Mammadov *et al.*, 2012), thus covering more of the genome, they perform better than microsatellite markers in terms of data quality, accuracy, reproducibility, robustness and cost-effectiveness (Telfer *et al.*, 2015). In addition, SNP data provides a resource for additional studies to be made into molecular breeding, genomic selection, population genomics and gene discovery by genome-wide association study (Silva-Junior *et al.*, 2015).

Myburg *et al.* (2014) assembled 605 Mb of the estimated 640 Mb *E. grandis* genome into 11 pseudomolecules, using Sanger sequencing, paired BAC-end sequencing and high-density genetic linkage-maps. A total of 4,941 scaffolds remained unanchored (totalling 85 Mb), corresponding mostly to repeat-rich sequences (as much as 44.5% of the genome is made up of repeat elements) and sequences containing haplotype variation. The genome was predicted to contain 36,376 protein coding genes, of which 84% share gene clusters with other rosid lineages and 34% were within tandem duplications (Myburg *et al.*, 2014). A second version of the *E. grandis* genome was released by Bartholome *et al.* (2015), which captures 88.6% (612.6 Mb) of the genome, mainly improving upon scaffolding errors. The high number of unanchored scaffolds demonstrates that assembly of near complete chromosome-scale plant genomes is a difficult task to accomplish.

New long-read sequencing and assembly strategies allow assembly of less fragmented reference genomes for many species, albeit only one such assembly is currently available for *Eucalyptus*. The genome of *E. pauciflora* was sequenced and assembled using a combination of Illumina short read sequencing (SRS) data and Oxford Nanopore Technologies (ONT) long read sequencing data. A total of 594.8 Mb was assembled into 416 contigs, contig N_{50} was 3.23 Mb and a 94.58% BUSCO completion score (Wang *et al.*, 2019). Even though *E. pauciflora* is not one of the major eucalypt species used in plantation forestry (Harwood, 2011; Rezende *et al.*, 2014), the genome of *E. pauciflora* provides a valuable genomic resource due to its potential use for studying structural and other variants that are

related to desirable attributes, such as cold and drought tolerance, in *Eucalyptus*. The availability of reference genomes for other eucalypt species (Low *et al.*, 2020) as well as multiple genomes from the same single species (Aucamp *et al.*, 2016; Li *et al.*, 2019; Alonge *et al.*, 2019; Song *et al.*, 2020), will allow dissection of haplotype and structural variants that are available for genomic improvement as has been done in tomato (Alonge *et al.*, 2020b).

1.3. State-of-the-art and challenges of plant genome assembly

More than 200 plant genomes have been sequenced and assembled to date (Michael & VanBuren, 2015; Kyriakidou *et al.*, 2018; Chen *et al.*, 2018, 2019). Most of these have been assembled using SRS platforms (i.e. Illumina sequencing) mainly due to the low cost associated with this technology. Although genome sequencing itself may not pose a problem, assembly of the resulting reads may be very difficult. Genome assemblies using SRS data are rarely assembled up to chromosome scale, and most of the current plant genome assemblies consist of many fragmented contigs and scaffolds, which are usually not mapped to chromosomal locations (Cao *et al.*, 2017). In addition, plant genomes are difficult to assemble due to their large genome size such as that of the loblolly pine 22 Gb (Neale *et al.*, 2014), highly repetitive due to transposable elements, polyploidy (Salman-Minkov *et al.*, 2016) and high levels of heterozygosity.

1.3.1. Difficulties associated with assembling plant genomes

Some plant genomes are very large (Jiao & Schneeberger, 2017), which has prevented high-quality genome assembly of many plant species (reviewed by Pellicer *et al.* 2018). The genome of loblolly pine, estimated at 21.6 Gb, is one of the largest assembled genomes to date. Successful assembly of this genome involved the use of haploid cells for DNA isolation and sequencing. Also, a novel computational tool, the MaSuRCA genome assembler, was developed to reduce short reads into a smaller more concise set of super-reads the assembly to a manageable scale. A final assembly of 20.1 Gb was obtained, with

a contig N_{50} = 8.2 kb and a scaffold N_{50} = 66.9 kb, using paired-end Illumina reads in combination with long-fragment mate-pair reads (Zimin *et al.*, 2014).

A second challenge to plant genome assembly is the high level of ploidy, especially prevalent within crop species (evaluated by Salman-Minkov *et al.* 2016), and this goes hand in hand with the problem of heterozygosity. Salman-Minkov *et al.* (2016) found that approximately 54% of monocot crops are polyploid, whereas 40% of related wild species are polyploid (of the 297 crop and 2,836 wild species evaluated). Ploidy complicates genome assembly by introducing heterozygosity (homologous chromosomes with two or more different alleles at a given locus). As a result, the higher the ploidy, the more heterozygosity can theoretically be expected in the genome (Kyriakidou *et al.*, 2018).

Ploidy and heterozygosity is difficult to resolve during the genome assembly process as multiple alleles from the same locus can be seen as sequences that originate from different loci by assembly algorithms (Huang *et al.*, 2017). In the case of SRS, reads are unlikely to span more than one haplotype, which causes the formation of separate contigs instead of a consensus sequence, resulting in decreased genome contiguity and inflated assembly size. As assembly algorithms try to generate a consensus sequence, rare variants may also be collapsed to obtain the greatest consensus sequence, missing important variants that may be related to species-specific traits.

Strategies that have been deployed for assembling polyploid plant genomes include reducing genome complexity via use of natural or *in vitro* generated haploids, or sequencing a diploid progenitor species to help assemble the genome of the domesticated species as seen in the case of the tetraploid peanut genome (Bertioli *et al.*, 2016). Inbred lines that are nearly homozygous can also be used, which essentially reduces the genome to a haploid state as was the case for the tetraploid upland cotton (Li *et al.*, 2015) and hexaploid wheat genomes (Brenchley *et al.* 2012). Lastly, haplotyping can be used, where

allelic variants are assigned to a specific chromosome or alleles that occur together can be defined, as in the case of the sweet potato genome (Yang *et al.*, 2017).

A third challenge to plant genome assembly is the fact that plant genomes contain many repetitive sequences. These are made up of transposable elements (TE), which proliferate within plant genomes (Pellicer *et al.*, 2018). TE make up 80 – 90% of the maize genome (Lisch, 2013) and 75% of the sunflower genome consist of long-terminal repeat retrotransposons (LTR, Badouin *et al.* 2017), which is a class of TEs. For *Eucalyptus*, 41.22% of the *E. grandis* and 44.77% *E. pauciflora* genome were made up of repeat elements of which 26.94% and 29.53% were LTR retrotransposons and 4.8% and 6.04% were DNA transposons for *E. grandis* and *E. pauciflora*, respectively (Wang *et al.*, 2020a). It is very difficult to solve repetitive regions with short reads, as reads do not span the entire repetitive sequences. As a result the assembly algorithms are unable to resolve the number of repeats and collapse them, or ends the contig when repeats are encountered, which lead to fragmented assemblies and/or mis-assemblies (Phillippy *et al.*, 2008). Most of the abovementioned assembly problems can be overcome by using longer reads, as they can span across the length of repetitive elements and connect haplotype variants.

1.4. Long read sequencing in plants

1.4.1. A brief history of genome sequencing

First generation sequencing (FGS, Sanger sequencing) played a crucial role in setting the stage for genome sequencing and assembly. Hundreds of DNA molecules could be sequenced simultaneously with high accuracy (99.999% accuracy). Unfortunately, FGS is very expensive, and had limited throughput (Liu *et al.*, 2012). Subsequently, the reduced cost of sequencing whole genomes with SRS technologies has facilitated assembly of many new genomes. Through resequencing and alignment of reads to a reference genome, studies analysing genomic diversity could also be performed at low cost

(reviewed by Koboldt *et al.*, 2013). In addition, transcription, gene regulation and epigenetic modifications could also be investigated in many species (Celniker *et al.*, 2009; Dunham *et al.*, 2012).

Even though SRS has enabled analysis of several plant and animal genomes, the limitations mentioned above have precluded assembly of complete genomes and have left many regions within assembled genomes unresolved (Chaisson *et al.*, 2015). In addition, the inability to span more than one haplotypic allele and, the generation of artefacts as a result of library preparation methods used may also contribute to the fragmented state of assembled reference genomes. A new and actively improving sequencing technique, single-molecule long-read sequencing (LRS), has enabled researchers to resolve some of these complex genomic regions due to the ability to sequence 10 – 100 kb routinely (Sedlazeck *et al.*, 2018a), as was demonstrated by Chaisson *et al.*, (2015a) for the human genome.

1.4.2. Long-read sequencing compared to second-generation sequencing

The availability of more affordable LRS technologies have presented the genomics community with opportunities to sequence and assemble high-quality genomes for any organism. As library preparation for LRS does not require amplification, it avoids the amplification biases associated with SRS technologies. Long read lengths also offer the advantage of having reads that span haplotypic variants and entire repetitive regions enabling phased assemblies and resulting in less fragmented assemblies (Jansen *et al.*, 2017). In addition, the greater read-length can span across large SV and thus enables identification of such SVs, which is a difficult task to accomplish using SRS data (Figure 1.1, Sedlazeck *et al.*, 2018b).

However, LRS does suffer from lower accuracy than SRS, and as such, many studies supplement long-read data with additional high-accuracy SRS data (Jiao & Schneeberger, 2017). Genomic features that are not identifiable using either LRS or SRS alone can be identified more effectively by using a combination of long- and short reads. For example in *Saccharomyces cerevisiae*, Nanopore-based hybrid

assemblies (incorporating Illumina and Nanopore sequencing data) were shown to have a greater number of completely assembled genes, and was able to assemble more telomeric repeats than assemblies based on Illumina sequencing data only (Istace *et al.*, 2017).

There are two main types of long-read sequencing approaches: synthetic approaches and single molecule approaches. Currently, two main synthetic systems are available: Illumina synthetic long reads (SLR) and 10X Genomics Chromium platforms. Both of these systems use short-read technologies to generate long reads *in silico*, by making use of barcodes for assembling larger fragments (Goodwin *et al.*, 2016). However, as assemblies performed using SLR hardly reach an N_{50} of greater than 100 kb and do not cover the DNA fragment from end-to-end and thus single-molecule long-read sequencing approaches are still more desirable (reviewed by Jiao and Schneeberger 2017). As such the next section of this review is focused on single-molecule LRS and the reader is referred to reviews by Goodwin *et al.*, (2016); Jiao & Schneeberger, (2017); Sedlazeck *et al.*, (2018a) and Jung *et al.*, (2019) if further information is required on either synthetic or single-molecule LRS platforms or genome assembly methods.

1.4.3. Single-molecule long-read sequencing

Single-Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio) is the most widely used long-read sequencing platform, partly because it has been commercially available for longer than other LRS technologies (Jansen *et al.*, 2017). It makes use of zero mode waveguides (ZMW), which are small wells with a DNA Polymerase enzyme attached to the bottom. DNA strands are allowed to pass through the ZMWs and the fixed polymerase enzyme allows visual tracking of nucleotide incorporation using a laser and camera system. The camera records the colour and duration of light emitted by the incorporated nucleotide at the bottom of the ZMW to determine the nucleotide (Goodwin *et al.*, 2016; Jiao & Schneeberger, 2017).

Another type of single-molecule long-read sequencing, Nanopore sequencing from Oxford Nanopore Technologies (ONT), detects electrical current changes as single-stranded DNA (ssDNA) moves through a protein nanopore to identify nucleotides. The ionic changes are measured and translated into DNA nucleotides (each shift in voltage is specific to a particular DNA sequence within the pore). Disadvantages include: high error rates due to five to six nucleotides occupying the pore simultaneously making it a challenge to identify which nucleotide is next within the ssDNA sequence, and a deletion-bias for homopolymer regions (demonstrated by Chin *et al.* 2016). A comparison between ONT and PacBio sequencing is provided in Table 1.1.

1.4.4. State of the art of long-read sequencing-based assembly of plant genomes

As a result of better accuracy and longer availability to the research community of SMRT sequencing, more plant reference genomes have been assembled using this technology. However, the average and maximum read length is lower than that offered by Nanopore sequencing and yields a lower proportion of longer reads (Belser *et al.*, 2018). Using the longest reads for assembly results in better assembly contiguity for both PacBio and ONT based genome sequencing, indicates that read length is more important than coverage for genome assembly (Schmidt *et al.*, 2017; Belser *et al.*, 2018).

In addition, using a combination of multiple types of sequencing data (i.e. short- and long-read sequencing data) for genome assembly results in better quality genome assemblies, as was demonstrated for *S. pennellii* (Schmidt *et al.*, 2017)*, A. thaliana* (Michael *et al.*, 2018) and *O. coarctata* (Mondal *et al.*, 2017). Using a hybrid assembly approach (incorporating Illumina short-reads as well as ONT long-reads for assembly), the *S. pennellii* and *A. thaliana* genome assemblies had greater contiguity than previous assemblies, and all three assemblies had an N_{50} value of > 1.8 Mb (Mondal *et al.*, 2017; Schmidt *et al.*, 2017; Michael *et al.*, 2018). The assembly of *A. thaliana* also showed that assembling the genome of an individual may be a simpler way to detect SVs that may have an impact on gene expression (Michael *et al.*, 2018; Wang *et al.*, 2020b; Alonge *et al.*, 2020b).

1.5. Software for genome assembly, annotation and phasing of ONT reads

It is easy enough to generate sequencing data if a suitable DNA extraction method has been found or developed, the real challenge lies in data analysis. Generating completely assembled and annotated genomes is a computationally intensive process, especially for plant genomes. Genome assembly involves identification of overlapping reads that can be built into contigs. The resulting assembly quality can be measured by its contiguity (how continuous the assembled fragments are), contig N_{50} size (the contig size for which all contigs of that size and larger cover 50% the genome), completeness (how much of the genes or genome is represented when looking at the proposed conserved orthologous gene set for the clade), base-level correctness and structural accuracy (Bradnam *et al.*, 2013). High-quality assemblies are desirable as they allow many insights to be made into a species, by allowing subsequent analysis such as gene annotation and identification of genomic features (Sedlazeck *et al.*, 2018a) like structural variants (Jiao & Schneeberger, 2019; Alonge *et al.*, 2020b).

An additional challenge is that assembly algorithms need to be aware of the characteristics of long reads, e.g., longer read lengths means looking for larger overlaps and the high error rate of Nanopore reads means adjusting to permit an amount error when looking for overlaps. As such, short-read assemblers may not work or need to be optimised for assembly with long-read data. In addition, the high error rate associated with long-reads require post-assembly bioinformatic solutions to handle low sequence identity (Jansen *et al.*, 2017). Reads can be corrected (or polished) using a hybrid sequencing approach (algorithms use short-read data to correct long-reads before or after assembly) or self-corrected (aligns long-reads to each other and increase long-read accuracy). Although, self-correction approaches often result in better contiguity than when correcting with short-reads, when enough coverage is available (Sedlazeck *et al.*, 2018a).

1.5.1. Assembly

There are two main algorithmic approaches for genome assembly. The first is the Overlap-Layout-Consensus (OLC) approach which looks for overlaps between sequences, creates graph layout of overlaps and reads, and generate consensus sequence (Basantani *et al.*, 2017). This method has read length flexibility and is robust against sequencing errors (Zimin *et al.*, 2013), but it is computationally intensive as it makes use of all-vs-all read comparisons (Sedlazeck *et al.*, 2018a). The second approach, called de Bruijn Graphs (DBG), is less computationally intensive (Zimin *et al.*, 2013) and replaces each read with overlapping set of fixed-length short sequences and merge short sequences that appear adjacently to form contigs, stopping at short length from repeat boundaries (Chaisson *et al.*, 2015). However, a review by Cherukuri and Janga (2016) compared nanopore-based assembly quality and the accuracy of different assemblers, and found that OLC based assemblers performed better in terms of contig N_{50}, mean contig values, number of contigs (fewer) and had lower computational run times (Cherukuri & Janga, 2016). The next section is focused on what has been found to work best, and for a detailed discussion on the programs available for genome assembly with LRS data, the reader is referred to the review by Jung *et al.*, (2019).

No single genome assembler is the absolute best, and all assemblers perform better for specific tasks or organisms, which needs to be considered when selecting a genome assembler. This was demonstrated by Istace *et al.* (2017), who found that SMARTdenovo could identify repeat regions very well, had good completeness, contiguity and speed compared to Canu and Miniasm, whereas ABruijn could assemble the mitochondrial genome of yeast completely. One may also consider a genome assembler which is has been used in other genome assemblies, to enable more accurate comparisons to be made between different genome assemblies (especially when comparing species or specific individuals in context of constructing a species pan-genome).

Genome assemblers can also be combined to produce the best assembly in a given time frame or to reduce computational time. This was demonstrated by Schmidt *et al.* (2017), where they compared different assembly methods (Canu, SMARTdenovo, Miniasm and a combination of Canu pre-corrected reads with SMARTdenovo assembly) to obtain the best assembly of the *S. pennellii* genome. Of the single assembler approaches, Miniasm had the highest N_{50}, but had the lowest alignment rate to the reference genome. Canu was second to Miniasm, however Canu required a greater amount of CPU time than either SMARTdenovo or Miniasm assemblers. The Canu and SMARTdenovo combined assembly delivered the most contiguous assembly, thus a combination of two assemblers performed better than any of the single assemblers (Schmidt *et al.*, 2017). There are also some new assemblers available such as Flye (Kolmogorov *et al.*, 2019), Shasta (Shafin *et al.*, 2020) and Necat (Chen *et al.*, 2021) that may be of interest to the reader.

After assembly, post-assembly polishing can be performed using either long-read (self-correction) or short-read (hybrid correction) data. Polishing after assembly is more effective as raw signal data and alignments can be evaluated for accuracy. However, polishing with short read data is limited, as repetitive regions cannot be confidently aligned using short reads and may lead to greater fragmentation of the genome (Sedlazeck *et al.*, 2018a). After polishing the genome may be annotated, phased (this can also be performed before polishing, but polishing may lead to collapsed heterozygosity within the genome, resulting in a mosaic assembly that is not representative of the haplogenomes within the genome) and the quality of the resulting genome assessed with BUSCO (Simão *et al.*, 2015) and/or QUAST (Gurevich *et al.*, 2013). Please refer to the review by Jung *et al.* (2019) for more information on genome polishing.

1.5.2. Phasing

Most current reference genomes do not reflect the heterozygosity present within the genome of the species, in which SVs or allelic variations between homologous chromosomes are excluded. This means

that the reference genome is not representative of spectrum of variation present in natural individuals in the species, leading to unannotated and missing genes (Kyriakidou *et al.*, 2018). Assembly of a representative heterozygous genome requires proper data handling, and in an ideal situation, reconstruction of all homologous chromosomes. The process whereby chromosomes containing variants (haplotypes) are reconstructed is called phasing (Jiao & Schneeberger, 2017). By distinguishing the maternal and paternal haplotypes (defined as sets of allelic variants that are inherited together), studies can be made into many processes associated with different allelic variants (Jiao & Schneeberger, 2017), as was found for disease resistance and fruit quality and quantity traits in tomato (Wang *et al.*, 2020b; Alonge *et al.*, 2020b). In addition, phasing haplotypes has been shown to have greater sensitivity for SVs detection (Cretu Stancu *et al.*, 2017; Garg *et al.*, 2018).

There are three main approaches to haplotype phasing: 1) based on the fact that shared haplotypes are inherited form common ancestors, phase can be inferred from genotypic information of large cohorts (statistical phasing, Browning & Browning, 2011), 2) similarly haplotypes can be inferred from genotypic data of related individuals (genetic haplotyping) and 3) haplotype sequences can be determined experimentally (molecular haplotyping, Garg *et al.*, 2016). Long-reads make haplotype assembly easier, as heterozygous variants can be phased when reads span them, since a read can span multiple variant (Jiao & Schneeberger, 2017; Sedlazeck *et al.*, 2018a).

Statistical methods for imputing haplotypes are very accurate for detecting common allelic variants, but does not conform well for rare and private variants (Jiao & Schneeberger, 2017; Sedlazeck *et al.*, 2018a). Falcon-Unzip is an example of a statistical method for haplotype imputation. Falcon-Unzip is a module implemented in the Falcon assembler that assembles long-read sequencing data into phased diploid genomes. Falcon first performs an initial assembly which is corrected with Falcon-Unzip which uses heterozygous SNPs and SVs, to identify haplotypes within the reads. The phased reads are then assembled into contigs and haplotigs (contigs representing individual chromosomes) to form a final

diploid assembly with phased SNPs and SVs. Falcon-Unzip allowed high continuity assemblies of contigs in i) an F_1 hybrid between two *Arabidopsis thaliana* strains, ii) cultivated *Vitis vinifera* cv Cabernet Sauvignon (a highly heterozygous F_1 hybrid) and iii) *Clavicorona pyxidata* (a coral fungus) that are comparable to assemblies of the individual parental genomes (Chin *et al.*, 2016).

Of the three strategies mentioned, genetic haplotyping is the most accurate and reliable approach, as genotyping of the parents or larger pedigree enables the direct identification of the parental origin of each variant (with the exception of homozygous regions), but this approach will increase the cost of the study (Sedlazeck *et al.*, 2018a). In one genetic haplotyping method employed by Koren *et al.* (2018), called trio-binning, allelic variation is resolved before genome assembly. Parental short reads are used to partition long-reads of the offspring into haplotype-specific sets. The haplotypes are subsequently assembled independently, resulting in two separately assembled haplotypes. Success of their method was tested in three organisms with varying levels of heterozygosity: i) an outbred F_1 hybrid between Angus and Brahman cattle species which resulted in two species-specific haplotypes of reference-genome quality for both species, ii) an F_1 hybrid of the same two *A. thaliana* strains used in the study by Chin *et al.* (2016) and iii) humans. The trio-binning haplotypes were found to have greater alignment identity than those generated with Falcon-Unzip (Koren *et al.*, 2018). This strategy has also been used to successfully assemble the haplotypes in flowering cherry Somei-Yoshino (Shirasawa *et al.*, 2019).

1.6. How can we use long-read data to aid molecular breeding?

A review by Bevan *et al.* (2017) highlighted four levels of sequencing approaches for crop improvement. Briefly: 1) Using LRS data to *de novo* assemble multiple reference genomes for whole-genome comparisons between species, cultivars and lineages (as was done for tomato, Alonge *et al.*, 2020) or 2) using linked-read sequencing technologies to identify SV (Saxena *et al.*, 2014), a database of haplotype- and structural variants can be constructed for the study population (pan-genome) (Bevan *et al.*, 2017). In addition, using assembled genome comparisons to find target variant regions for further investigation,

can be used in combination with more cost effective 3) low coverage sequencing (5-10X Illumina short-read skimming) to identify variation present in the study population. Lastly 4) SNPs and allelic variants that define a particular haplotype- or SV can be detected by using genotype-by-sequencing approaches to capture variation in gene-coding regions (Bevan *et al.*, 2017). The four strategies can be used in combination or singly in a top-down or bottom-up direction, depending on the resources available. The following section focuses on the use of two of these levels of sequencing, long-read sequencing and whole genome skimming, to improve crop yield and production.

1.6.1. Using deep sequencing of parents to impute offspring haplotypes in molecular breeding

Current breeding strategies require multiple (often more than six) generations of backcrossing to purge undesired variation in diploid crops (Bevan *et al.*, 2017). As trees are a long-lived species (taking up to 9 years before maturity), it is unfeasible to remove allelic variants with multi-generation breeding strategies. In addition, making desirable crosses and holding field trails are expensive and time consuming. But, by incorporating genomic resources and tools (such as SNPs and pan-genomic variant information) the process can be sped up considerably, as it allows for early identification of individuals with desirable genomic variants (Grattapaglia & Kirst, 2008). This allows breeders to focus their resources on those individuals that produce desirable products reducing the cost somewhat.

Currently, eucalypt breeding is making use of SNPs and other molecular markers for crop improvement. However improvement in sequencing technologies have enabled the use of haplotype information for crop improvement, which was shown to be more accurate it's in predictive ability for crop performance (Ogawa *et al.*, 2018, 2019; Zhang *et al.*, 2019). In addition, using haplotype and structural variants in a pan-genomic context, instead of molecular markers, has also led to the discovery of causal variants related to important breeding traits in tomato (Wang *et al.*, 2020b; Alonge *et al.*, 2020b) and allows identification and exploitation of new gene variants that are not in a single reference genome (Marschall *et al.*, 2018).

To move towards the use haplotype information for eucalypt improvement, high-quality phased reference genomes are required for the species of interest (in this case *E. urophylla* and *E. grandis*). When a reference genome is available, additional *de novo* genomes can be assembled and compared to the reference to create a database of haplotype and structural diversity (pan-genome) (Figure 1.1A and B). After the diversity panel has been constructed, one can look for associations between haplotype variants and traits of interest to identify haplotypes containing genic variants that are associated with desired phenotypes (Figure 1.1B and C). Subsequently, offspring may be screened with markers that define a haplotype (allowing statistical imputation of the offspring haplotype from parental data, Motazedi *et al.*, 2017) or with low coverage sequencing, and those with desired haplotype combinations selected for propagation (Figure 1.1D). Early selection and identification of individuals with the desired haplotypes will speed up the process of selection by eliminating the need for field trials before suitable genetic (and haplotype) combinations can be deployed (Bevan *et al.*, 2017).

1.7. Conclusion and future prospects

Climate change is expected to alter crop yield, plant-pathogen dynamics and to create more variable environments. As such, breeders need to develop crops that have a reduced impact on the environment, whilst having a higher yield and greater environmental adaptability. However, before we can improve crops, we need a clear understanding of how the genomic, phenotypic and environmental factors interact to give us a desired product. As a first step, we need to understand the genetic variation present within the crop and how these variants contribute to desired crop characteristics.

A central aim of crop plant genomics has been to assemble accurate plant genomes that represent the entire spectrum of genetic variation found within the study population (Bevan *et al.*, 2017). Plants are an extremely diverse group of organisms, as is reflected by the variation in their genome size (Bennett & Leitch, 2011), repeat content, ploidy and heterozygosity (Kyriakidou *et al.*, 2018). These factors have

made the assembly of plant genomes very challenging, and have made it necessary to develop novel strategies to tackle the problem of genome assembly, such as reducing the complexity of the genome to be sequenced and assembled (Zimin *et al.*, 2014). The emergence of LRS technologies offers a new method of taking on the plant genome assembly challenge.

LRS technologies offer the genomics community with a solution to many of the current plant genome assembly problems. Of importance to this study is the ability of reads to contain more than one haplotypic variants and span repetitive regions allowing for phased genome assemblies. Additionally, longer read-lengths allows for organellar genome assembly (Wang *et al.* 2018) and also allows sequencing of full transcripts, enabling isoform identification in transcriptomic studies (reviewed by Weirather *et al.* 2017). Nanopore sequencing specifically also offers the ability to identify base modifications (epigenetic marks (Jain *et al.*, 2018)) which has been shown to be influenced by SV and directly impact on gene expression and chromosomal recombination (Jiao & Schneeberger, 2020; Alonge *et al.*, 2020b).

Long-read sequencing of large numbers of genomes is also becoming feasible, with an expected increase in throughput and decreased cost of sequencing with the ONT PromethION platform. The genome of banana has been sequenced with the PromethION platform, and the estimated cost of sequencing a 500-600 Mb genome was $6,500 compared to the estimated $16,300 it would have cost using MinION flow-cells (Belser *et al.*, 2018). The lower cost allows genomes to be assembled for many new plant species and, the additional genomes that are becoming available for different species as well as multiple individuals of a particular species will allow us to establish a knowledge base for the types of genetic variation at disposal for crop improvement (Bevan *et al.*, 2017; Jiao & Schneeberger, 2020; Wang *et al.*, 2020b; Alonge *et al.*, 2020b).

In addition to the long-read sequencing strategies discussed above, there are long-range genome mapping strategies that allow assembly of near complete chromosomes. The BioNano optical mapping system, in which stretched DNA molecules tagged with fluorescent markers is imaged to generate a physical map that can be used to scaffold a genome and to identify SVs and haplotypes (Lam *et al.*, 2012). Chromosome conformation capture also forms part of these technologies, as it provides 3D-proximity information of genomic loci depending on how far these genomic loci are from each other (Lieberman-Aiden *et al.*, 2009). Chromosome conformation capture can be used to uncover how chromosomes fold, how genic regions interact (Lieberman-Aiden *et al.*, 2009) and can be used in combination with other sequencing technologies to order and orientate reads from long-read sequencing (Mascher *et al.*, 2017).

However, all the genotypic information needs to be combined with phenotypic and environmental data, to fully understand how these elements interact to present a given phenotype. To measure and integrate genotype-environment interactions and generate multi-dimensional datasets of these interactions, there is a need to complement genomics with large scale precision phenotyping (Bevan *et al.*, 2017). New methods for phenotyping are becoming available and offer increased resolution, precision and measurement scales for crop growth and developmental measurements (reviewed by Araus and Cairns 2014; Araus *et al.* 2018). This will also allow crop phenotype standardisation and maintenance (Zamir, 2013). Once we understand how the genome and environment interact to produce a specific phenotype, we are better armed to prepare crops for the future and to provide sustainable resources for the growing human population.

In conclusion, advances in genome sequencing technologies has enabled the use haplotype-based molecular breeding strategies, that have been shown to be more effective and accurate than molecular-marker based breeding strategies (Ogawa *et al.*, 2018, 2019). In order to prepare for changing environments and the larger human population, plantation forestry needs more accurate and effective

breeding strategies, such as haplotype-based molecular breeding. However, before haplotype-based molecular breeding can be deployed, high-quality reference genomes are required to provide a baseline for comparison. Recently, the trio-binning approach has been shown to be capable of generating two reference quality genomes for both parental species, by assembling and phasing the genome of an F_1 interspecific hybrid sequence (Koren *et al.*, 2018). As such, this project aims to assemble and phase the genome of an F_1 *E. urophylla* x *E. grandis* interspecific hybrid, to generate high-quality reference genomes for both parental species. In addition, we aim to provide the first direct whole-genome comparison between these species to gain a preview of the extent of SVs between different eucalypts within the same subsection. The availability of a high-quality reference genome can later be used to generate a pan-genome of haplotype and structural variants for the study population which will later enable haplotype-based molecular breeding in *E. urophylla* and *E. grandis*.

1.8. Tables

Table 1.1 Comparison between Pacific Biosciences (PacBio) single-molecule real-time sequencing (SMRT) and Oxford Nanopore Technologies (ONT) MinION long-read sequencing platforms.

Feature	SMRT (PacBio)	MinION (ONT)
Release date	2011	2014
Sequencer size[a]	92.7 x 86.4 x 167.6 cm	USB sized
Cost	$700K start-up, $300 / Gb data[b]	$1000 start-up, < $300 / Gb data[c]
Detection method	Optical detection of nucleotide incorporation[d]	Electrical current changes[e]
Size range	20-60 kb[f]	Limited by DNA fragment size[e]
Error rates[a]	<1% (PacBio HiFi)	15% (R9.4 chemistry) <1% (R10.4 chemistry)
Accuracy with correction[a]	99.9%[a]	> 99.3%
Advantages	Stability, high accuracy, less bias in error profile[g]	Unlimited read length, both DNA strands can be sequenced[d]

[a] Manufacturer data
[b] 2016 NGS Field Guide
[c] Giordano *et al.,* (2017)
[d] Jiao and Schneeberger (2017)
[e] de Lannoy *et al.*, (2017)
[f] Vanburen *et al.*, (2015)
[a] Jansen *et al*., (2017)

1.9. Figures

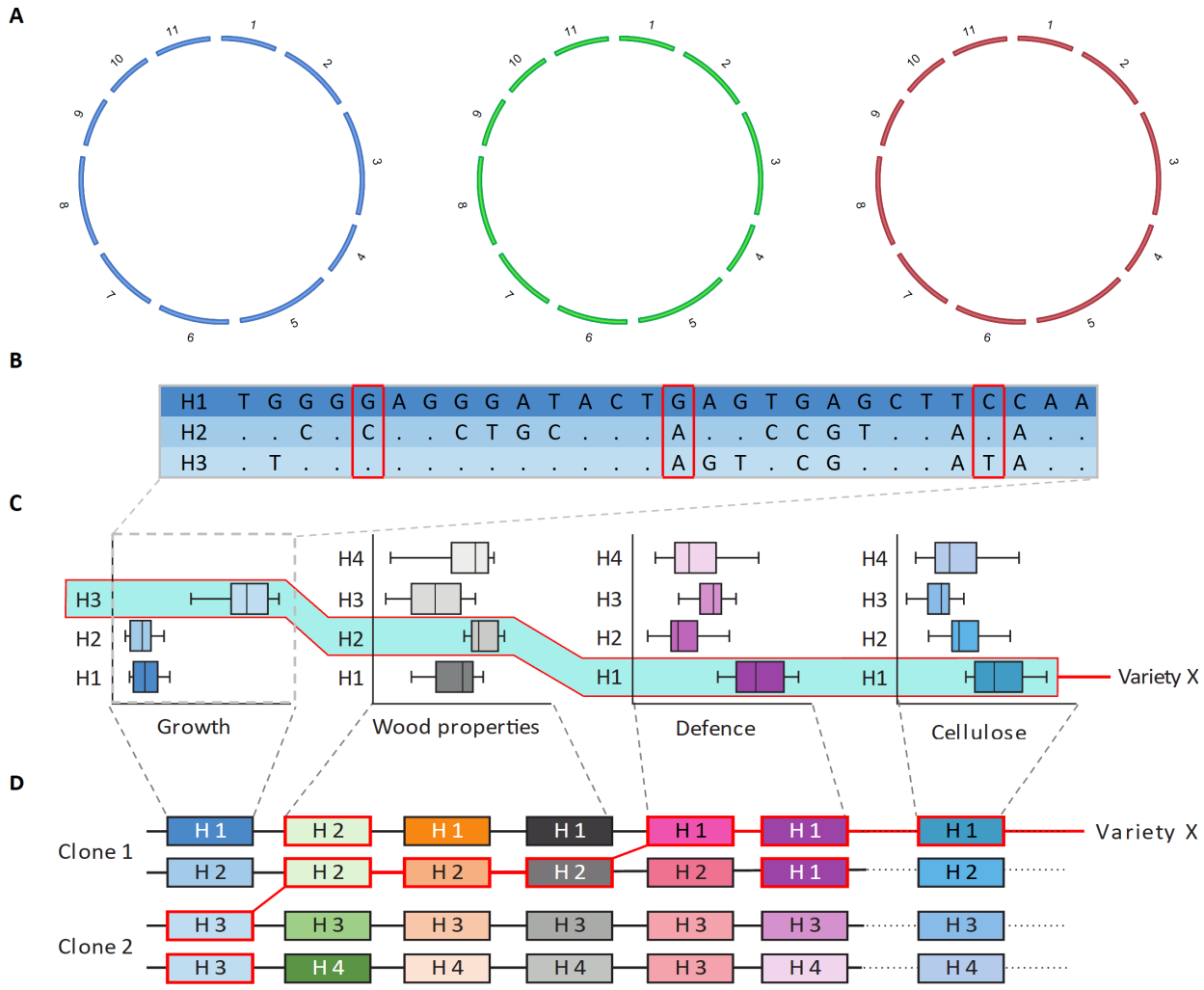


Figure 1.1 Model of how haplotypes can be used for crop improvement. (A) Using a species reference genome, along with genome sequencing data of multiple species a **(B)** haplotype diversity panel containing multiple haplotype variations (H1, H2 and H3) can be constructed. The haplotype diversity panel will be a representative pan-reference genome that can be used in combination with **(C)** phenotypic data to identify haplotype-phenotype associations. In addition, genomic structure, diversity, and functions of haplotypes can be established by re-sequencing of clones and analysis of quantitative trait loci. **(D)** Once haplotype-phenotype associations have been confirmed, desired haplotypes can be selected during breeding, by using markers specific for each clonal haplotype (red box in B) to produce new clones (variety X) that perform well for different traits of interest (modified from Bevan *et al.* 2017)**.**

1.10. References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. **2019**. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*: **20**: 224

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, *et al.* 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**: 145-161.e23.

Araus JL, Cairns JE. **2014**. Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science* **19**: 52–61.

Araus JL, Kefauver SC, Zaman-Allah M, Olsen MS, Cairns JE. **2018**. Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science* **23**: 451–466.

Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ. **2016**. A historical and evolutionary perspective on the biological significance of circulating DNA and extracellular vesicles. *Cellular and Molecular Life Sciences* **73**: 4355–4381.

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Brière C, Owens GL, Carrère S, Mayjonade B, *et al.* 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**: 148–152.

Bartholome J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion J-M. **2015**. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* **4**: 1283–1296.

Basantani MK, Gupta D, Mehrotra R, Mehrotra S, Vaish S, Singh A. **2017**. An update on bioinformatics resources for plant genomics research. *Current Plant Biology* **11**–**12**: 33–40.

Bauhus J, Pokorny B, Van der Meer P, Kanowski PJ, Kanninen M. **2010**. *Ecosystem goods and services from plantation forests*. Bauhus J, van der Meer PJ, Kanninen M, eds. Earthscan, London, UK. Washington DC: Earthscan, 205–227.

Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R, *et al.* 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* **4**: 879–887.

Bennett MD, Leitch IJ. **2011**. Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. *Annals of Botany* **107**: 467–590.

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, *et al.* 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**: 438–446.

Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD. **2017**. Genomic innovation for crop improvement. *Nature* **543**: 346–354.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, *et al.* 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**: 2047-217X.

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, *et al.* 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

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

Brooker MIH, Kleinig DA. **1983**. *Field guide to eucalypts: South-eastern Australia.* Melbourne: Inkata Press.

Browning S, Browning B. **2011**. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**: 703–714.

Byrne M. **2008**. Phylogeny, diversity and evolution of eucalypts. In: Sharma A, Sharma A, eds. *Plant genome: biodiversity and evolution, Vol. 1, Part E: Phanerogams - Angiosperm*. Enfield, NH, USA: Science Publishers, 303–346.

Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJM. **2017**. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications* **8**: 1–10.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, *et al.* 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, *et al.* 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608–611.

Chase MW, Clarke M, Grierson CS, Grierson D, Edwards KJ, Jellis GJ, Barnes SR, Chase MW, Clarke M, Grierson D, *et al.* 2011. One hundred important questions facing plant science research. *New Phytologist* **192**: 6–12.

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L. 2018. The sequenced Angiosperm genomes and genome databases. *Frontiers in Plant Science* **9**: 1–14.

Chen F, Song Y, Li X, Chen J, Mo L, Zhang X, Lin Z, Zhang L. 2019. Genome sequences of horticultural plants: past, present, and future. *Horticulture Research* **6**: 1-23.

Chen, Y., Nie, F., Xie, S. Q., Zheng, Y. F., Dai, Q., Bray, T., Wang, Y. X., Xing, J. F., Huang, Z. J., Wang, D. P., et al. 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* **12**:1–10.

Cherukuri Y, Janga SC. 2016. Benchmarking of *de novo* assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. *BMC Genomics* **17**: 95-105.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, *et al.* 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**: 1050–1054.

Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, *et al.* 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications* **8**: 1-13.

de Assis TF. 2000. Production and use of *Eucalyptus hybrids* for industrial purposes. In: *Hybrid breeding and genetics of forest trees*. 63–74.

de Lannoy C, de Ridder D, Risse J. 2017. A sequencer coming of age: *De novo* genome assembly using MinION reads. *F1000Research* **6**.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Dvorak WS, Hodge GR, Payn KG. 2008. The conservation and breeding of *Eucalyptus urophylla*: A case study to better protect important populations and improve productivity. *Southern Forests* **70**: 77–85.

Eldridge K, Davidson J, Harwood C. 1993. *Eucalypt domestication and breeding*. K Eldridge, Ed. Oxford, England: Clarendon Press.

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

Gaiotto FA, Bramucci M, Grattapaglia D. 1997. Estimation of outcrossing rate in a breeding population of *Eucalyptus urophylla* with dominant RAPD and AFLP markers. *Theoretical and Applied Genetics* **95**: 842–849.

Garg S, Martin M, Marschall T. 2016. Read-based phasing of related individuals. *Bioinformatics* **32**: i234–i242.

Garg S, Rautiainen M, Novak AM, Garrison E, Durbin R, Marschall T. 2018. A graph-based approach to diploid genome assembly. *Bioinformatics* **34**: i105–i114.

Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, Tischler G, Jackson DK, Keane TM, Li J, *et al.* 2017. *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports* **7**: 1–10.

Glenn TC. 2016. 2016 NGS Field Guide: Overview. *The Molecular Ecologist*. URL: http://www.molecularecologist.com/next-gen-fieldguide-2016/. [accessed 13 April 2019].

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies.

Nature Reviews Genetics **17**: 333–351.

Goulet BE, Roda F, Hopkins R. **2017**. Hybridization in plants: old ideas, new techniques. *Plant Physiology* **173**: 65–78.

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

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

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

Gunn B V, McDonald MW. **1991**. *Eucalyptus urophylla* seed collections. *Forest Genetic Resources Information* **19**: 34–37.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. **2013**. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Harwood C. **2011**. Introductions: Doing it right. In: Developing a eucalypt resource workshop proceedings. Wood technology research centre, 43–54.

Huang M, Tu J, Lu Z. **2017**. Recent advances in experimental whole genome haplotyping methods. *International Journal of Molecular Sciences* **18**: 1944.

Iglesias I, Wiltermann D. Eucalyptologics information resources on eucalypt cultivation worldwide. Available from http://git-forestry-blog.blogspot.com/2009/10/global-eucalyptus-map-2009-in-buenos.html. *GIT Forestry Consulting*.

Istace B, Friedrich A, D'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, *et al.* 2017. *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience* **6**: 1–13.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, *et al.* 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**: 338–345.

Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien FA, Swinkels W, Koelewijn A, Palstra AP, Pelster B, Spaink HP, *et al.* 2017. Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports* **7**: 1–13.

Jiao WB, Schneeberger K. **2020**. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications* **11**: 1–10.

Jiao WB, Schneeberger K. **2017**. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* **36**: 64–70.

Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P. **2019**. Tools and strategies for long-read sequencing and *de novo* assembly of plant genomes. *Trends in Plant Science* **24**: 700–724.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. **2013**. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27-38.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology.* **37**:540–546.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. **2018**. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**: 1174–1182.

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik M V. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* **9**: 1–15.

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

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, *et al.* 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology* **30**: 771–776.

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, *et al.* 2015. Genome sequence of cultivated upland cotton (*Gossypium hirsutum TM-1*) provides insights into genome evolution. *Nature Biotechnology* **33**: 524–530.

Li R, Tian X, Yang P, Fan Y, Li M, Zheng H, Wang X, Jiang Y. **2019**. Recovery of non-reference sequences missing from the human reference genome. *BMC genomics* **20**: 1-11.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, *et al.* 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. **2012**. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* **2012**: 1–11.

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, *et al.* 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications* **11**: 1-14.

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. **2012**. SNP Markers and their impact on plant breeding. *International Journal of Plant Genomics* **2012**: 1–11.

Marschall T, Marz M, Abeel T, Dijkstra L, Dutilh BE, Ghaffaari A, Kersey P, Kloosterman WP, Mäkinen V, Novak AM, *et al.* 2018. Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics* **19**: 118–135.

Martin B, Cossalter C. **1975**. The Eucalypts of the Sunda islands. *Bois et Forets des Tropiques* **163**: 3–25.

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, *et al.* 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**: 427–433.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. **2018**. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications* **9**: 1–8.

Michael TP, VanBuren R. **2015**. Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology* **24**: 71–81.

Mondal TK, Rawal HC, Gaikwad K, Sharma TR, Singh NK. **2017**. First *de novo* draft genome sequence of *Oryza coarctata*, the only halophytic species in the genus *Oryza*. *F1000Research* **6**.

Moran GF, Bell JC, Griffin AR. **1989**. Reduction in levels of inbreeding in a seed orchard of *Eucalyptus regnans* F. Muell. compared to with natural populations. *Silvae Genetica* **38**: 32–35.

Motazedi E, Finkers R, Maliepaard C, de Ridder D. **2017**. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in Bioinformatics*: 387-403.

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.

Neale DB, Wegrzyn JL, Stevens KA, Zimin A V., Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, *et al.* 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**: 1–13.

Ogawa D, Nonoue Y, Tsunematsu H, Kanno N, Yamamoto T, Yonemaru J. **2019**. Discovery of QTL alleles for grain shape in the Japan-MAGIC rice population using haplotype information. *G3 Genes|Genomes|Genetics* **8**: 3559–3565.

Ogawa D, Yamamoto E, Ohtani T, Kanno N, Tsunematsu H, Nonoue Y, Yano M, Yamamoto T, Yonemaru JI. **2018**. Haplotype-based allele mining in the Japan-MAGIC rice population. *Scientific Reports* **8**: 1–11.

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. **2018**. Genome size diversity and its impact on the evolution of land plants. *Genes* **9**: 88.

Phillippy AM, Schatz MC, Pop M. **2008**. Genome assembly forensics: Finding the elusive mis-assembly. *Genome Biology* **9**: 1-13.

Pryor LD, Williams ER, Gunn B V. **1995**. A morphometric analysis of *Eucalyptus urophylla* and related taxa with descriptions of two new species. *Australian Systematic Botany* **8**: 57–70.

Retief ECL, Stanger TK. **2009**. Genetic parameters of pure and hybrid populations of *Eucalyptus grandis* and *E. urophylla* and implications for hybrid breeding strategy. *Southern Forests* **71**: 133–140.

Rezende GDSP, de Resende MD V., de Assis TF. **2014**. *Eucalyptus* breeding for clonal forestry (T Fenning, Ed.). *Challenges and opportunities for the world's forests in the 21st century* **81**: 393–424.

Salman-Minkov A, Sabath N, Mayrose I. **2016**. Whole-genome duplication as a key factor in crop domestication. *Nature Plants* **2**: 1–4.

Saxena RK, Edwards D, Varshney RK. **2014**. Structural variations in plant genomes. *Briefings in Functional Genomics and Proteomics* **13**: 296-307.

Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, *et al.* 2017. *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *The Plant Cell* **29**: 2336-2348.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. **2018**. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* **19**: 329–346.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. **2018**. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **15**: 461–468.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*. **38**:1044–1053.

Shirasawa K, Esumi T, Hirakawa H, Tanaka H, Itai A, Ghelfi A, Nagasaki H, Isobe S. **2019**. Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (*Cerasus* × *yedoensis*). *DNA Research* **26**: 379–389.

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. *The New Phytologist* **206**: 1527–1540.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. **2015**. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, *et al.* 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants* **6**: 34-45.

Telfer EJ, Stovold GT, Li Y, Silva-Junior OB, Grattapaglia DG, Dungey HS. **2015**. Parentage reconstruction in *Eucalyptus nitens* using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness (C Chen, Ed.). *PLOS ONE* **10**: e0130601.

Vanburen R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, *et al.* 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**: 508–511.

Vignal A, Milan D, SanCristobal M, Eggen A. **2002**. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**: 275–305.

Vigneron P, Bouvet J-M, Gouma R, Saya AR, Gion J-M, Verhaegen D. **2000**. Eucalypt hybrids breeding in Congo. In: Dungey HS, Dieters MJ, Nikles DG, eds. Hybrid breeding and genetics of forest trees: Proceedings of the QFRI-CRS Symposium, 9-14 April 2000, Noosa, Queensland, Australia. Brisbane: Department of Primary Insdustries, 14–26.

Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R. **2020**. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. *GigaScience* **9**: giz160.

Wang W, Schalamun M, Morales-Suarez A, Kainer D, Schwessinger B, Lanfear R. **2018**. Assembly of chloroplast genomes with long- and short-read data: A comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC*

Genomics **19**: 1–15.

Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, Zhang J, Mainiero S, Strickler SR, Catala C, *et al.* 2020. Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nature Communications* **11**: 1-11.

Yang J, Moeinzadeh M-H, Kuhl H, Helmuth J, Xiao P, Haas S, Liu G, Zheng J, Sun Z, Fan W, *et al.* 2017. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants* **3**: 696–703.

Zamir D. **2013**. Where have all the crop phenotypes gone? *PLoS Biology* **11**: e1001595.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. **2019**. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**: 833-845.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, *et al.* 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology* **34**: 303–311.

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. **2013**. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, *et al.* 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196**: 875–890.

Chapter 2

Haplotype-resolved genome assembly of an F_1 hybrid of *Eucalyptus urophylla* x *E. grandis*

Anneri Lötter[1], Eshchar Mizrachi[1], Tuan A. Duong[1], Jill L. Wegrzyn[2], Alexander A. Myburg[1]

[1] *Department of Biochemistry, Genetics and Microbiology, Forestry and Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria 0028, South Africa*

[2] *Department of Ecology and Evolutionary Biology, Institute for Systems Genomics: Computational Biology Core, University of Connecticut, 67 N. Eagleville Road, Storrs, Connecticut, USA*

I performed all analyses in the manuscript and prepared the manuscript. Dr T.A. Duong and Prof J.L. Wegrzyn provided bioinformatic and technical support and along with Prof. A.A. Myburg advised on data analysis and interpretation throughout the project. Dr T.A. Duong, Prof. E. Mizrachi and Prof. J.L. Wegrzyn co-supervised the project. Prof. A.A. Myburg conceived and supervised the project.

2.1. Abstract

De novo haplotype phased genome assemblies based on long-read sequencing technologies have improved the detection and characterization of structural variants (SVs) in plant and animal genomes. As long-reads are able to span across haplotypes, they also allow phased (haplo) assemblies of highly heterozygous genomes such as those of forest trees. Knowledge of SV function and their resulting impact on gene expression can be used by breeders to guide tree improvement. *Eucalyptus* species and hybrids are some of the most widely planted hardwood trees. Hybrids are often preferred as they combine the genetic background of two species to produce more resilient trees that can inhabit a wider environmental deployment range. For example, *E. urophylla* x *E. grandis* hybrids combines disease resistance of *E. urophylla* with fast growth and desirable wood properties of *E. grandis*. However, to use such a strategy in eucalypt breeding firstly requires a high-quality reference genome (preferably phased) with which additional *de novo* assembled genomes can be compared. The aim of this study was to assemble high-quality haplotype phased genomes for *Eucalyptus urophylla* and *E. grandis*. Using Nanopore sequencing data generated for an *E. urophylla* x *E. grandis* F_1 hybrid and a trio-binning approach, we successfully assembled 544.51 Mb of the *E. urophylla* haplogenome (contig N_{50} of 1.93 Mb) and 566.75 Mb of the *E. grandis* haplogenome (contig N_{50} of 2.42 Mb) with a BUSCO completion score of 98.8%. Using high-density SNP genetic linkage maps of both parents, more than 88% of the haplogenome contigs could be anchored to one of the eleven chromosomes (scaffold N_{50} of 42.45 Mb and 43.82 Mb for the *E. urophylla* and *E. grandis* haplogenome assemblies, respectively). We also provide the first genome-wide comparison between the *E. urophylla* and *E. grandis* using the Synteny and Rearrangement Identifier (SyRI) to identify SVs, leading to the discovery of 48,729 SVs between the two haplogenomes. This study is the first step towards implementing haplotype-informed molecular breeding of *Eucalyptus* tree species.

2.2. Introduction

There is considerable pressure to improve crop yields to provide food, fibre, shelter and renewable energy for the growing human population (Chase *et al.*, 2011) in a sustainable manner. Fast-growing *Eucalyptus* tree species provide an important renewable feedstock for biomaterial (timber, fibre and lignocellulosics) and bioenergy production, relieving pressure on native forests (Grattapaglia & Kirst, 2008). These species, commonly referred to as eucalypts, constitute the most widely planted hardwood fibre crop globally. The most productive plantation areas are planted with interspecific F_1 hybrid clones that combine favourable characteristics of parental species and generally lead to increased forest productivity and product quality, and reduced production costs (de Assis, 2000; Grattapaglia & Kirst, 2008). The most commonly planted hybrid combination, *E. grandis* x *E. urophylla,* is primarily bred to merge the disease resistance of the tropical species *E. urophylla* with the fast growth of the subtropical *E. grandis*. However to further improve plantation productivity, wood quality and resilience, better breeding and deployment strategies are needed (Rezende *et al.*, 2014).

Our ability to develop accelerated breeding strategies for growth and climate resilience will play a crucial role in the sustainability of future plantation forestry. Current crop breeding strategies require many (often more than six) generations of backcrossing to introduce desirable allelic variation and remove undesired allelic variation in annual crops (Bevan *et al.*, 2017). As trees are outcrossed, suffer from inbreeding depression, have long breeding cycles and require large, expensive field trials, it is unfeasible to remove allelic variants using backcross breeding. By incorporating genomic resources and genome-wide molecular markers the breeding process can be sped up considerably and the cost associated with tree breeding can be reduced (Grattapaglia & Kirst, 2008). However, to prepare for the future, even more accurate and fast molecular breeding strategies are needed.

Haplotype-based molecular breeding has been shown to be a very accurate and effective breeding strategy (Ogawa *et al.*, 2018, 2019) compared to SNP based strategies. Discriminating the maternal and

paternal chromosome copies (defined as haplotypes or blocks of allelic variants that are inherited together by Zheng *et al.*, 2016) allows for identification of causal haplotype variants related to crop productivity and diseases resistance associated with different allelic/structural variants (Jiao & Schneeberger, 2017; Alonge *et al.*, 2020b). Parental haplotypes within the population can be identified and defined and, gene regions from 10,000s of genes from multiple individuals can be identified and sequenced to generate a set of genome-wide markers (e.g. single nucleotide polymorphisms or SNP tag-markers) defining particular haplotypes associated with a desired quantitative trait. By converting SNP data to haplotype data (based on two or more adjacent SNPs), quantitative trait locus (QTL) positions and effect sizes can be estimated more accurately (Ogawa *et al.*, 2018, 2019). Because the extent of genomic variation in the population is known (as all haplotypes of all the parents in the population is known), haplotypes can be inferred accurately for offspring, by using previously defined SNP tag-markers and imputing the rest of the haplotype with statistical methods (Motazedi *et al.*, 2017). These SNP tag-markers can then be used to aid the selection of the individuals to be used in further breeding or deployment.

Before haplotypes can be identified and defined a high-quality reference genome is needed. The golden standard of genome sequencing has been using short-read sequencing (SRS) platforms, due to their low cost (Michael & VanBuren, 2015; Kyriakidou *et al.*, 2018; Chen *et al.*, 2018). However, using short reads exclusively to assemble genomes may lead to shorter contigs and fragmented scaffolds that are usually not assembled up to chromosome scale (Cao *et al.*, 2017). As a result, the reference genomes of many species are an unrealistic representation of other individuals from the same species and represent a flat DNA sequence without variants between homologous chromosomes. Consequently, many of these reference sequences (in the case of outbred organisms) do not reflect levels of heterozygosity and the presence of unannotated or missing genes that differ between homologs due to pan-genome variation (Kyriakidou *et al.*, 2018). Long-read sequencing (LRS) technologies can mitigate the challenges associated with SRS-based plant genomes.

Currently, there are two LRS platforms available, of which Nanopore (ONT) sequencing offers many advantages including unlimited read length (Oxford Nanopore Technologies) and a lower cost than Pacific Biosciences sequencing (Glenn, 2016). As read lengths are longer, they can span across multiple homozygous regions and connect allelic variants between them, allowing us to sort and store multiple haplotype and structural variant alternatives in the assembly. This concept, with which one can store multiple genomes containing the spectrum of genomic variation, is referred to as a reference pan-genome, whereby variation would represent the dispensable component of the genome and the homozygous regions the core-genome. The growing number of assembled genomes, especially those assembled with LRS data, is making it clear to researchers that a single flat reference genome misses a substantial component of the genotypic and phenotypic diversity within a species (Sherman & Salzberg, 2020). As such, there is a movement towards assembly of a pan-reference genome, a concept that incorporates haplotype- and structural variants from multiple individuals in humans (reviews by Sherman & Salzberg, 2020) and plants (reviewed by Bayer *et al.*, 2020) into a single reference genome.

Studies on pan-genomic (including haplotype and structural) variation in *Eucalyptus* are limited, however several studies on genome synteny have been conducted. These genome synteny studies are based on genetic linkage maps constructed from a variety of molecular markers and have shown that there is high collinearity between the multiple different species, including *E. grandis* and *E. urophylla* (Brondani *et al.*, 1998; Marques *et al.*, 2002; Hudson *et al.*, 2012; Bartholome *et al.*, 2015). However, the degree of fine scale synteny between the *E. grandis* and *E. urophylla* is unknown as there is not a genome available for *E. urophylla*, one of the most important hybrid parent partners. Although three genomes have been published to date, two genomes, *E. grandis* (Myburg *et al.*, 2014) and *E. camaldulensis* (Hirakawa *et al.*, 2011) have been sequenced with a combination of Sanger and SRS. These sequencing technologies, have limited haplotype and structural variant identification capabilities (reviewed by Ho *et al.*, 2020). The third genome is that of *E. pauciflora,* which is not a species used in plantation forestry, and was assembled using a combination of SRS and LRS. As a result of the more

fragmented state of the other two species genomes and a lack of other LRS based genomes for *Eucalyptus*, studies regarding pan-genome variation for *Eucalyptus* is not possible.

Combining SRS and LRS data with a parent-offspring trio-sequencing approach has been demonstrated to allow assembly of high-quality haplo-reference genomes for the two parents, at a lower cost than generating two independent reference quality genomes (Koren *et al.*, 2018; Shirasawa *et al.*, 2019; Zhu *et al.*, 2019). Similarly, trio-sequencing of an interspecific F_1 hybrid of *E. grandis* and *E. urophylla,* paired with LRS technologies will generate high-quality assemblies of the haplogenomes contained in the F_1 hybrid. These high-quality phased genome assemblies will ultimately provide a basis for pursuing haplotype-based molecular breeding of eucalypt trees and will provide preliminary insights into the abundance and distribution of structural variants (SVs) of consequence to breeding. Thus, the aim of this study is to create a starting point for defining pan-genome, haplotype and structural variation in *Eucalyptus urophylla* and *E. grandis* parents used for hybrid breeding in South Africa.

2.3. Materials and Methods

2.3.1. Sample background

Leaf tissues of an F_1 *E. urophylla* x *E. grandis* hybrid offspring and its parents (*E. urophylla* seed parent and *E. grandis* pollen parent) were collected and used for DNA extractions. These individuals form part of a large nested association mapping trial and SNP data was used to generate high-density genetic linkage maps for both the *E. grandis* and *E. urophylla* parents (Candotti *et al.* unpublished). Sequencing both parents will enable a) inference of both haplotypes for the parental genomes and b) haplotype binning for genome phasing (Figure 2.1).

2.3.2. DNA isolation

Illumina sequencing

Genomic DNA was extracted from 50 mg of leaf tissue for the *E. urophylla* and *E. grandis* parents using the NucleoSpin® Plant II Kit (Machery-Nagel, Germany). Gel electrophoresis was performed using a 0.8% w/v agarose gel to assess DNA quality. DNA quality was also assessed using a NanoDrop® ND-1000 spectrophotometer (Thermo Fisher Scientific) and quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). Whole-genome sequencing of the F_1 hybrid and its parents was performed on an Illumina NovaSeq6000 platform by Macrogen (Macrogen Inc., Seoul, Korea).

High molecular weight DNA extraction

There is a trade-off between the read length and amount of data that can be obtained from a single ONT flow cell. To determine which DNA isolation method would yield the best combination of depth and read length for ONT PromethION sequencing, two DNA isolation methods were tested on MinION flow cells before PromethION sequencing. These methods were the 100/G Genomic-Tip (Qiagen) and a modified SDS based DNA extraction protocols. High molecular weight (HMW) DNA from the DNA isolation method yielding the best amount of data while still having a longer read length (close to 20 kb) was then sent for PromethION sequencing. The two DNA isolation methods are discussed below.

100/G Genomic-tip DNA extraction (Qiagen): Genomic DNA was extracted using 1.2 g of flash frozen ground leaf tissue. The ground material was suspended in 25 ml Guanidine buffer (20 mM EDTA, 100 mM NaCl, 1% Trition® X-100, 500 mM Guanidine-HCl and 10 mM Tris, pH 7.9), supplemented with 50 mg cellulase (Sigma-Aldrich) and 50 mg lysing enzyme (Sigma-Aldrich) incubated at 42 °C with gentle agitation. After 2.5 h, 10 µl RNase A (20 µg/ml) was added and the sample was incubated for 30 min at 37 °C, after which 50 mg proteinase K was added and the mixture was incubated for another 2 h at 50 °C. The mixture was then centrifuged for 20 min at 12 000 x *g* and the clarified lysate transferred to an appropriate buffer QBT-equilibrated Genomic-tip column (Qiagen), after which the column was

washed three times with 7 ml Buffer QC and HMW DNA was eluted with 5 ml Buffer QF. The DNA was precipitated by adding 0,7 V of isopropanol and centrifuged at 12 000 *g* for 20 min. The DNA pellet was washed twice with 70% Ethanol and resuspended in an appropriate volume of low salt TE (10 mM Tris-HCL pH 8.0; 0.1 mM of EDTA). Gel electrophoresis was performed using a 0.8% w/v agarose gel to assess DNA quality, and DNA quantity was assessed using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific).

SDS-based (Cornelissen and Ranketse, unpublished): Genomic DNA was extracted from 1 g of flash frozen ground leaf tissue. The ground leaf tissue was added to 10 ml of preheated lysis buffer (0.5 ml 1% SDS, 100 mM Tris-HCl, pH 8.0, 2.8 ml 1.4 M NaCl, 0.4 ml 20 mM EDTA, 2 ml 0.04% PVP, 0.05 ml Beta-Mercaptoethanol, 0.2 µg Proteinase K and ddH2O to a final volume of 10 ml) and incubated for 30 min with shaking at 55 °C. After incubation, the mixture was centrifuged for 30 min at 3000 *g* at room temperature (RT), after which 0.5 V of chloroform was added to the supernatant and gently mixed. The mixture was centrifuged again as before, after which 1 V of 24:1 chloroform: iso-amyl alcohol was added. The centrifugation step was repeated, 0.1 V of 5 M NaCl and 2.5 V of 100% ethanol was added and incubated overnight at -20 °C. The centrifugation step was repeated, the supernatant discarded, and the pellet washed twice with 70% ice cold ethanol. The mixture was centrifuged for 1 min at 12000 x *g* and the supernatant removed. The airdried pellet was resuspended in TE (0.1 m; 10 mM Tris-HCl, pH 8.0, 0.02 ml 1 mM EDTA, pH 8.0, 5 ul/350 ul RNAseA (10 mg/ml)/TE, and water to a final volume of 10 ml) by incubating for 15 min at 37 °C followed by 4 °C overnight. The process was repeated from the 1 V 24:1 chloroform: iso-amyl alcohol step twice, however with the last resuspension step, the pellet was resuspended in water and 2 µl of RNase A (20 µg/ml) was added. Gel electrophoresis was performed using a 0.8% w/v agarose gel to assess DNA quality, and DNA quantity was assessed using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific).

Nanopore sequencing

HMW DNA from both HMW DNA isolation methods were prepared for MinION sequencing following the manufacturers protocol using the genomic sequencing kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK). Approximately 3.3 µg of HMW DNA from was used without exogenous shearing or size selection. HMW DNA was first repaired with NEBNext FFPE Repair Mix (New England Biolabs) and 3'-adenylated with NEBNext Ultra II End Repair/dA-Tailing Module (NEB). The DNA was then purified with AMPure XP beads (Beckmann Coulter) and ligated with sequencing adapters (ONT) using NEBNext Quick T4 DNA Ligase (NEB). After purification with AMPure XP beads (Beckman Coulter), the library was mixed with sequencing buffer (ONT) and library loading beads (ONT) and loaded on primed MinION R9.4 SpotOn flow cells (FLO-MIN106). MinION sequencing was performed with a MinION Mk1B sequencer running for 48 h.

The resulting FAST5 files were base-called and reads with a QV < 7 were removed with Oxford Nanopore Technologies' Guppy base-calling software version 3.4.5 (ONT) using parameters for FLO-MIN106 and SQK-LSK109 library type. RStudio was used summarise and visualise statistics for both DNA isolation methods based on the sequencing summary file generated by the Guppy base-caller. The Guppy base-caller may not remove all of the sequence adapters so to ensure all sequence adapters are removed PoreChop version 0.2.4 (Wick, 2018) was used. All scripts used in this study is available online (https://gitlab.com/PlantGenomicsLab/eucalyptus-genome/-/tree/master/Final%20Thesis%20methods). The resulting adapter-less reads of both DNA isolation methods were combined into a single FASTQ file for further use.

HMW DNA from the DNA isolation method yielding the best amount of data while still having a longer read length (close to 20 kb) was used for PromethION sequencing. PromethION sequencing was performed by the Centre for Genome Innovation (University of Connecticut, Connecticut, USA) on a FLO-PRO002 PromethION flow cell as per the PromethION sequencing protocol (ONT) using the SQK-

LSK109 (ONT) sequencing kit with Circulomics Short Read Eliminator XS (Circulomics Inc.) size-selection. The flow cell was washed and reloaded after 38 h and run for an additional 6 h of sequencing. Base-calling was performed using the Guppy v3.4.5 basecaller and adapter removal was performed as stated above.

2.3.3. Genome assembly

Trio-binning and haplogenome assembly

Illumina short-reads were used for k-mer based genome size estimation was performed using Jellyfish v2.2.6 (Marçais & Kingsford, 2011) for 21-mers and visualised with GenomeScope v2.0 (Ranallo-Benavidez *et al.*, 2020). Long-reads of the F_1 hybrid were binned into *E. urophylla* and *E. grandis* haplotype bins (corresponding to parental short-reads) using the Trio-Canu module in Canu v1.8 (Koren *et al.*, 2017). Read contaminants were identified from the binned reads using Centrifuge v1.0.4-beta (Kim *et al.*, 2016) and removed with a custom script. Similarly, contaminant reads were also identified and removed from short read data with Kraken v2.0.8-beta (Wood *et al.*, 2019). The remaining raw reads were used for all assembly and alignment steps.

The binned reads corresponding to each of the parents were assembled separately, along with the corresponding parental short reads, using the MaSuRCA v3.3.4 (Zimin *et al.*, 2017) genome assembler. MaSuRCa was chosen as initial testing of multiple genome assemblers (based on the BUSCO completion score, contig N50 and total assembly size) indicated that the MaSuRCA genome assembler performed the best. Quality of the resulting assemblies was assessed using QUAST v5.0.2 (Gurevich *et al.*, 2013; Mikheenko *et al.*, 2018) and BUSCO v4.0.2 (Simão *et al.*, 2015; Seppey *et al.*, 2019). To verify genome coverage of the assemblies, Illumina reads from each of the parental haplotypes were mapped to the corresponding and alternative assembled haplogenomes using BWA v0.7.5a-r405 (Li & Durbin, 2009) and mapping rate calculated using the flagstat module from Samtools v1.9 (Li *et al.*, 2009).

Genome scaffolding

To improve assembly contiguity, scaffolding was performed for the MaSuRCa assembled *E. urophylla* and *E. grandis* genomes using high-density genetic linkage maps previously constructed for each of the parents (Candotti *et al*., unpublished). To resolve possible chimeric contigs that were assembled by MaSuRCa, Polar_Star (Phase Genomics, 2020) was used to infer breakpoints in contigs based on identification of read-depth outliers from the binned long-reads. After breakpoints were inferred, all contigs smaller than 3 kb were removed before scaffolding with high-density genetic linkage maps. A BLAST database was created for both assembled haplogenomes to identify the position of 1,588 *E. grandis* and 1,575 *E. urophylla* SNP probes used to construct the genetic maps. A consensus map was constructed with ALLMAPS (Tang *et al.*, 2015), consisting of SNPs that mapped to the assembled haplogenomes, and was used to perform genome scaffolding. For the consensus map construction, a weight of two was given to the parental genetic linkage map corresponding to species haplogenome to be scaffolded, while a weight of one was given for the alternative parental linkage map. Chromosome scaffold sizes from the two haplogenomes were compared to one another and to the *E. grandis* v2.0 genome to see whether there is a size difference between the *E. grandis* v2.0 reference potential bias in scaffolding of particular chromosomes. To validate if unplaced scaffolds were from a particular chromosome, unplaced scaffolds were aligned to the *E. grandis* v2.0 genome using MiniMap2 (Li, 2016) and alignments visualized with D-Genies (Cabanettes & Klopp, 2018).

2.3.4. Sequence based structural variant identification

To check for regions that are unassembled in the haplogenome assemblies compared to the *E. grandis* v2.0 reference genome, the *E. grandis* and *E. urophylla* haplogenomes were each aligned to the *E. grandis* v2.0 genome, with MiniMap2 (Li, 2016) and alignments visualised using D-Genies (Cabanettes & Klopp, 2018). Using the same method, the eleven assembled *E. grandis* and *E. urophylla* chromosomes were aligned to each other to visually identify genomic regions with possible large structural variants (SVs). To identify structural rearrangements (inversions, translocations and

duplications) and local variations (SNPs, InDels, copy gains/losses, highly diverged regions and tandem repeats) between *E. grandis* and *E. urophylla*, haplogenome assemblies were aligned to each other using nucmer from the MUMmer3 toolbox (Kurtz *et al.*, 2004) with alignment parameters "--maxmatch –c 100 -b 500 -l 50" . The resulting alignments were further filtered for alignment length (>100) and identity (>90). Identification of structural rearrangements and local variations was performed using the Synteny and Rearrangement Identifier (SyRI) pipeline (Goel *et al.*, 2019). The same method was also used to identify regions that have been altered between the *E. grandis* haplogenomes, and the *E. grandis* v2.0 reference genome. As the linear visualisation of syntenic regions and variants from SyRI prohibits us to see inter-chromosomal events, synteny and variants of greater than 10 kb were visualised with Circos (Krzywinski *et al.*, 2009).

2.3.5. Repeat element analysis

Custom libraries of repetitive elements were constructed for the *E. urophylla* and *E. grandis* haplogenomes with RepeatModeler v1.0.8 (Smit & Hubley, 2008). Repetitive elements were annotated with RepeatMasker v4.0.9 (Smit *et al.*, 2013) for the haplogenome assemblies. To eliminate the chance of missing repeat elements in either haplogenome due to not being identified, the combined species library was used as input for RepeatMasker. Lastly, to identify the abundance of LTR retrotransposons, LTR retrotransposon candidates were identified with LTR retriever (Ou & Jiang, 2018) for both haplogenomes and their distribution visualised with Circos.

2.4. Results

2.4.1. Illumina sequencing

To produce short-read data for genome size estimation and for trio-binning of the long-read data, we performed Illumina sequencing of the F_1 hybrid individual (SAP_F1_FK118) and its pure-species *E. grandis* (SAP_GRA_FK1758) and *E. urophylla* (SAP_URO_FK1756) parents (Sappi Forest Research, South Africa). This resulted in more than 116 Gb of PE150 data (Supplementary Table 2.1). Using

GenomeScope2.0, we estimated the genome size to be 443.19 Mb, 482.27 Mb and 477.76 Mb for the *E. urophylla, E. grandis* parents and the F_1 hybrid respectively (Supplementary Figure 2.1). This was substantially smaller than previously estimated based on flow cytometry (Grattapaglia & Bradshaw Jr., 1994) and reported for the *E. grandis* reference genome (Myburg et al. 2014), but a similar smaller estimate (408.16 Mb) has been reported for *E. pauciflora* based on short-read data (Wang *et al.*, 2020a). Levels of heterozygosity in the short-read data were 2.14%, 2.63% and 3.46% for the *E. grandis*, *E. urophylla* and F_1 hybrid based on GenomeScope2.0 analysis (Supplementary Figure 2.1). As expected, the estimated level of heterozygosity of the F_1 hybrid (3.46%) was higher than that of either parent. To further investigate the smaller than expected short-read estimates of genome size, we further explored the range of genome size estimates using short-read datasets for additional *E. grandis*, *E. urophylla* and hybrid individuals (see Supplementary Note 2.1). This analysis suggests that the genomes of these three trees are smaller than expected from published flow cytometry estimates, but within the range of short-read estimates for individuals and hybrids of the same species.

2.4.2. HMW DNA extraction and Nanopore sequencing

To find the best DNA extraction method for optimal ONT sequencing data, we compared two ONT sequencing libraries, named after the DNA extraction method used (referred to as the 100/G library and the SDS library). Of the two libraries, the 100/G library produced more data, resulting in 11.18 Gb of base-called sequence (in 2,169,209 reads), of which 9.3 Gb (75.85%) passed QC (Q-value > 7). For reads passing QC, the mean read length was 5.8 kb (read N_{50} of 18.96 kb) and the average read quality value was Q9.8. In comparison, for the SDS library, only 2.55 Gb (429,541 reads) of base-called sequence was generated, of which 2.28 Gb (84.92%) passed QC (Supplementary Table 2.2). The mean read length for reads passing QC was 6.4 kb (read N_{50} of 23.86 kb) and the average quality value was 10.3. In total, approximately 11.8 Gb of sequence data corresponding to approximately 17X coverage was obtained from the two MinION flow cells.

As DNA extracted with the 100/G tip method delivered more sequencing data based on MinION sequencing, we used DNA extracted with the 100/G tip for PromethION sequencing. This resulted in a total of 61.59 Gb of base called PromethION sequencing data was generated (read N_{50} of 28 kb), of which 56.57 Gb (91.85%) passed QC. Thus, a total 68.15 Gb of Nanopore sequencing data was generated for use in trio-binning corresponding to approximately ~105X coverage of the F_1 hybrid genome and ~50X coverage per haplogenome in the hybrid (Supplementary Table 2.2).

2.4.3. Genome assembly

Phased hybrid genome assembly using trio-binning

To separately assemble the long reads originating from the two haplogenomes in the F_1 hybrid, we first performed trio-binning with Canu using the Illumina short-read data for the parents and the long-read data for the F_1 individual. We were able to bin 1,876,816 reads (32.66 Gb) for the *E. urophylla* haplogenome and 1,998,860 reads (35.11 Gb) for the *E. grandis* haplogenome corresponding to 50X and 54X coverage of the two haplotypes, respectively (Figure 2.1, Supplementary Table 2.3). Only 6,693 reads could not be binned. We excluded these from further analysis as they made up less than 10 Mb (0.014%, much less than the 5% cut-off recommended by Koren *et al.*, 2018) of the total amount of reads.

Assembly of the binned reads for the *E. urophylla* haplogenome resulted in 654 contigs and a total size of 546.1 Mb, with a contig N_{50} size of 4.41 Mb and L_{50} of 36 (Table 2.1). A BUSCO completeness score of 99.2% was obtained using the embryophyte dataset (for 1,614 BUSCO groups tested), of which 95.2% were single-copy genes and only 4.0% were duplicate-copy genes (Supplementary Figure 2.2). We assembled the reads binned for the *E. grandis* haplogenome into 793 contigs with a total size of 568.5 Mb, with a contig N_{50} size of 3.91 Mb and L_{50} of 38 (Table 2.1). For this assembly we obtained a BUSCO completeness score of 99.0%, of which 94.4% is single copy and 4.6% duplicate (Supplementary Figure

2.2). The low duplicate-copy percentage for both assemblies confirm the efficiency of trio-binning to separate the long reads into haplogenome bins.

Next, we investigated whether the difference between the smaller genome assembly size may be as a result of reads that were excluded from the assembly by mapping parental Illumina reads to the corresponding parental haplogenome. Mapping the parental *E. urophylla* and *E. grandis* Illumina reads to the corresponding assembled haplogenomes resulted in mapping rates of 98.73% and 99.10% (93.79% and 92.91% properly paired) respectively (Supplementary Table 2.3), indicating that the smaller than expected genome assembly sizes were not due to exclusion of reads in the genome assembly process. We also mapped the alternative parental reads to the haplogenomes and found mapping rates of 98.11% and 97.67% (85.03% and 84.85% properly paired) for the *E. urophylla* and *E. grandis* haplogenomes respectively. This slightly lower mapping rate was expected as certain reads are not present in the alternative parental haplogenome due to species specific genomic variation.

Genome scaffolding

To curate incorrectly assembled contigs, contig breakpoints were inferred based on long-read depth support. An additional 764 breakpoints were inferred for the *E. urophylla* haplogenome assembly and 785 breakpoints were inferred for the *E. grandis* haplogenome assembly retaining 544.5 Mb and 566.7 Mb, respectively, after removal of all contigs of less than 3 kb. This resulted in lowered assembly contiguity for both haplogenome assemblies, while retaining the high BUSCO completeness scores (Table 2.1). To improve genome contiguity, parental SNP genetic linkage maps (Candotti *et al*., unpublished) were used to anchor contigs to linkage groups. The parental genetic linkage maps yielded a set of 3,125 (for the *E. urophylla* haplogenome) and 3,129 (*E. grandis* haplogenome) unique SNP markers to anchor contigs into pseudo-chromosome level scaffolds. The anchoring rate for both haplogenome assemblies was greater than 88.0% (Table 2.2) and a BUSCO completeness score of at least 96.3% was obtained for contigs anchored to one of the eleven chromosomes. At least two markers

are required per contig for ALLMAPS to be able to orientate a contig, which was the case for 299 *E. urophylla* and 261 *E. grandis* contigs. Contigs that only have one marker, are placed, but the orientation is unknown. There were 52 such contigs for *E. urophylla* and 49 for *E. grandis* (Table 2.2). A total of 1,067 contigs (corresponding to 63.37 Mb) of the *E. urophylla* and 1,268 contigs (67.78 Mb) of the *E. grandis* haplogenome assembly could not be anchored (Table 2.2) of which 863 (9.71 Mb) and 1,051 contigs (11.86 Mb) were smaller than 50 kb (Supplementary Table 2.4). As these contigs are small, most of them contain no markers (only 18 *E. urophylla* and 26 *E. grandis* contigs contain markers, average of one marker every 2 Mb) and cannot be anchored onto a particular chromosome. The lack of markers within them could be due to the manner in which we selected SNP markers for the parental maps and may have some properties in common.

Most of the anchored assembly had a high level of congruence between the genetic and physical maps as indicated by the Pearson's correlation coefficient (ρ) being close to -1 or 1, with the weakest correlation being $\rho = 0.965$ (Supplementary Figure 2.3) for *E. urophylla* and $\rho = 0.938$ for *E. grandis* (Supplementary Figure 2.4). We observed some genomic regions with gaps in the marker positions of the genetic linkage map for one of the parents (i.e., there are no markers in the assembly for that region). For example, on Chromosome 3 of the *E. urophylla* haplogenome there was a large region with no corresponding SNP markers in the *E. grandis* parental map, but many SNP markers in the *E. urophylla* linkage map (Supplementary Figure 2.4). In addition, for *E. grandis*, there was one region on Chromosome 6 that had SNP markers mapping to linkage group 5 (LG5) of both parental maps. We inspected this region by mapping raw long reads to the conflicting contig, which revealed that was in fact a misassembled contig. We subsequently split the conflicted contig by inferring a breakpoint based on a MUMmer3 (Kurtz *et al.*, 2004) alignment to the *E. grandis* v2.0 reference assembly before re-scaffolding of all contigs with ALLMAPS as described previously (Supplementary Figure 2.5, Supplementary Figure 2.6, named "*E. grandis* corrected" in Table 2.2 which was used for all further analyses). When comparing chromosome sizes, Chromosome 3 and 5 differed from the reference *E.*

grandis v2.0 genome by more than 20 Mb (Supplementary Figure 2.7). To investigate this, we aligned unplaced scaffolds to the *E. grandis* v2.0 reference genome. Dot-plot alignments of unplaced scaffolds to the *E. grandis* v2.0 reference genome did not reveal any chromosomal preference for unplaced scaffolds. Rather, unplaced scaffolds were distributed throughout the genome, with some aligning to multiple chromosomes (Supplementary Figure 2.8). This suggest that the chromosomal size difference is not due to unplaced scaffolds not being placed onto their respective chromosomes.

2.4.4. Identification of structural variants

E. grandis and *E. urophylla* are in the same section (*Latoangulatae*) and subgenus *Symphyomyrtus* but have non-overlapping natural ranges with unique adaptations such as greater resistance to fungal pathogens in *E. urophylla,* which has a more tropical distribution. Genetic linkage mapping has suggested high collinearity of their genomes (Hudson *et al.*, 2012; Kullan *et al.*, 2012; Bartholome *et al.*, 2015), but a direct fine-scale comparison of genome synteny has not been possible to date. We investigated the synteny of the two assembled haplogenomes using the whole-genome comparison tool SyRI, to identify structural rearrangements and other local sequence differences. SyRI works in a hierarchical manner, firstly identifying syntenic regions, then structural rearrangements and lastly genome sequence divergence in colinear regions (syntenic and rearranged regions that align to each other).

A total of 318 Mb was syntenic between the two assemblies, while 386.6 Mb and 213.5 Mb were identified as rearranged between the *E. grandis* v2.0 and *E. grandis* haplogenome, respectively (Supplementary Table 2.5 and Supplementary Figure 2.10). In comparison, 257 Mb was syntenic (Figure 2.3, Figure 2.4 and Figure 2.5A), while 262.2 and 374.9 Mb were identified as rearranged (ranging in size from 100 bp to 4.91 Mb in size, Figure 2.5B) in the *E. grandis* and *E. urophylla* haplogenomes, respectively (Supplementary Table 2.5, Figure 2.5A and Figure 2.5B). As expected, there was greater synteny between the two *E. grandis* assemblies than between the *E. grandis* and *E. urophylla*

haplogenomes, however due to the difference in assembly size this does not translate to genomic proportion. The rearranged regions included 167 and 189 inversions and 9,260 and 10,526 translocations for the *E. grandis* v2.0 vs *E. grandis* haplogenome and *E. grandis* haplogenome vs *E. urophylla* haplogenome comparisons (Figure 2.5A and B, Supplementary Figure 2.9, Supplementary Figure 2.10, Supplementary Table 2.5 and Supplementary Table 2.7). In addition, there were 29,596 duplications in the *E. grandis* v2.0 and 17,519 duplications in the *E. grandis* haplogenome, compared to 16,865 duplications in *E. grandis* and 21,149 duplications in *E. urophylla* (Supplementary Table 2.5 and Supplementary Table 2.8). Together these results suggest that although there is high collinearity between the *E. grandis* and *E. urophylla* haplogenomes, finer scale synteny is lower than previously suggested.

Next, we investigated genome sequence divergence in colinear regions, named local variants, which made up 56.5 and 69.5 Mb the across all comparisons. The size of local variants between *E. grandis* and *E. urophylla* haplogenomes (excluding SNPs) ranged from 1 bp to 3.09 Mb (Figure 2.5C). In both comparisons, SNPs were the most prevalent class of local variants in terms of number, with 8.3 million SNPs between the *E. grandis* and *E. urophylla* haplogenomes and 6.3 million between the *E. grandis* v2.0 reference and the *E. grandis* haplogenome, followed by insertions and deletions (Supplementary Table 2.6). However, in terms of the total bases affected, highly diverged regions and copy gain/losses had the greatest impact, as they made up 9.5 Mb and 38 – 40 Mb of the haplogenome assemblies and 11 Mb and 31 – 45 Mb of the *E. grandis* v2.0 and *E. grandis* haplogenome assemblies. Although there is a greater number of local variants compared to SV, local variants made up 13.8% of the *E. urophylla* and 13.1% of the *E. grandis* chromosomal assembly compared to 54.5% and 75.1% in SV. This suggests that although local variants are more numerous, structural variants have a larger impact. This was also confirmed in previous studies in tomato (Alonge *et al.*, 2020a).

2.4.5. Annotation of repeat elements

To further examine whether the smaller haplogenome assembly size is due to a difference in repeat content, we annotated repeat elements with RepeatMasker. A total of 48.34% of the *E. urophylla* haplogenome assembly comprised of repetitive elements, whereas it was 49.09% for the *E. grandis* haplogenome (Table 2.3). In both cases, LTR retrotransposons were the most prevalent repetitive element, making up more than 21.06% of the assembled haplogenomes (Table 2.3). DNA transposons made up ~6% of the haplogenomes. These results are similar to previous repeat annotations for *Eucalyptus* (Myburg *et al.*, 2014). To characterize and visualize the distribution of various LTR retrotransposons (in bins of 300 kb), we used LTR retriever, which is more sensitive for detection of LTR retrotransposons. We found that the total percentage of LTR retrotransposons is greater according to LTR retriever, with 29.08% and 29.25% of the *E. grandis* (Supplementary File 1) and *E. urophylla* haplogenomes (Supplementary File 2) detection of retrotransposons with LTR retriever has also been seen previously (Wang *et al.*, 2020a). The increased percentage of LTR retrotransposons identified by LTR retriever suggests that some LTR elements may not be identified by RepeatModeler or may have been identified but not classified as LTR elements, but rather as unknown and future studies should rather incorporate a combined library as input for RepeatMasker. Unfortunately, direct comparison of the LTR retrotransposon distribution pattern between *E. grandis* and *E. urophylla* is not possible as the assembled chromosomes differ in size, but a general overview can be seen in Figure 2.3.

2.5. Discussion

We have assessed the use of a trio-binning read separation strategy to assemble high-quality haplogenomes for two important eucalypt tree species as a starting point towards investigating pan-genome variation within and between these species. The high level of heterozygosity in the F_1 hybrid genome enabled discrimination of almost all parental long reads and independent assembly of the parental haplogenomes present in the F_1 hybrid. These haploid assemblies are the first of their kind for any forest tree and allowed us to circumvent the problem of co-assembly of alternative haplotypes which

has presented a challenge for the assembly of highly heterozygous tree genomes, especially in intergenic DNA where complex structural variation from partially overlapping haplotypes may be co-assembled (Myburg *et al.*, 2014; Bartholome *et al.*, 2015). Furthermore, the high coverage of long reads (50X per haplogenome) allowed us to assemble across complex repeat structures leading overall to highly contiguous assemblies (contig N_{50} of 2.4 Mb for *E. grandis* and 1.9 Mb for *E. urophylla*). To improve accuracy of the assembled contigs in the two long-read assemblies, contigs were verified based on long-read coverage support and misassembled contigs broken if there was low read-depth support. Intriguingly, we find that, despite having very high BUSCO completeness scores (>98%), the assembled haplogenomes were substantially smaller than previous diploid reference genome assembly of 691.4 Mb (Myburg *et al.*, 2014; Bartholome *et al.*, 2015) and the ~640 Mb genome size estimates based on flow cytometry (Grattapaglia & Bradshaw Jr., 1994). We used high-density SNP genetic linkage maps to further improve haplogenome assembly contiguity by scaffolding contigs onto chromosomal linkage groups. Finally, we performed a genome-wide structural comparison of the *E. grandis* and *E. urophylla* haplogenomes, the first direct, fine structure comparison for any two eucalypt genomes, and show that SVs are more prevalent than detected in previous studies but follow a similar class distribution pattern than previously observed, where inversion events are the least frequent, followed by translocation events and duplications are the most frequent (Goel *et al.*, 2019; Jiao & Schneeberger, 2020).

2.5.1. Trio-binning of a highly heterozygous F1 hybrid genome

To assemble the separate (phased) haplogenomes that make up the genome of the F_1 hybrid individual, we used the trio-binning strategy described by Koren *et al.* (2018) to separate the long-reads derived from the F_1 hybrid into *E. urophylla* and *E. grandis* haplotype bins before genome assembly (Figure 2.1). This approach allowed successful binning of the *E. urophylla* and *E. grandis* haplogenome derived long reads. A total of 99.98% of the sequenced read bases could be assigned to one of the two parental haplo-bins, with only a small proportion (0.014%) of mostly shorter nanopore reads not assigned to bins (N_{50} = 1,385 bp vs N_{50} ~ 27.5 kb for binned reads). The long-read data was split almost evenly per

haplotype (51.80% and 48.18% of long read bases assigned to the *E. grandis* and *E. urophylla* haplotypes respectively, Supplementary Table 2.3), as one would expect in a diploid organism where the two haplogenomes are similar in size. Furthermore, we performed cross-mapping of the parental short-read data to the two haplogenomes and found, as expected, lower mapping rates to the opposite haplogenome (average 93.35% vs 84.94%, individual rates shown in Supplementary Table 2.3) supporting the expectation that we have efficiently separated the haplogenome reads from the two species.

Using the binned long reads, we assembled 544.1 Mb of the *E. urophylla* haplogenome and 566.7 Mb of the *E. grandis* haplogenome (contig N_{50} of 1.9 Mb and 2.4 Mb, respectively) with BUSCO completion scores of greater than 98.7% (Table 2.1). The low level of BUSCO duplication in the assembled haplogenomes, less than 3.8% (Supplementary Figure 2.2) compared to 13.9% observed previously after haplotig removal for the recent diploid *E. pauciflora* assembly (Wang *et al.*, 2020a), further supports our conclusion that the haplotype binning was highly efficient. We further validated the size of phased blocks, as well as phase origin (Supplementary Note 2.2) and found the haplogenome assemblies had very low haplotype switch error rates (lower than 0.033%) confirming the accuracy of haplotype separation. Together these results suggest that the trio-binning approach was highly efficient and accurate in the heterozygous F_1 hybrid genome.

Haplotype separation is known to improve with higher levels of heterozygosity (Koren *et al.*, 2018; Rhie *et al.*, 2020). We observed high heterozygosity for both pure-species parents (2.14% for *E. grandis* and 2.63% for *E. urophylla*), and as expected, heterozygosity was substantially higher in the F_1 hybrid offspring (estimated to be 3.46%; Supplementary Figure 2.1). Such high heterozygosity levels are expected for outcrossed organisms such as eucalypts (Moran *et al.*, 1989; Gaiotto *et al.*, 1997). Successful haplotype separation of an F_1 hybrid of species within the same section of Myrtaceae (*Latoangulatae*) suggests that application of trio-binning for haplotype separation should be successful for other intrasectional and intersectional *Eucalyptus* F_1 hybrid combinations. In addition, the high

heterozygosity observed in the pure species parents suggests that haplotype binning will also be successful in intraspecific crosses of *Eucalyptus* as the trio-binning strategy has been demonstrated to be efficient at much lower levels of heterozygosity (0.9% in the case of a F_1 Brahman x Angus cattle hybrid and 1.36% for *A. thaliana*, Koren *et al.*, 2018).

We note that the haplogenome assembly sizes, 546/481 Mb for *E. urophylla* and 568/498 Mb for *E. grandis* (total/scaffolded size) were much smaller than that of the current *E. grandis* v2.0 reference genome (691/612 Mb, Myburg *et al.*, 2014; Bartholome *et al.*, 2015) and previous estimates (~ 640 Mb) based on flow cytometry (Grattapaglia & Bradshaw Jr., 1994). K-mer based genome size estimates of the parental reads predicted diploid genome sizes of 443 Mb for *E. urophylla,* 482 Mb for *E. grandis* and 478 Mb for the F_1 hybrid (Supplementary Figure 2.1), which agreed with the scaffolded genome sizes of the two haplogenome assemblies. This apparent discrepancy was also observed in *E. pauciflora*, where k-mer based estimates were 408 Mb compared to the final 595 Mb assembly (Wang *et al.*, 2020a). Further exploration of k-mer based genome size estimates showed that our genome size estimates fall within that expected for *E. grandis* (Supplementary Note 2.1) and that this discrepancy is observed for multiple previously sequenced individuals and not unique to the sequencing data used in this study. The total assembly sizes of the two haplogenomes were therefore approximately 70 - 100 Mb smaller than previous flow cytometry estimates for the two species and the total scaffolded sizes were 140 - 160 Mbp smaller than expected. This size discrepancy may be explained by several factors, which we explore below.

First, to exclude the possibility that the smaller assembly size was due to a portion of sequencing reads not being assembled, i.e. that we failed to assemble parts of the haplogenomes, we aligned the parental Illumina reads to the corresponding parental haplogenome assembly. We also aligned the raw short- and long-reads and the haplogenome assemblies to the *E. grandis* v2.0 reference genome to make sure all v2.0 genomic regions had sequencing coverage (Supplementary Note 2.3). We noted that some regions

had very high sequencing depth when aligning reads to the *E. grandis* v2.0 reference genome and explored this further in Supplementary Note 2.3. More than 98.7% of parental Illumina reads aligned to their corresponding parental haplogenome, which suggests that almost all of the sequences in the parental genomes (that are amenable to Illumina sequencing) are represented in the haplogenomes (Supplementary Table 2.2), although it is possible that these may in some cases map to highly repetitive regions that are collapsed in the haplogenome assemblies. To further investigate this possibility, we confirmed that the repeat content of the haplogenomes were not lower than that reported in the *E. grandis* v2.0 reference assembly. In fact, the repeat content for the *E. urophylla* and *E. grandis* haplogenomes (48.16% and 48.91%, respectively) was higher than that reported for the *E. grandis* v2.0 assembly (44.50%) and for the more recent *E. pauciflora* assembly (44.77%, Table 2.3) (Myburg *et al.*, 2014; Wang *et al.*, 2020a). This suggests that the observed size difference is most probably not due to the collapse of repetitive regions during haplogenome assembly. Rather, the slightly higher repeat content of our haplogenome assemblies probably reflect our ability to better assemble across such repeats using long-read technology in haplo-assemblies vs short-read/Sanger sequencing derived from highly heterozygous genomes. Although derived from a partially inbred individual (S1), it is possible that the *E. grandis* v2.0 reference assembly is somewhat inflated in size due to the possible co-assembly of partially overlapping alternative haplotypes in the highly heterozygous regions of the genome. Accordingly, our analysis showed that Chromosomes 3 and 5 in the haplogenome assemblies were 20 Mb smaller than the corresponding chromosomes in the diploid *E. grandis* v2.0 assembly.

2.5.2. Genetic linkage maps support high scaffolding rates

We used high-density SNP genetic linkage maps of the parents to order and orient scaffolds from the draft haplogenome assemblies of *E. urophylla* and *E. grandis* with ALLMAPS, using a greater weight for the genetic linkage map of the parent from which the haplogenome originated. Overall, 88.4% and 88.0% of the assembly was anchored into 11 pseudo-chromosomes for *E. urophylla* and *E. grandis*, corresponding to the haploid chromosome number. A similar percentage of the haplogenomes could be

anchored using both parental genetic linkage maps as has been found in other plant species where a range of 69.7% to 98.8% of contigs could be ordered and orientated with genetic linkage maps only (Raymond *et al.*, 2018; Morrissey *et al.*, 2019; Li *et al.*, 2020; Langdon *et al.*, 2020). The high percentage of anchored bases is in part due to the level of contiguity of the haplogenome assemblies before scaffolding (N_{50} of 1.9 Mb and 2.4 Mb), as well as the high density of SNP markers (averaging more than 6.3 markers per Mb) used for anchoring (Table 2.2). The fact that the genetic linkage maps were from the exact same parents from which the two haplogenomes were derived, also would have contributed to higher anchoring rates.

There are some limits to using ALLMAPS for genome scaffolding as the program cannot identify and separate duplicated regions that are misassembled or collapsed by the genome assembler due to high similarity (Tang *et al.*, 2015). We found one such misassembly by MaSuRCa on Chromosome 6 of the *E. grandis* haplogenome assembly, where multiple SNP markers mapped to linkage group 5 (LG5) of both parental maps (Supplementary Figure 2.5 and Supplementary Figure 2.6). By aligning raw long reads to the region, we could infer the breakpoint in the misassembled contig, resulting in a 3 Mb contig that was subsequently correctly anchored to Chromosome 5 (Table 2.2, Supplementary Figure 2.5 and Supplementary Figure 2.6).

In addition, most genetic linkage maps contain regions such as centromeres with no or very low recombination and few DNA markers for anchoring and orientation of contigs. Thus, integration of additional proximity ligation or optical mapping data may lead to inclusion of some of the remaining unplaced contigs that had few markers to place or orient them (average 0.4 and 0.5 markers per Mb for unanchored vs 6.5 and 6.3 markers per Mb for anchored *E. urophylla* and *E. grandis* contigs, respectively, Table 2.2). Many of the unanchored contigs may contain difficult to assemble, centromeric or other non-recombinogenic regions devoid of mapped DNA markers. The N_{50} of the unanchored contigs was 324 kb which was smaller than the average marker spacing in those regions (Supplementary

Table 2.4). Improved anchoring of contigs using optical mapping or proximity ligation approaches should result in more accurate assembly of these complex genomic regions than can be achieved through the use of genetic linkage maps alone. However, despite the limitations of only using genetic linkage maps for contig placement, we were able to produce eleven pseudo-chromosome scaffolds for each of the haplogenomes owing to the high density of SNP markers in the parental maps and the quality of the genetic maps as evidenced in collinearity of markers between the genetic map and the de novo assembled contigs, as well as high collinearity between the scaffolded assembly and the genetic linkage maps (Pearson's correlation of $\rho = 0.938$ to $\rho = 1.00$; Supplementary Figure 2.4, Supplementary Figure 2.5 and Supplementary Figure 2.6).

2.5.3. Structural variants between *E. urophylla* and *E. grandis*

To our knowledge, this is the first genome-wide comparison of synteny and structural rearrangements between two eucalypt species. In addition, we had the advantage of being able to directly compare the two haplogenomes from the same F_1 hybrid individual assembled using the same method. Using SyRI we found that 53.39% (256.9 Mb) of the 481.16 Mb chromosomal assembly of *E. urophylla* and 51.45% (256.7 Mb) of the 498.97 Mb chromosomal assembly of *E. grandis* was syntenic (Supplementary Table 2.5). We were able to identify 48,729 SVs between the two haplogenomes, with a 103.62 Mb difference between the two haplogenomes due to duplications (Supplementary Table 2.5). This seems to be an artifact of the chosen reference as using *E. urophylla* haplogenome as reference resulted in an increase in duplications for *E. urophylla*. As seen in previous studies using SyRI for SV calling, we found that inversions were the smallest group of SVs in terms of number, followed by translocations, with duplications being the most abundant (189 inversions, 10,526 translocations and 38,014 duplications, Supplementary Table 2.5; Goel *et al.*, 2019; Jiao & Schneeberger, 2020). A previous study by Zhou *et al.*, (2019) identified SVs between two Chardonnay haplotypes, Chardonnay and Cabernet Sauvignon (Cab08) as well as a variety of grapevine cultivars and also found a lower number of inversions and translocations compared to duplications. To identify SV between Chardonnay haplotypes of the FPS 04

clone, they assembled a haplotype resolved *de novo* primary assembly for Chardonnay (Char04) and mapped all long-read sequence data generated they generated to Char04 (Zhou *et al.*, 2019). SV between Char04 and Cab08 haplotypes were identified and verified using three methods of SV detection: 1) alignment of long-read sequencing data of Cab08 to Char04, 2) whole-genome alignment of the Char04 and Cab08 assemblies and, 3) alignment of Illumina short-read sequencing data from Cab08 to Char04 (Zhou *et al.*, 2019). Only 62% of SV detected by whole-genome alignment and long-read alignment methods could be detected based on short-read alignment, which confirms the limited ability of short-read alignment methods for SV detection. As a result, when SVs were identified for 50 grapevine cultivars and 19 wild relatives through short-read alignment SV were limited to those confirmed through short- and long-reads between Char04 and Cab08. Using unfolded site frequency spectrum of the SV, Zhou *et al.*, (2019) found that there is purifying selection against SVs, and that there is stronger purifying selection against inversions and translocations compared to duplications as they have a more deleterious effect compared to duplications (Zhou *et al.*, 2019). Stronger purifying selection against inversions and translocations in our haplogenome assemblies may therefore explain the lower frequency of these two classes of SV, however this will need to be tested in future sequencing projects including population-wide tracking of SVs.

With additional genome sequences for *E. grandis* and *E. urophylla*, a pan-genome of genomic (structural and local) variants can be reconstructed as was done for *Arabidopsis* (Jiao & Schneeberger, 2020) and tomato (Wang *et al.*, 2020b; Alonge *et al.*, 2020b). Although there are multiple different tools and manners in which to identify SV, we made use of the whole genome alignment tool SyRI. SyRI identifies SVs and local variants using three main steps: 1) identify syntenic alignments, 2) identify inverted, duplicated and translocated alignments and 3) identify "local variants" within alignment blocks. As such, there is a hierarchy of variation where local variants are found within alignment blocks, be they syntenic or rearranged regions. However, when looking for the functional effects of local and larger structural variants, it is important to note the hierarchy of genomic rearrangements, as local variants within

rearranged regions show different inheritance patterns to those in syntenic regions. SVs can influence recombination studies as rearrangement hotspots have lower synteny and reduced recombination rates (Jiao & Schneeberger, 2020). In addition, SVs can influence gene expression directly or indirectly, for example duplications versus epistatic interactions such as proximity of the promotor to the gene (Alonge *et al.*, 2020b) which makes their functional interpretation harder (Goel *et al.*, 2019).

2.6. Conclusions and future perspectives

To improve crop yield, a clear understanding of how genomic and environmental factors interact to produce desired phenotypes is required. The first step towards understanding these interactions is to identify the genetic variation present within the crop and understand how these variants contribute to trait variation. As a prelude to haplotype-based molecular breeding in *Eucalyptus*, we successfully applied a trio-binning approach to assemble approximately 545 Mb of the *E. urophylla* haplogenome and 567 Mb of the *E. grandis* haplogenome contained in an F_1 *E. urophylla* x *E. grandis* hybrid and obtained two high-quality genomes, each with a BUSCO completeness of greater than 98.8% and a chromosome scaffold N_{50} of greater than 42 Mb. Surprisingly, despite the high completeness, we found that the total assembled size of each of the haplogenomes is substantially smaller than that of the *E. grandis* v2.0 reference genome and previous flow cytometry estimates. We propose that the size difference is not due to collapse of the repeat content of the haplogenome assemblies, but rather due to possible overestimation of the *E. grandis* v2.0 genome assembly as a result of inclusion of partially overlapping alternative haplotypes in highly heterozygous regions of the diploid genome assembly. However, resolving this discrepancy will require further *de novo* genome assemblies for *E. grandis*, possibly including resequencing using long read technology for the reference BRASUZ1 individual.

The success of the trio-binning strategy for discrimination of long reads originating from different parental haplotypes indicate that the approach can be used to build a reference pan-genome comprising haplotype and structural variation for parental species used in intra- and intersectional eucalypt hybrid

combinations. In addition, the use of high-density genetic linkage maps allowed placement of more than 88% of the haplogenome contigs onto one of the eleven chromosomes. This is comparable to previous studies where genetic linkage maps were used (Raymond *et al.*, 2018; Morrissey *et al.*, 2019; Li *et al.*, 2020; Langdon *et al.*, 2020), but genome contiguity and the percentage of placed contigs could be improved if long-range sequencing data is incorporated in the future (Tang *et al.*, 2015).

Finally, we provide the first whole-genome comparison between *E. urophylla* and *E. grandis*. We identify 48,729 SVs, ranging in size from 100 bp to 4.91 Mb. Some of these variants are large enough to be able to cover multiple genes and future studies will focus on understanding the genomic context and functional implications of these variants. An added advantage of this study is the reduced false discovery of SVs which may be introduced when comparing genomes assembled using different pipelines. In conclusion, this study was a successful exploration and foundation for a pan-reference genome for eucalypts and in the near future we aim to use the same strategy to expand the available haplogenomes for *E. grandis* and *E. urophylla*. Once multiple haplogenomes are available, we will study genotype-phenotype associations in segregating experimental populations to move toward incorporating haplotype- and structural variation in breeding strategies as was recently proposed for tomato (Alonge *et al.*, 2020b).

2.7. Tables

Table 2.1 Genome assembly statistics of currently available reference genomes and newly assembled *E. urophylla* and *E. grandis* haplogenomes.

	E. grandis v2.0	*E. grandis*	*E. urophylla*	*E. pauciflora*
Type of sequencing	BAC end cloning (ABI)	Illumina + ONP	Illumina + ONP	Illumina + ONP
Genome coverage[a]	6.73x	54x (ONP)	50x (ONP)	150x (ONP)
Number of contigs[b]	32,724	793	654	465
Total number bases in contigs[b]	691.43 Mb	568.46 Mb	546.11 Mb	594.53 Mb
Contig N50 length[b]	67.2 kb	3.9 Mb	4.4 Mb	2.99 Mb
Contig L50[b]	2,261	38	36	59
Total contigs > 50 kb[b]	288	387	368	NA
Number of contigs[c]	-	1,579	1,418	-
Total number bases in contigs[c]	-	566.72 Mb	544.51 Mb	-
Contig N50 length[c]	-	2.4 Mb	1.9 Mb	-
Contig L50[c]	-	74	83	-
Total contigs > 50 kb[c]	-	522	547	-
BUSCO completion[d]	98.8%	98.7%	99.2%	94.5%
Number scaffolds	4,951	1,279	1,078	415
Total number of bases scaffolded[e]	612.60 Mb	498.98 Mb	481.16 Mb	NA
Scaffold N50	53.80 Mb	43.82 Mb	42.45 Mb	3.23 Mb
Scaffold L50	5	6	6	58
BUSCO completion[f]	98.80%	98.80%	99.20%	94.50%
GC content	39.99%	39.46%	39.44%	39.40%
Repeat content	44.50%	49.06%	48.34%	44.77%

[a] Coverage based on 650 Mb genome size for *E. grandis* and *E. urophylla* and 500 Mb for *E. pauciflora*.
[b] Number of contigs reported for the haplogenome assemblies are before splitting of contigs and genome scaffolding.
[c] Number of contigs reported for the haplogenome assemblies are after splitting contigs with Polar Star.
[d] BUSCO completion score of contig level assembly.
[e] Total number of bases scaffolded onto one of the eleven chromosomes.
[f] BUSCO completion score of all scaffolds (including unplaced scaffolds).

Table 2.2 Summary statistics for each of the two component maps (gra_allmap and uro_allmap) and final consensus anchoring of the *E. urophylla* and *E. grandis* haplogenomes. A greater weight (indicated with w) was given to the component map of the species whose haplogenome was to be scaffolded. Scaffolds that contain no markers or have ambiguous placements are counted as unplaced. Marker density (measured as number of markers per Mb) represents the sum of unique markers from both input datasets.

E. urophylla	**gra_allmap (w=1)**	**uro_allmap (w=2)**	**Anchored**	**Unplaced**
Linkage Groups	11	11	11	n.a.
Markers (unique)	1,577	1,573	3,125	25
Average markers per Mb	3.5	3.5	6.5	0.4
N50 Scaffolds	76	79	81	2
Scaffolds	311	299	351	1,067
Scaffolds with 1 marker	83	80	52	13
Scaffolds with 2 markers	51	53	42	4
Scaffolds with 3 markers	41	37	44	0
Scaffolds with >=4 markers	136	129	213	1
Total bases	**448,984,013**	**447,297,011**	**481,132,251**	**63,374,165**
Percent of genome	82.5%	82.1%	88.4%	11.6%

E. grandis	**gra_allmap (w=2)**	**uro_allmap (w=1)**	**Anchored**	**Unplaced**
Linkage groups	11	11	11	n.a.
Markers (unique)	1,588	1,575	3,129	34
Average markers per Mb	3.3	3.4	6.3	0.5
N50 Scaffolds	71	71	72	1
Scaffolds	282	262	310	1,268
Scaffolds with 1 marker	62	60	49	21
Scaffolds with 2 markers	46	33	26	3
Scaffolds with 3 markers	32	32	30	1
Scaffolds with >=4 markers	142	137	205	1
Total bases	**477,075,775**	**464,179,728**	**498,948,047**	**67,775,781**
Percent of genome	84.2%	81.9%	88.0%	12.0%

***E. grandis* corrected**	**gra_allmap (w=2)**	**uro_allmap (w=1)**	**Anchored**	**Unplaced**
Linkage groups	11	11	11	n.a.
Markers (unique)	1,588	1,575	3,129	34
Average markers per Mb	3.3	3.4	6.3	0.5
N50 Scaffolds	72	72	73	1
Scaffolds	283	263	311	1,268
Scaffolds with 1 marker	62	60	49	21
Scaffolds with 2 markers	46	33	26	3
Scaffolds with 3 markers	32	32	30	1
Scaffolds with >=4 markers	143	138	206	1
Total bases	**477,075,775**	**464,179,728**	**498,948,047**	**67,775,781**
Percent of genome	84.2%	81.9%	88.0%	12.0%

Table 2.3 Repeat element content of assembled haplogenomes.

	E. grandis			*E. urophylla*		
Repeat element type	**Number of elements**	**Length occupied (bp)**	**Percentage of sequence**	**Number of elements**	**Length occupied (bp)**	**Percentage of sequence**
SINEs	1,898	573,397	0.10%	1,850	604,955	0.11%
ALUs	0	0	0.00%	0	0	0.00%
MIRs	163	22,957	0.00%	144	21138	0.00%
LINEs	17,799	16,126,661	2.85%	16,914	15,186,973	2.79%
LINE1	12,470	14,349,121	2.53%	11,657	13,377,778	2.46%
LINE2	2191	452919	0.08%	2,133	458,588	0.08%
L3/CR1	598	282,979	0.05%	687	437,955	0.08%
LTR elements	112,614	121,835,381	21.50%	107,567	114,678,058	21.06%
ERVL	0	0	0.00%	0	0	0.00%
ERVL-MaLRs	0	0	0.00%	0	0	0.00%
ERV_classI	910	823,749	0.15%	860	760,999	0.14%
ERV_classII	80	39039	0.01%	92	46,510	0.01%
DNA elements	101,418	33,813,982	5.97%	98,074	33,335,491	6.12%
hAT-Charlie	1,630	440,706	0.08%	1564	422328	0.08%
TcMar-Tigger	0	0	0.00%	0	0	0.00%
Unclassified	281,973	96,109,077	16.96%	267,902	90,086,398	16.54%
Total interspersed repeats		**268,458,498**	**47.37%**		**253,891,875**	**46.62%**
Small RNA	2,042	953,359	0.17%	2,000	1,088,940	0.20%
Satellites	1,162	821,255	0.14%	1142	777,368	0.14%
Simple repeats	9,016	6,982,108	1.23%	8,685	6,561,653	1.20%
Low complexity	0	0	0.00%	0	0	0.00%
Total	545,964	562,085,188	48.91%	11,827	262,319,836	48.16%

2.8. Figures

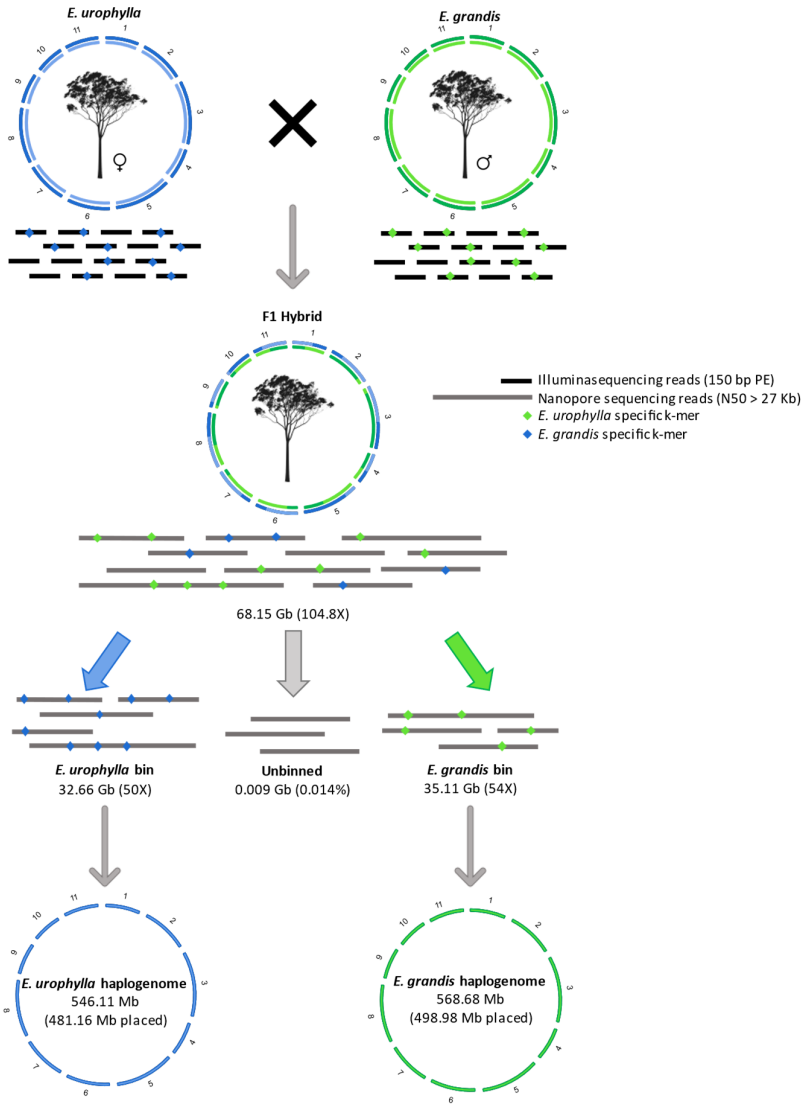


Figure 2.1 Separation of *E. urophylla* and *E. grandis* haplogenomes in the F_1 hybrid using a trio-binning strategy. Using whole-genome Illumina short-read sequencing data of the parental genomes, long-read sequencing data of the F_1 hybrid offspring is separated based on unique parental k-mers into *E. urophylla* and *E. grandis* haplotype bins (amount of Nanopore sequencing data is indicated in gigabases (Gb) below each bin, as well as the estimated genome coverage). Reads that contain no unique k-mers were unbinned and kept in their own bin. Long reads were subsequently assembled independently, resulting in fully assembled *E. urophylla* and *E. grandis* haplogenome (total assembly size is shown below the relevant haplogenome and size of assembly scaffolded into eleven chromosomes are indicated in brackets). This figure is adapted from Koren *et al.*, (2018), and tree images are from https://rooweb.com.au/.

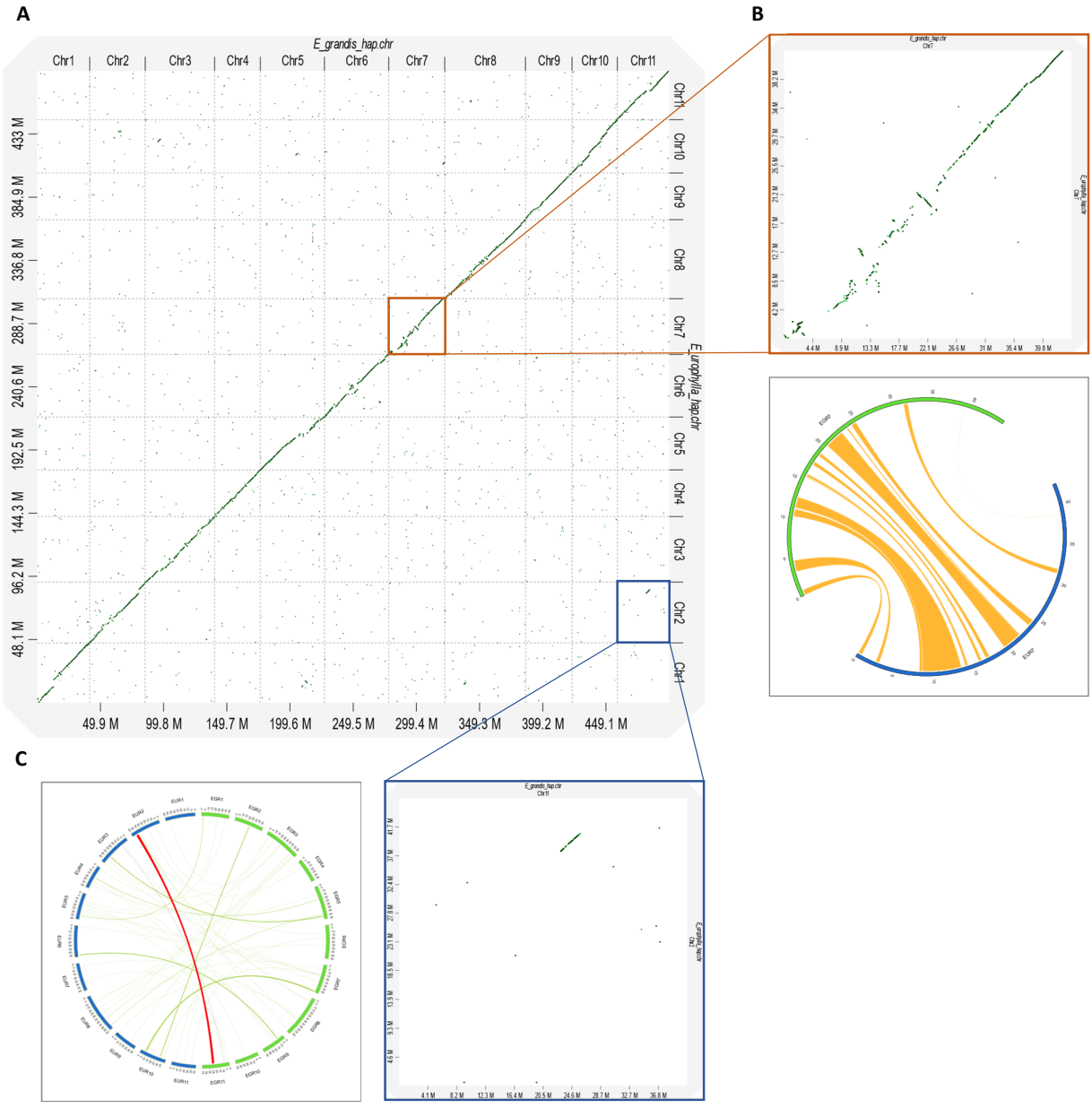


Figure 2.2 Alignment between the *E. grandis* and *E. urophylla* scaffolded haplogenome assemblies.

(A) The *E. grandis* scaffolded haplogenome assembly (498.98 Mb) is shown on the x-axis and the *E. urophylla* scaffolded haplogenome assembly (481.16 Mb) on y-axis and is arranged by chromosome (from one to eleven). **(B)** The right-hand panel (orange block) is a zoom-in of an inversion on chromosome seven as seen with D-Genies (top), and a corresponding Circos visualization of the inversions called by SyRI (bottom). **(C)** The bottom panel (blue block) is a D-Genies zoom-in of a translocation from chromosome eleven in *E. grandis* to chromosome two in *E. urophylla* (on the right), and the corresponding event in a circus plot (highlighted in red). Alignment size is measured in megabases (M in the figure).

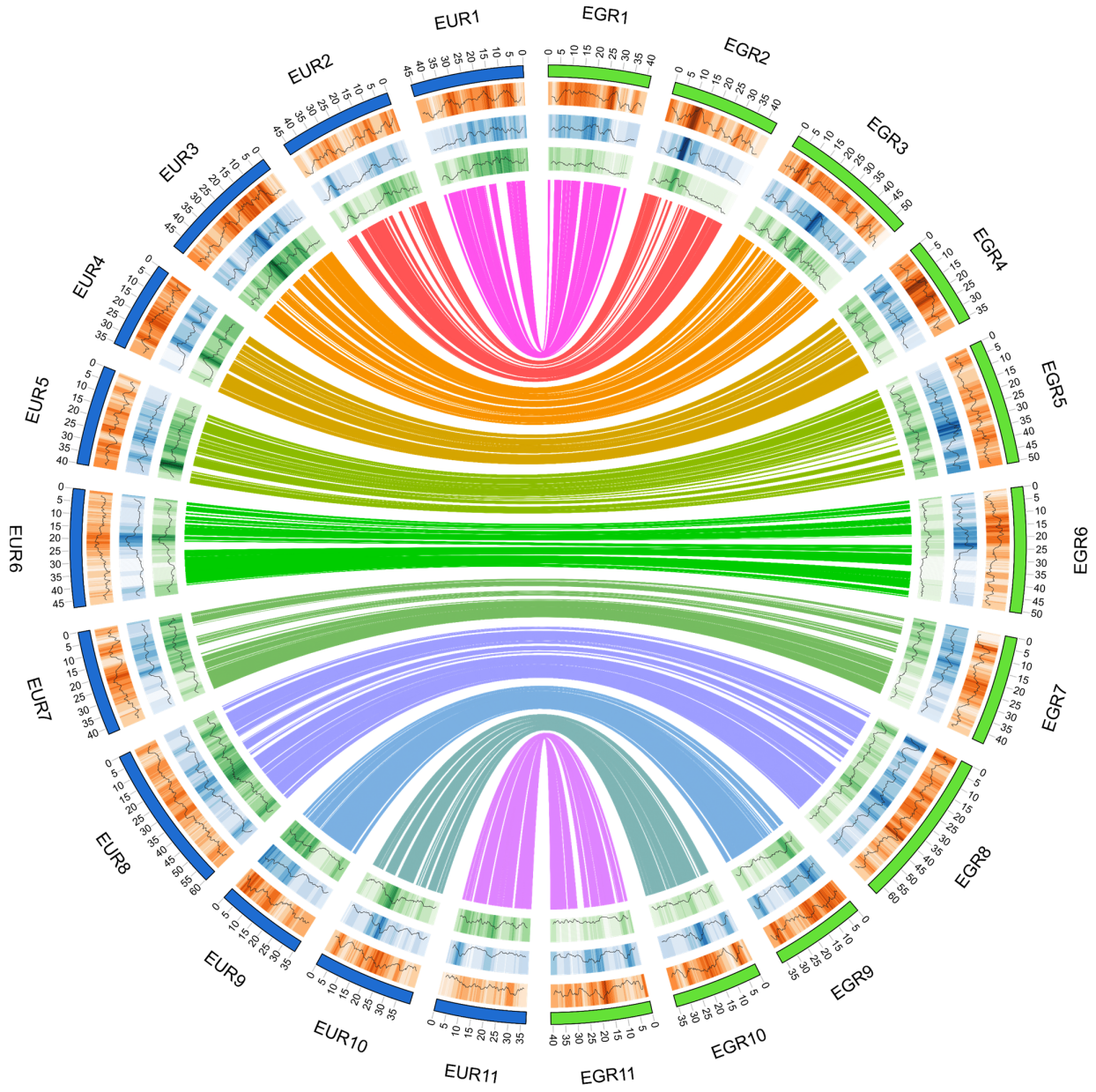
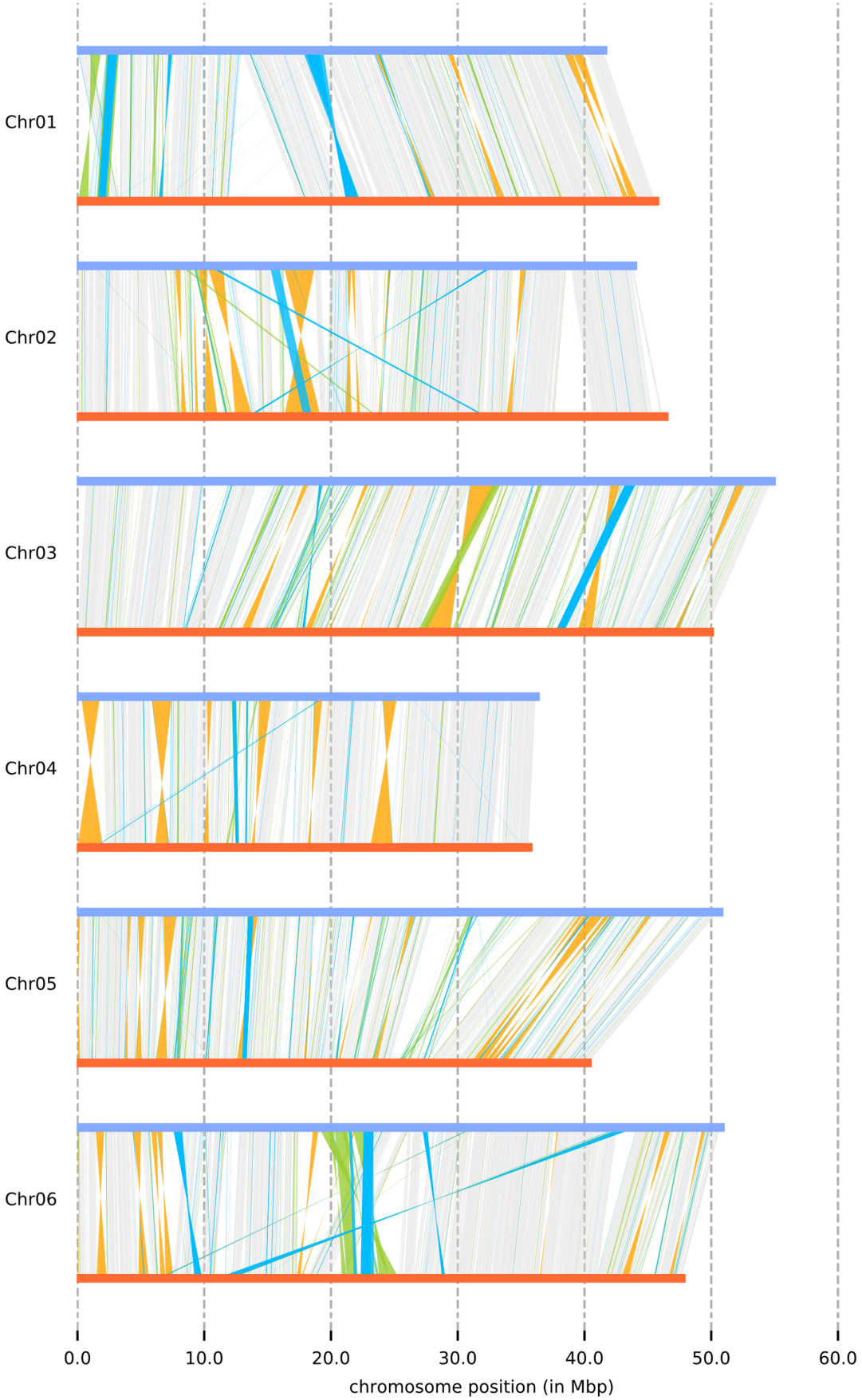


Figure 2.3 Synteny and distribution of LTR retrotransposons along the *E. grandis* and *E. urophylla* haplogenome assemblies for eleven scaffolded chromosomes. Syntenic regions are shown between the *E. urophylla* and *E. grandis* haplogenomes in the middle, based on SyRI (see Figure 2.4). LTR retrotransposon distribution is shown for the *E. urophylla* (EUR) and the *E. grandis* (EGR) haplogenome assemblies. From outside to inside, the heatmaps show the distribution of Copia (orange, ranging from 6 to 21.5%), Gypsy (blue, ranging from 1.3 to 26.5%) and unknown (green, ranging from 2.8 to 16.6%) LTR retrotransposons, with darker shades representing a higher percentage of retrotransposons within the bin (see Supplementary File 1 and 2). Chromosome number and size is indicated on the outer circle in megabases.


Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
0.0
10.0
20.0
30.0
40.0
50.0
60.0
chromosome position (in Mbp)

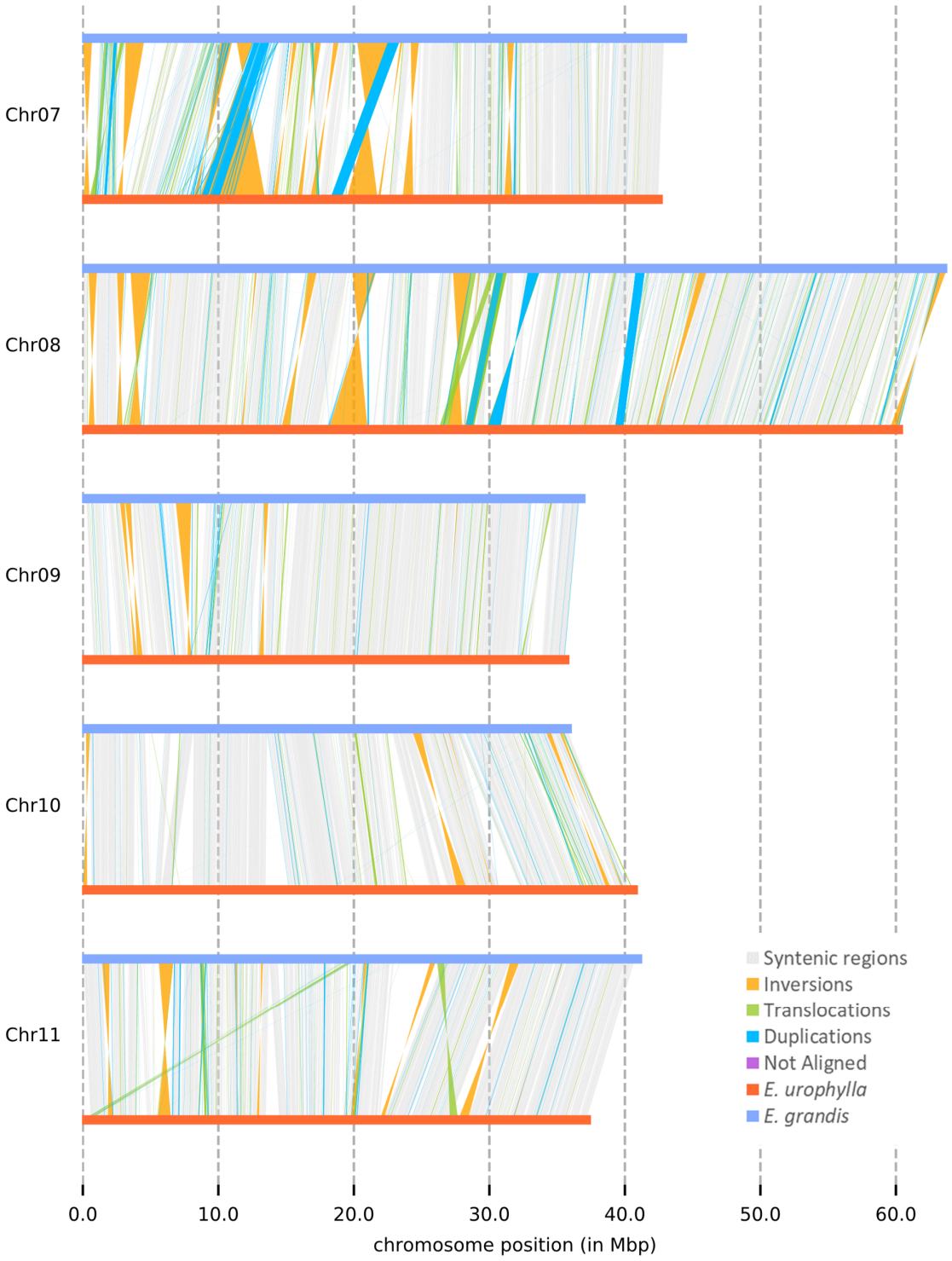


Figure 2.4 Synteny and structural rearrangements between the *E. grandis* and *E. urophylla* haplogenomes for all eleven chromosomes. Position and size of syntenic and rearranged genomic regions between the *E. grandis* haplogenome (blue) and the query genome is the *E. urophylla* haplogenome (orange), for the eleven scaffolded chromosomes. Syntenic regions are indicated in grey, translocations in green, inversions in yellow-

orange and duplications in light blue. Chromosome number is indicated on the y-axis, while chromosome position is shown on the x-axis in megabase-pairs (Mbp).

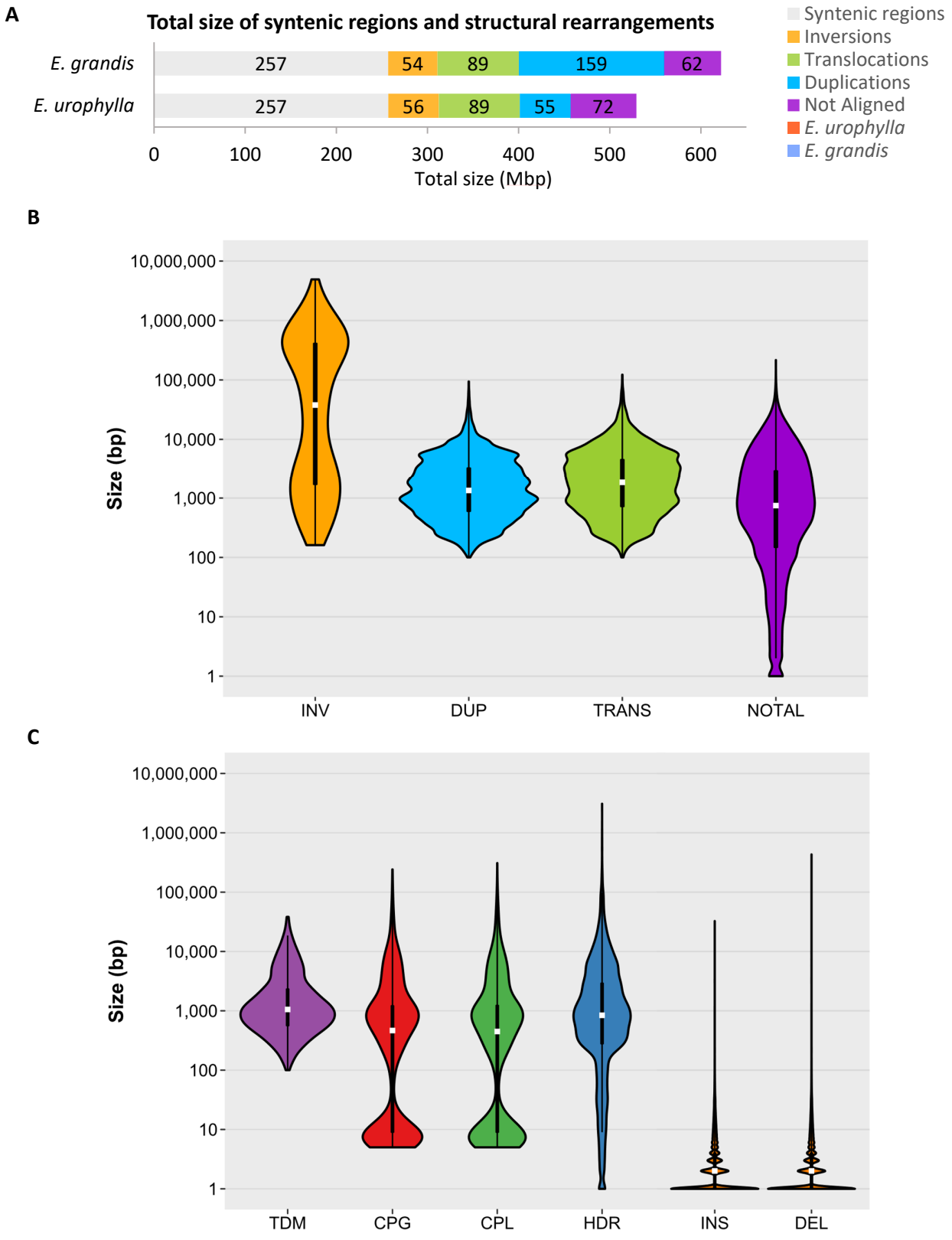


Figure 2.5 Size and distribution of structural rearrangements and local variants between the *E. grandis* and *E. urophylla* haplogenomes. (A) Total size of syntenic and rearranged regions in megabases (Mbp) for the *E. grandis* and *E. urophylla* haplogenome. The size of syntenic or rearranged regions are indicated within the bar

in Mbp, while the bar colour represents the rearrangement type. **(B)** Size distribution of rearranged regions between the *E. grandis* and *E. urophylla* haplogenomes. Size is indicated in base pairs on the y-axis (ranging from one to 4.91 Mbp), and the rearrangement type on the x-axis; INV are inversions, DUP are duplications, TRANS are translocatoins and NOTAL are regions that are not aligned. **(C)** Size distribution of local variations within syntenic and rearranged genomic regions. Size is indicated in base pairs on the y-axis (ranging from one to 3.09 Mbp) and the local variant type on the x-axis: TDM are tandem repeats, CPG and CPL are copy gains/losses, HDR are highly diverged regions, INS are insertions and DEL are deletions.

2.9. References

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, *et al.* 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**: 145-161.e23.

Bartholome J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion J-M. 2015. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* **4**: 1283–1296.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nature Plants* **6**: 914–920.

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, *et al.* 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**: 438–446.

Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD. 2017. Genomic innovation for crop improvement. *Nature* **543**: 346–354.

Brondani RP V, Brondani C, Tarchini R, Grattapaglia D. 1998. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theoretical and Applied Genetics* **97**: 816–827.

Cabanettes F, Klopp C. 2018. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**: e4958.

Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJM. 2017. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications* **8**: 1–10.

Chase MW, Clarke M, Grierson CS, Grierson D, Edwards KJ, Jellis GJ, Barnes SR, Chase MW, Clarke M, Grierson D, *et al.* 2011. One hundred important questions facing plant science research. *New Phytologist* **192**: 6–12.

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L. 2018. The sequenced angiosperm genomes and genome databases. *Frontiers in Plant Science* **9**: 1–14.

de Assis TF. 2000. Production and use of *Eucalyptus hybrids* for industrial purposes. In: *Hybrid breeding and genetics of forest trees*. 63–74.

Dvorak WS, Hodge GR, Payn KG. 2008. The conservation and breeding of *Eucalyptus urophylla*: A case study to better protect important populations and improve productivity. *Southern Forests* **70**: 77–85.

Gaiotto FA, Bramucci M, Grattapaglia D. 1997. Estimation of outcrossing rate in a breeding population of *Eucalyptus urophylla* with dominant RAPD and AFLP markers. *Theoretical and Applied Genetics* **95**: 842–849.

Glenn TC. 2016. 2016 NGS Field Guide: Overview. *The Molecular Ecologist*. URL http://www.molecularecologist.com/next-gen-fieldguide-2016/. [acessed 13 April 2019]

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**: 1–13.

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

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

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Hirakawa H, Nakamura Y, Kaneko T, Isobe S, Sakai H, Kato T, Hibino T, Sasamoto S, Watanabe A, Yamada M, *et al.* 2011. Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. *Plant Biotechnology* **28**: 471–480.

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nature Reviews Genetics* **21**: 171–189.

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

Jiao WB, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* **36**: 64–70.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of

rearrangements with altered evolutionary dynamics. *Nature Communications* **11**: 1–10.

Kim D, Song L, Breitwieser FP, Salzberg SL. **2016**. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research* **26**: 1721–1729.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. **2018**. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**: 1174–1182.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. **2017**. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**: 722–736.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. **2009**. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639–1645.

Kullan ARK, van Dyk MM, Jones N, Kanzler A, Bayley A, Myburg AA. **2012**. High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* × *E. urophylla*. *Tree Genetics and Genomes* **8**: 163–175.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg S. **2004**. Versatile and open software for comparing large genomes. *Genome Biology* **5**: 1–9.

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik M V. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* **9**: 1–15.

Langdon KS, King GJ, Baten A, Mauleon R, Bundock PC, Topp BL, Nock CJ. **2020**. Maximising recombination across macadamia populations to generate linkage maps for genome anchoring. *Scientific Reports* **10**: 1-15.

Li H. **2016**. Minimap and miniasm: Fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110.

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

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

Li W, Zhu X, Zhang Q-J, Li K, Zhang D, Shi C, Gao L-Z. **2020**. SMRT sequencing generates the chromosome-scale reference genome of tropical fruit mango, *Mangifera indica*. *Biorxiv*.

Marçais G, Kingsford C. **2011**. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Marques CM, Brondani RPV, 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.

Michael TP, VanBuren R. **2015**. Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology* **24**: 71–81.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. **2018**. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142–i150.

Moran GF, Bell JC, Griffin AR. **1989**. Reduction in levels of inbreeding in a seed orchard of *Eucalyptus regnans* F. Muell. compared to with natural populations. *Silvae Genetica* **38**: 32–35.

Morrissey J, Stack JC, Valls R, Motamayor JC. **2019**. Low-cost assembly of a cacao crop genome is able to resolve complex heterozygous bubbles. *Horticulture Research* **6**: 1-13.

Motazedi E, Finkers R, Maliepaard C, de Ridder D. **2017**. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in Bioinformatics*: 387-403.

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.

Ogawa D, Nonoue Y, Tsunematsu H, Kanno N, Yamamoto T, Yonemaru J. **2019**. Discovery of QTL alleles for grain shape in the Japan-MAGIC rice population using haplotype information. *G3 Genes|Genomes|Genetics* **8**: 3559–3565.

Ogawa D, Yamamoto E, Ohtani T, Kanno N, Tsunematsu H, Nonoue Y, Yano M, Yamamoto T, Yonemaru JI. **2018**. Haplotype-based allele mining in the Japan-MAGIC rice population. *Scientific Reports* **8**: 1–11.

Ou S, Jiang N. **2018**. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176**: 1410–1422.

Phase Genomics. **2020**. phasegenomics/polar_star. Available at https://github.com/phasegenomics/polar_star.

Pinard D, Myburg AA, Mizrachi E. **2019**. The plastid and mitochondrial genomes of Eucalyptus grandis. *BMC Genomics*

20: 1–14.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. **2020**. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**: 1-10.

Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, Vergne P, Moja S, Choisne N, Pont C, *et al.* 2018. The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics* **50**: 772–777.

Rezende GDSP, de Resende MD V., de Assis TF. **2014**. *Eucalyptus* breeding for clonal forestry. In: T Fenning, Ed. *Challenges and opportunities for the world's forests in the 21st century.* Forestry Sciences, Dordrecht: Springer, **81**: 393–424.

Rhie A, Walenz BP, Koren S, Phillippy AM. **2020**. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**: 1–27.

Seppey M, Manni M, Zdobnov EM. **2019**. BUSCO: Assessing genome assembly and annotation completeness. *Methods in Molecular Biology* **1962**: 227–245.

seqtk, Toolkit for processing sequences in FASTA/Q formats. Available at https://github.com/lh3/seqtk.

Sherman RM, Salzberg SL. **2020**. Pan-genomics in the human genome era. *Nature Reviews Genetics* **4**: 243-254.

Shirasawa K, Esumi T, Hirakawa H, Tanaka H, Itai A, Ghelfi A, Nagasaki H, Isobe S. 2019. Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (*Cerasus* × *yedoensis*). *DNA Research* **26**: 379–389.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. **2015**. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Smit A, Hubley R. **2008**. RepeatModeler Open-1.0. Available at http://www.repeatmasker.org/RepeatModeler/.

Smit AFA, Hubley R, Green P. **2013**. RepeatMasker Open-4.0. Available at http://www.repeatmasker.org.

Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. **2015**. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* **16**: 1-15.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. **2017**. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202–2204.

Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R. **2020a**. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing *de novo* assemblies. *GigaScience* **9**: 1–12.

Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, Zhang J, Mainiero S, Strickler SR, Catala C, *et al.* 2020b. Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nature Communications* **11**: 1-11.

Wick R. 2018. Porechop: adapter trimmer for Oxford Nanopore reads. Available at https://github.com/rrwick/Porechop.

Wood DE, Lu J, Langmead B. **2019**. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**: 1–13.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, *et al.* 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology* **34**: 303–311.

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. **2019**. The population genetics of structural variants in grapevine domestication. *Nature Plants* **5**: 965–979.

Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, Li J, Chakraborty S, Balan B, Jiang C-Z, *et al.* 2019. Sequencing a *Juglans regia* × *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. *Horticulture Research* **6**: 1-16.

Zimin A V., Luo M-C, Marçais G, Salzberg SL, Yorke JA, Puiu D, Koren S, Zhu T, Dvořák J. **2017**. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* **27**: 787–792.

2.10. Supplementary Tables

Supplementary Table 2.1 Illumina sequencing results. Raw read statistics are given followed by the mapping rate of reads to the main scaffolds of the *E. grandis* v2.0 reference genome (Myburg *et al.*, 2014) as well as the *E. grandis* mitochondrial and plastid genomes (Pinard *et al.*, 2019) after read contaminants have been removed. The total amount of sequencing data generated is given in gigabases (Gb).

Sample ID	Total bases (Gb)	Total reads	Q20 (%)	Q30 (%)	Mapping rate[a] (%)	Properly paired[b] (%)
E. urophylla parent	127.5	884,340,104	97.097	92.515	94.34	82.61
E. grandis parent	141.6	937,939,296	96.357	90.786	95.07	86.55
F_1 hybrid	116.1	769,097,570	97.361	93.182	94.78	84.16

[a] Mapping rate of Illumina short reads with contaminates removed to the *E. grandis* v2.0 main scaffolds along with the mitochondrial and plastid reference genomes.
[b] Mapping rate of Illumina short reads with contaminates removed to *E. grandis* v2.0 main scaffolds along with the mitochondrial and plastid reference genomes that are properly paired reads.

Supplementary Table 2.2 Nanopore sequencing results for the F_1 hybrid individual. The 100/G tip and SDS-based runs are both MinION sequencing runs (using different DNA isolation methods as indicated by the name). kb – kilobase, Gb – gigabase, Q7 – quality score of seven.

Sequencing run	Number of reads	Total bases (Gb)	Number reads > Q7[a]	Total bases > Q7 (Gb)[b]	Longest read > Q7 (kb)	Read N50 (kb)
100/G Tip	2,169,269	11.18	1,645,369 (75.85%)	9.30 (83.18%)	182.95	18.96
SDS-based	429,264	2.55	364,541 (84.92%)	2.28 (89.41%)	304.89	23.86
PromethION	3,290,284	61.59	2,875,796 (87.40%)	56.57 (91.85%)	221.38	28.00
Total[c]	**5,888,817**	**75.32**	**4,885,706 (82.97%)**	**68.15 (90.48%)**	**304.89**	

[a] Percentage of reads passing QC are shown in brackets.
[b] Percentage bases passing QC are shown in brackets.
[c] 99.45% of basecalled reads mapped to the *E. grandis* v2.0 main scaffolds along with the mitochondrial and plastid reference genomes, with 99.94% mapping of the *E. urophylla* read bin and 99.93% of the *E. grandis* read bin.

Supplementary Table 2.3 Summary statistics for long-read binning using the parental short reads.

Bin	Reads[a]	L50	N50	Max read length	Sum	Percentage of total	Mapping rate[b]	Properly paired[c]	Mapping rate[d]	Properly paired[e]
E. grandis	1,999,561	465,851	27,604	262,889	35.11 Gb	51.80%	98.73%	93.79%	98.11%	85.03%
E. urophylla	1,877,139	433,849	27,548	304,871	32.66 Gb	48.18%	99.10%	92.91%	97.67%	84.85%
Unknown	6,693	2,553	1,385	7,631	9.59 Mb	0.014%	NA	NA	NA	NA
Total	**3,883,393**				**67.78 Gb**					

[a] Only reads greater than 500 base pairs were considered.
[b] Mapping rate of corresponding parental species Illumina short reads to haplogenome assembly.
[c] Mapping rate of parental species Illumina short reads to haplogenome assembly that are properly paired reads.
[d] Mapping rate of alternative parental species Illumina short reads to haplogenome assembly.
[e] Mapping rate of alternative parental species Illumina short reads to haplogenome assembly that are properly paired reads.

Supplementary Table 2.4 Summary statistics of placed and unplaced contigs after scaffolding with ALLMAPS for the *E. urophylla* and *E. grandis* haplogenomes respectively. # indicates number.

Assembly	Unplaced *E. urophylla*	Placed *E. urophylla*	Unplaced *E. grandis*	Placed *E. grandis*
# contigs (>= 0 bp)	1,067	11	1,268	11
# contigs (>= 1000 bp)	1,067	11	1,268	11
# contigs (>= 5000 bp)	796	11	968	11
# contigs (>= 10000 bp)	519	11	599	11
# contigs (>= 25000 bp)	308	11	325	11
# contigs (>= 50000 bp)	204	11	217	11
Total length (>= 0 bp)	63,374,165	481,166,251	67,775,781	498,977,947
Total length (>= 1000 bp)	63,374,165	481,166,251	67,775,781	498,977,947
Total length (>= 5000 bp)	62,333,808	481,166,251	66,626,180	498,977,947
Total length (>= 10000 bp)	60,386,045	481,166,251	64,045,556	498,977,947
Total length (>= 25000 bp)	57,251,799	481,166,251	59,766,617	498,977,947
Total length (>= 50000 bp)	53,663,936	481,166,251	55,918,898	498,977,947
# contigs	1,067	11	1,268	11
Largest contig	2,720,265	60,186,531	2,440,300	63,773,828
Total length	63,374,165	481,166,251	67,775,781	498,977,947
GC (%)	39.5	39	39.52	39
N50	324,500	45,562,418	324,100	44,251,077
N75	118,300	40,242,915	96,704	40,936,616
L50	46	5	53	5
L75	127	8	141	8
# N's per 100 kbp	0.78	7.3	1.64	6.31

Supplementary Table 2.5 Number and total length of syntenic and rearranged regions in the *E. grandis* and *E. urophylla* haplogenomes. Regions are shown between the *E. grandis* v2.0 reference genome and the *E. grandis* haplogenome as well as the *E. grandis* haplogenome and the *E. urophylla* haplogenome. Rearrangements were called with SyRI (Synteny and rearrangement identifier) with a minimum 100 bp size, using the *E. grandis* v2.0 or *E. grandis* haplogenome as the reference in the two analyses. Length is indicated in basepairs (bp). Only the eleven scaffolded chromosomes were compared for identification of rearranged regions.

***E. grandis* v2.0 (reference) vs *E. grandis* haplogenome (query)**

Variation type[a]	SYN	INV	TRANS	DUP (v2.0)	DUP (*E. grandis*)	Not aligned (v2.0)	Not aligned (*E. grandis*)
Count	13,463	167	9,290	29,596	17,519	28,761	18,904
Length (v2.0)	317,981,657	57,482,207	75,974,491	141,439,740	-	111,692,400	-
Length (*E. grandis*)	317,513,455	45,151,373	75,544,141	-	50,831,969	-	41,963,576

***E. grandis* haplogenome (reference) vs *E. urophylla* haplogenome (query)**

Variation type	SYN	INV	TRANS	DUP (*E. grandis*)	DUP (*E. urophylla*)	Not aligned (*E. grandis*)	Not aligned (*E. urophylla*)
Count	15,236	189	10,526	21,149	16,865	24,700	22,770
Length (*E. grandis*)	256,747,807	54,233,806	89,269,151	159,115,873	-	72,234,120	-
Length (*E. urophylla*)	256,876,296	55,605,620	88,801,518	-	55,495,066	-	62,322,133

[a] SYN: syntenic region, INV: inversion, TRANS: translocation, DUP: duplication in the genome indicated in brackets, where v2.0 is the *E.* grandis v2.0 reference genome, *E. grandis* the *E. grandis* haplogenome and *E. urophylla* the *E. urophylla* haplogenome, Not aligned: unaligned regions in *E. grandis* or *E. urophylla* (query) haplogenome.

Supplementary Table 2.6 Number and total length of local sequence variation in syntenic and rearranged region between the *E. grandis* v2.0 reference genome and *E. grandis* haplogenome as well as between the *E. grandis* and *E. urophylla* haplogenomes. Local sequence variants were called with SyRI using the *E. grandis* v2.0 and *E. grandis* haplogenome as the reference genome respectively. Only the eleven scaffolded chromosomes were compared for local sequence variant identification.

***E. grandis* v2.0 (reference) vs *E. grandis* haplogenome (query)**

Variation type	SNPs	Insertions	Deletions	Copygains	Copylosses	Highly diverged	Tandem repeats	Total
Count	6,373,115	539,605	578,616	1,759	1,665	9,202	321	7,504,283
Length *E. grandis* v2.0	6,373,115	-	5,615,561	-	10,885,284	45,970,827	685,226	69,530,013
Length *E. grandis*	6,373,115	6,116,677	-	11,424,778	-	31,670,584	870,720	56,455,874

***E. grandis* haplogenome (reference) vs *E. urophylla* haplogenome (query)**

Variation type	SNPs	Insertions	Deletions	Copygains	Copylosses	Highly diverged	Tandem repeats	Total
Count	8,376,569	676,636	704,383	2,172	2,127	8,018	268	9,770,173
Length *E. grandis*	8,376,569	-	8,412,691	-	9,689,158	38,129,322	656,872	65,264,612
Length *E. urophylla*	8,376,569	7,629,721	-	9,578,692	-	40,181,747	680,693	66,447,422

Supplementary Table 2.7 Inversions between the *E. grandis* and *E. urophylla* haplogenomes that are larger than 50 kb.

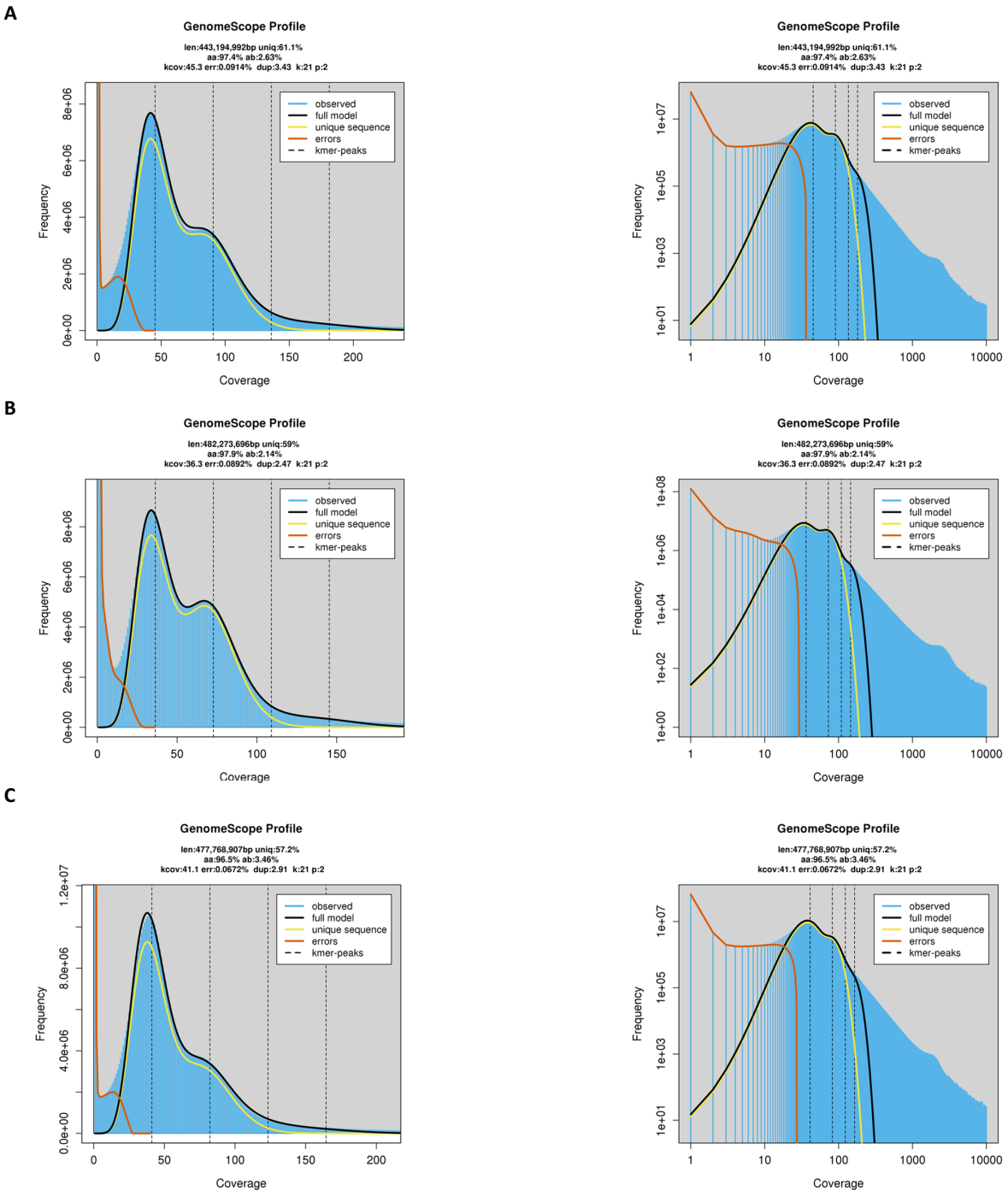
E. grandis haplogenome				*E. urophylla* haplogenome			
Chr	**Start**	**End**	**Reference length**	**Chr**	**Start**	**End**	**Query length**
Chr01	23,388,498	23,774,716	386,218	Chr01	27,866,247	28,333,853	467,606
Chr01	29,088,860	29,535,584	446,724	Chr01	33,200,026	33,802,257	602,231
Chr01	38,325,932	38,988,219	662,287	Chr01	43,142,878	43,537,862	394,984
Chr01	38,989,424	39,795,647	806,223	Chr01	43,575,027	44,294,133	719,106
Chr02	10,256,034	11,495,357	1,239,323	Chr02	12,449,219	13,806,880	1,357,661
Chr02	16,268,629	18,735,729	2,467,100	Chr02	16,317,615	19,174,190	2,856,575
Chr02	21,273,193	21,613,433	340,240	Chr02	21,133,426	21,629,094	495,668
Chr02	21,614,540	21,866,290	251,750	Chr02	22,058,000	22,319,351	261,351
Chr02	24,530,674	24,753,625	222,951	Chr02	23,723,944	23,839,960	116,016
Chr02	34,946,307	35,416,437	470,130	Chr02	33,832,261	34,243,757	411,496
Chr02	7,710,307	8,128,604	418,297	Chr02	8,216,904	8,593,150	376,246
Chr02	9,120,701	9,141,333	20,632	Chr02	9,309,778	9,532,539	222,761
Chr02	9,543,813	10,251,800	707,987	Chr02	10,024,991	11,058,528	1,033,537
Chr03	17,924,896	18,361,694	436,798	Chr03	12,844,626	13,544,499	699,873
Chr03	22,750,923	23,075,176	324,253	Chr03	17,872,261	18,383,259	510,998
Chr03	24,894,740	25,020,737	125,997	Chr03	19,732,494	19,842,712	110,218
Chr03	26,582,840	26,695,689	112,849	Chr03	22,211,437	22,302,731	91,294
Chr03	29,450,360	29,499,020	48,660	Chr03	25,057,971	25,153,753	95,782
Chr03	31,007,465	33,222,991	2,215,526	Chr03	27,113,759	29,387,690	2,273,931
Chr03	42,077,277	42,865,022	787,745	Chr03	39,421,192	40,598,627	1,177,435
Chr03	52,056,591	52,768,850	712,259	Chr03	46,996,755	47,400,472	403,717
Chr04	10,204,799	10,608,432	403,633	Chr04	9,959,942	10,372,400	412,458
Chr04	14,353,739	15,314,090	960,351	Chr04	13,693,800	13,989,776	295,976
Chr04	18,680,518	19,330,061	649,543	Chr04	18,130,979	18,386,712	255,733
Chr04	19,856,681	19,916,877	60,196	Chr04	18,916,777	18,948,259	31,482
Chr04	24,074,346	25,225,372	1,151,026	Chr04	23,089,700	24,877,693	1,787,993
Chr04	285,506	1,781,371	1,495,865	Chr04	8	1,997,000	1,996,992
Chr04	5,839,102	7,451,977	1,612,875	Chr04	6,099,540	7,272,200	1,172,660
Chr05	1,139	191,613	190,474	Chr05	1	199,651	199,650
Chr05	13,767,410	14,264,679	497,269	Chr05	12,544,177	13,230,121	685,944
Chr05	18,731,556	18,772,986	41,430	Chr05	17,841,218	18,037,838	196,620
Chr05	23,214,319	23,331,219	116,900	Chr05	20,525,113	20,682,115	157,002
Chr05	26,312,453	26,639,176	326,723	Chr05	23,326,012	23,619,675	293,663
Chr05	39,764,872	40,079,538	314,666	Chr05	31,002,806	31,460,527	457,721
Chr05	3,977,112	4,219,061	241,949	Chr05	3,704,167	3,953,788	249,621
Chr05	40,147,672	40,620,311	472,639	Chr05	31,467,622	32,031,945	564,323
Chr05	40,664,102	41,683,031	1,018,929	Chr05	32,044,329	32,499,398	455,069
Chr05	41,852,810	42,212,844	360,034	Chr05	32,508,317	32,842,567	334,250
Chr05	42,231,666	42,497,444	265,778	Chr05	33,225,046	33,730,213	505,167
Chr05	45,116,621	45,471,128	354,507	Chr05	36,775,424	37,162,018	386,594
Chr05	4,790,253	5,284,213	493,960	Chr05	4,587,258	5,054,184	466,926
Chr05	6,762,823	7,811,079	1,048,256	Chr05	6,186,655	7,075,904	889,249
Chr06	1,422,135	2,104,051	681,916	Chr06	1,501,336	2,307,495	806,159
Chr06	17,562	116,068	98,506	Chr06	1	105,034	105,033
Chr06	18,585,323	19,047,661	462,338	Chr06	17,287,053	17,586,309	299,256

Chr06	21,901,981	22,015,039	113,058	Chr06	17,587,701	17,694,441	106,740
Chr06	37,843,975	37,953,970	109,995	Chr06	36,451,444	36,543,431	91,987
Chr06	4,348,505	5,014,910	666,405	Chr06	4,809,501	5,634,648	825,147
Chr06	46,380,498	46,880,027	499,529	Chr06	42,887,607	43,469,053	581,446
Chr06	49,293,183	49,600,931	307,748	Chr06	46,607,955	46,889,702	281,747
Chr06	5,790,165	6,245,897	455,732	Chr06	6,531,092	6,952,534	421,442
Chr06	6,276,170	6,681,490	405,320	Chr06	6,952,991	7,550,358	597,367
Chr07	10,195,846	11,089,184	893,338	Chr07	8,134,852	8,493,558	358,706
Chr07	11,303,891	12,661,932	1,358,041	Chr07	8,550,716	13,463,227	4,912,511
Chr07	13,073	685,800	672,727	Chr07	53,539	498,307	444,768
Chr07	15,398,973	15,469,228	70,255	Chr07	14,007,680	14,077,253	69,573
Chr07	15,548,182	15,826,775	278,593	Chr07	14,078,345	14,337,355	259,010
Chr07	17,183,241	17,610,391	427,150	Chr07	15,818,970	16,167,843	348,873
Chr07	18,528,803	18,920,067	391,264	Chr07	16,770,493	17,235,509	465,016
Chr07	20,211,368	22,642,279	2,430,911	Chr07	19,458,012	21,757,921	2,299,909
Chr07	23,468,568	23,626,990	158,422	Chr07	21,768,463	21,937,670	169,207
Chr07	24,191,002	24,827,475	636,473	Chr07	23,584,113	24,382,712	798,599
Chr07	31,325,104	31,831,058	505,954	Chr07	31,095,389	31,656,188	560,799
Chr07	3,157,850	4,559,613	1,401,763	Chr07	2,510,914	2,920,972	410,058
Chr08	12,725,330	12,863,580	138,250	Chr08	11,688,255	11,819,476	131,221
Chr08	16,700,735	17,318,973	618,238	Chr08	14,631,661	15,206,380	574,719
Chr08	19,875,397	21,709,986	1,834,589	Chr08	18,154,032	21,017,051	2,863,019
Chr08	2,510,912	3,087,516	576,604	Chr08	2,492,921	2,934,491	441,570
Chr08	27,297,239	28,670,928	1,373,689	Chr08	26,916,592	28,009,638	1,093,046
Chr08	29,225,629	29,316,542	90,913	Chr08	28,143,627	28,201,928	58,301
Chr08	3,501,184	5,038,276	1,537,092	Chr08	3,372,176	4,298,000	925,824
Chr08	419,214	1,055,841	636,627	Chr08	407,448	877,416	469,968
Chr08	45,309,936	45,408,786	98,850	Chr08	42,198,433	42,388,897	190,464
Chr08	45,522,689	46,091,114	568,425	Chr08	42,436,168	42,559,996	123,828
Chr08	63,156,082	63,769,512	613,430	Chr08	59,525,804	60,126,627	600,823
Chr09	13,376,823	13,702,567	325,744	Chr09	12,998,893	13,338,127	339,234
Chr09	2,707,324	3,062,759	355,435	Chr09	3,777,976	4,028,769	250,793
Chr09	27,619,973	27,697,783	77,810	Chr09	26,834,362	26,876,310	41,948
Chr09	3,101,186	3,530,067	428,881	Chr09	4,029,319	4,440,544	411,225
Chr09	6,826,654	8,005,141	1,178,487	Chr09	7,719,031	8,059,633	340,602
Chr10	233,352	527,104	293,752	Chr10	20,102	323,104	303,002
Chr10	24,242,636	24,889,751	647,115	Chr10	27,630,162	28,354,048	723,886
Chr10	26,475,377	26,532,239	56,862	Chr10	29,859,221	29,888,792	29,571
Chr10	26,902,198	26,970,937	68,739	Chr10	30,235,379	30,306,797	71,418
Chr10	34,082,797	34,441,174	358,377	Chr10	38,683,430	39,016,229	332,799
Chr10	35,068,172	35,243,214	175,042	Chr10	39,806,630	39,909,104	102,474
Chr11	11,262,172	11,396,228	134,056	Chr11	11,417,392	11,457,497	40,105
Chr11	13,155,091	13,298,496	143,405	Chr11	12,829,340	13,017,471	188,131
Chr11	1,408,650	1,977,286	568,636	Chr11	1,848,260	2,259,527	411,267
Chr11	20,764,115	20,973,486	209,371	Chr11	19,775,853	20,065,496	289,643
Chr11	25,701,549	26,104,207	402,658	Chr11	21,898,345	22,199,580	301,235
Chr11	31,583,111	32,300,126	717,015	Chr11	27,699,796	28,409,715	709,919
Chr11	35,437,224	35,496,143	58,919	Chr11	31,711,764	31,775,563	63,799
Chr11	5,589,851	6,679,990	1,090,139	Chr11	5,522,667	6,499,452	976,785

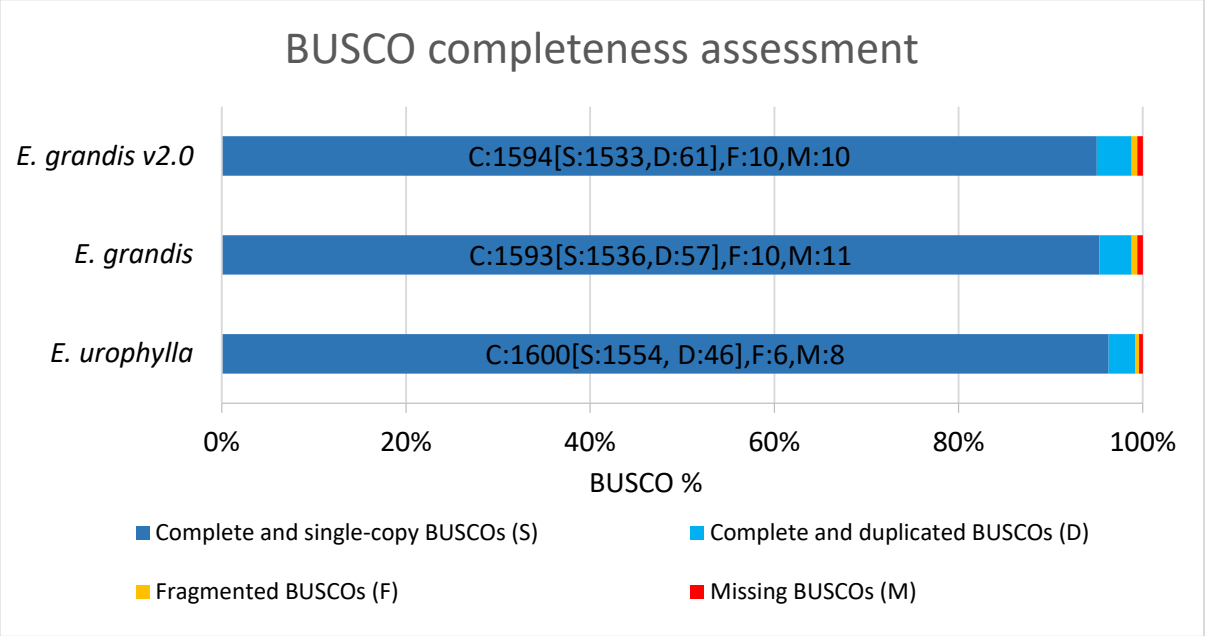
Supplementary Table 2.8 Translocations between the *E. grandis* and *E. urophylla* haplogenomes that are larger than 50 kb.

E. grandis haplogenome				*E. urophylla* haplogenome			
Chr	**Start**	**End**	**Reference length**	**Chr**	**Start**	**End**	**Query length**
Chr01	2,749,813	2,819,595	69,782	Chr01	1,917,560	1,987,365	69,805
Chr02	8,220,591	8,275,256	54,665	Chr02	23,641,937	23,696,439	54,502
Chr04	12,396,028	12,452,029	56,001	Chr04	12,703,390	12,759,400	56,010
Chr06	20,361,544	20,482,641	121,097	Chr06	21,352,697	21,473,569	120,872
Chr06	20,682,722	20,765,954	83,232	Chr06	21,659,542	21,742,402	82,860
Chr07	43,614,628	43,679,697	65,069	Chr10	7,513,443	7,578,449	65,006
Chr07	43,753,217	43,808,068	54,851	Chr10	7,644,800	7,699,782	54,982
Chr07	43,934,853	44,028,924	94,071	Chr10	7,818,126	7,912,179	94,053
Chr07	44,029,147	44,090,096	60,949	Chr10	7,912,179	7,972,966	60,787
Chr08	28,793,518	28,854,941	61,423	Chr08	26,691,561	26,752,939	61,378
Chr08	28,960,832	29,054,796	93,964	Chr08	26,841,892	26,935,863	93,971
Chr09	8,771,435	8,829,203	57,768	Chr06	45,043,500	45,101,519	58,019
Chr09	8,837,050	8,896,455	59,405	Chr06	45,101,515	45,160,936	59,421
Chr10	7,148,524	7,214,985	66,461	Chr10	6,545,748	6,612,181	66,433
Chr11	23,115,517	23,174,363	58,846	Chr02	37,900,157	37,958,643	58,486
Chr11	24,151,203	24,203,956	52,753	Chr02	39,003,539	39,056,219	52,680
Chr11	24,238,142	24,290,314	52,172	Chr02	39,079,522	39,131,773	52,251
Chr11	24,387,150	24,509,566	122,416	Chr02	39,238,929	39,361,120	122,191
Chr11	24,591,384	24,648,523	57,139	Chr02	39,436,979	39,493,981	57,002
Chr11	24,855,007	24,924,836	69,829	Chr02	39,719,434	39,789,354	69,920
Chr11	24,936,151	25,001,436	65,285	Chr02	39,800,932	39,866,164	65,232
Chr11	25,001,695	25,072,755	71,060	Chr02	39,866,476	39,937,490	71,014
Chr11	25,123,131	25,186,141	63,010	Chr02	39,980,961	40,044,235	63,274
Chr11	25,240,856	25,293,593	52,737	Chr02	40,102,319	40,155,043	52,724

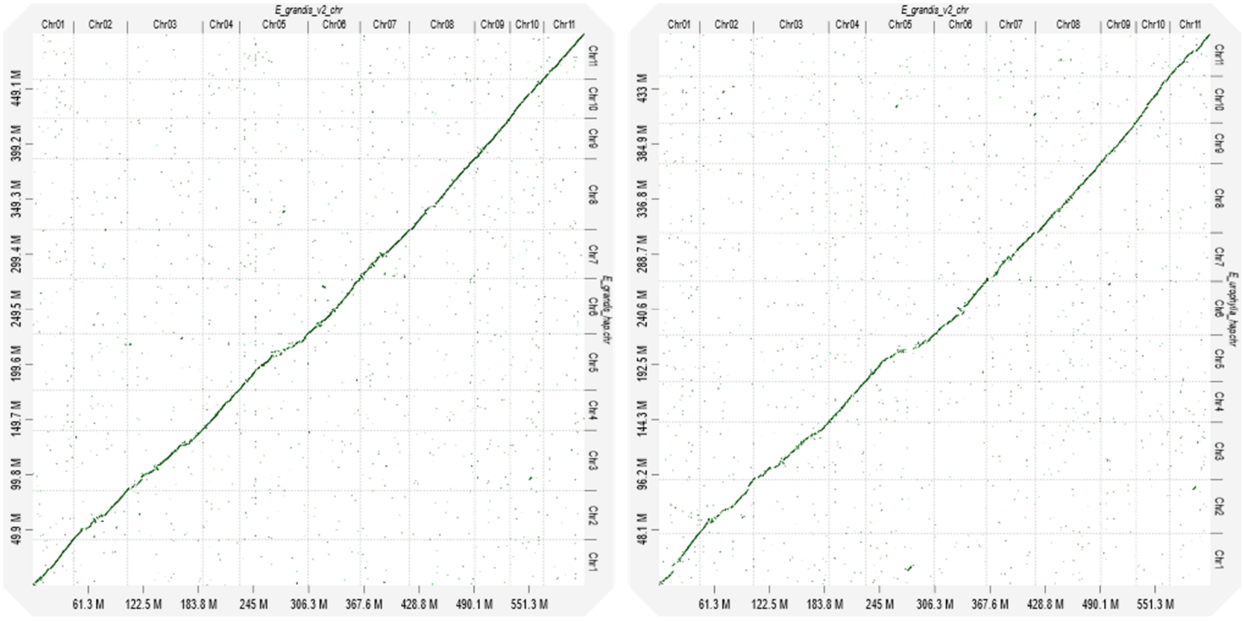
2.11. Supplementary Figures



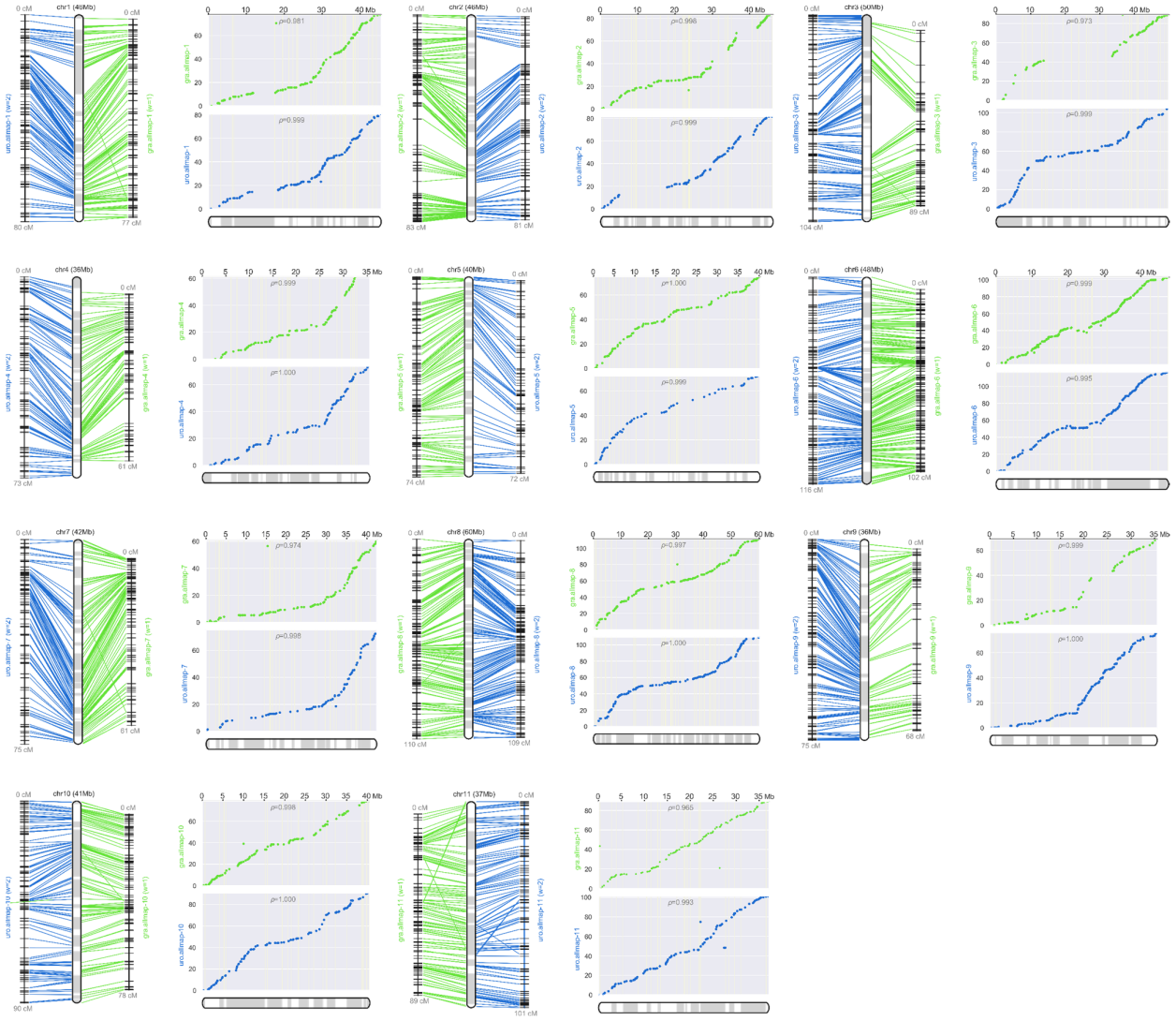
Supplementary Figure 2.1 Genome size estimates. Genome size was estimates for the **(A)** *E. urophylla*, **(B)** *E. grandis* and **(C)** the *E. urophylla* x *E. grandis* F_1 hybrid genomes. Genome size was estimated at k = 21 with GenomeScope2.0.



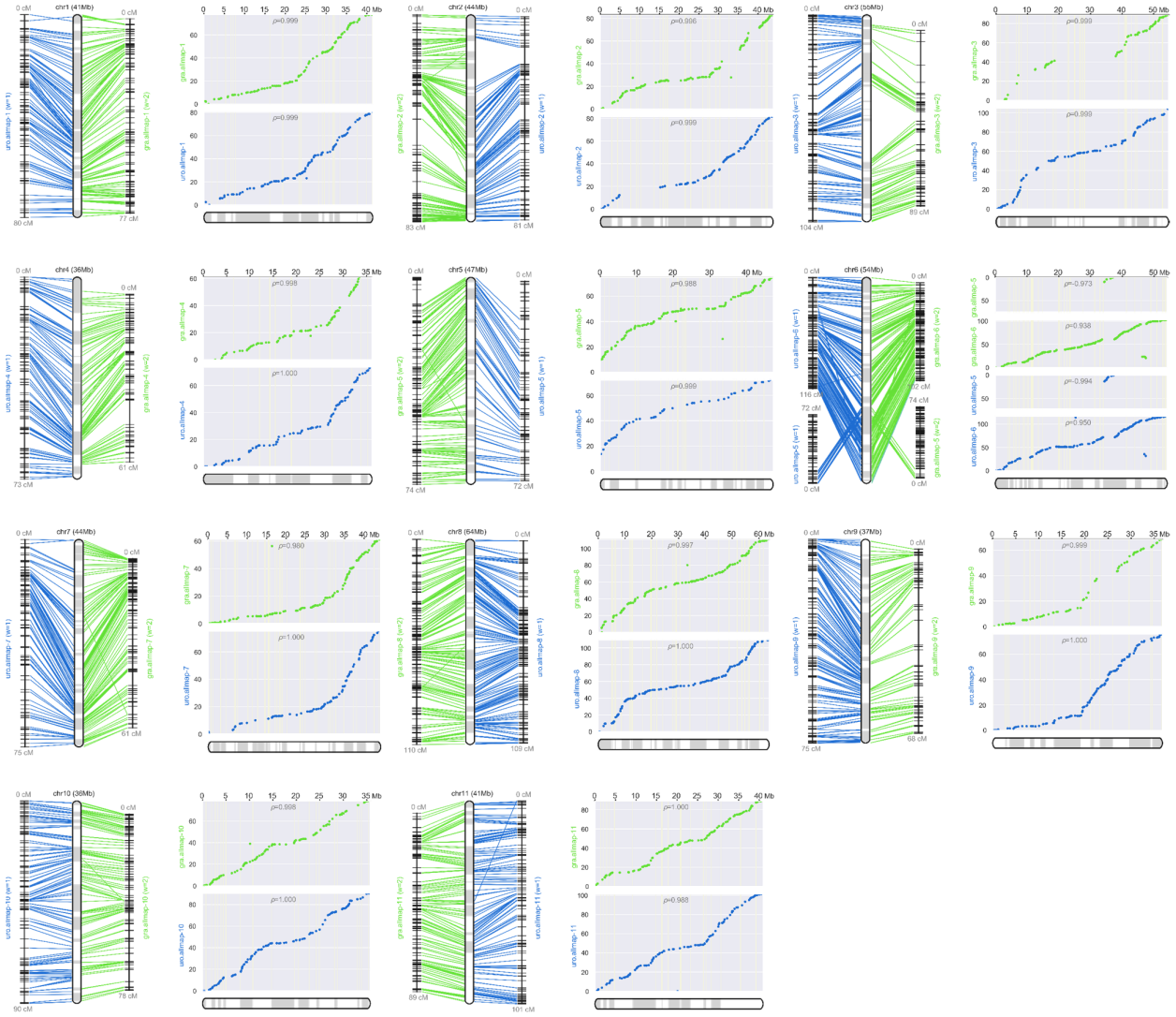
Supplementary Figure 2.2 Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness scores for both haplogenome assemblies as well as the currently available *E. grandis* v2.0 reference genome. A set of 1,614 embryophyte gene groups were used to calculate completeness of assembled genome. The bar indicates the percentage of genes belonging to categories as indicated by colour. The number of gene groups that are present (S - complete and single-copy, D - complete and duplicate-copy or F - fragmented) or absent (M - missing) are indicated by the numbers within the bar.



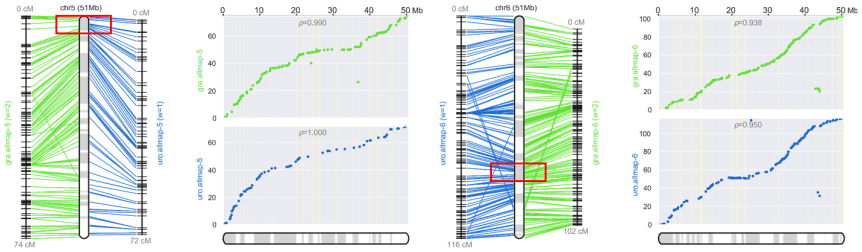
Supplementary Figure 2.3 Alignment of placed haplogenome scaffolds to the *E. grandis* v2.0 reference genome. Alignments are shown for the *E. grandis* scaffolded haplogenome (y-axis) against the *E. grandis* v2.0 reference genome (x-axis) on the left and the *E. urophylla* scaffolded assembly (y-axis) against the *E. grandis* v2.0 reference genome (x-axis) on the right and is arranged by chromosome (from one to eleven). Alignment size is measured in megabases (M).



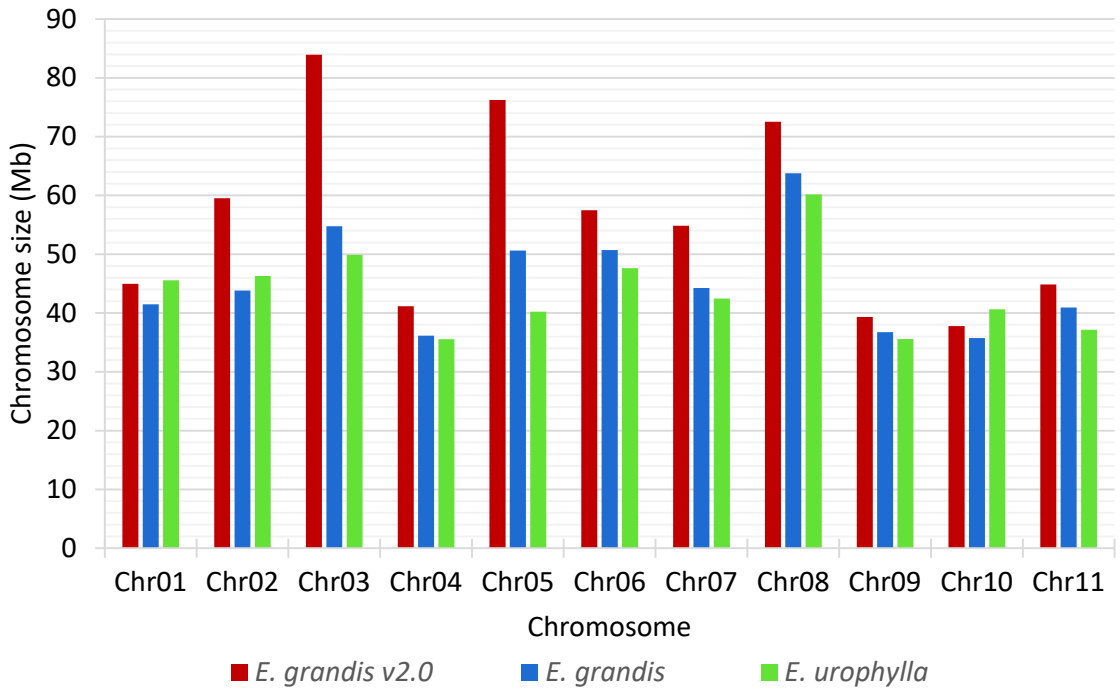
Supplementary Figure 2.4 Pseudochromosomes of *E. urophylla* haplogenome, reconstructed from two genetic linkage input maps – uro.allmap and gra.allmap, with unequal weights (2 and 1 respectively). The left-hand panels for each chromosome represent CMAP-style presentation with lines connecting physical positions on the reconstructed chromosomes and genetic map positions of SNP markers used. Boxes alternating between grey and white in the CMAP-representations represent alternating scaffolds within the reconstructed chromosomes and mark scaffold boundaries. The right-hand panel has a set of two scatter plot, where dots on the x-axis represent the physical position on the chromosomes and the y-axis the map location for the *E. urophylla* (blue) and *E. grandis* (green). Pearson's correlation coefficient is indicated as the ρ-value, and values range from -1 to 1 (where values closer to -1 and 1 indicates near-perfect collinearity).



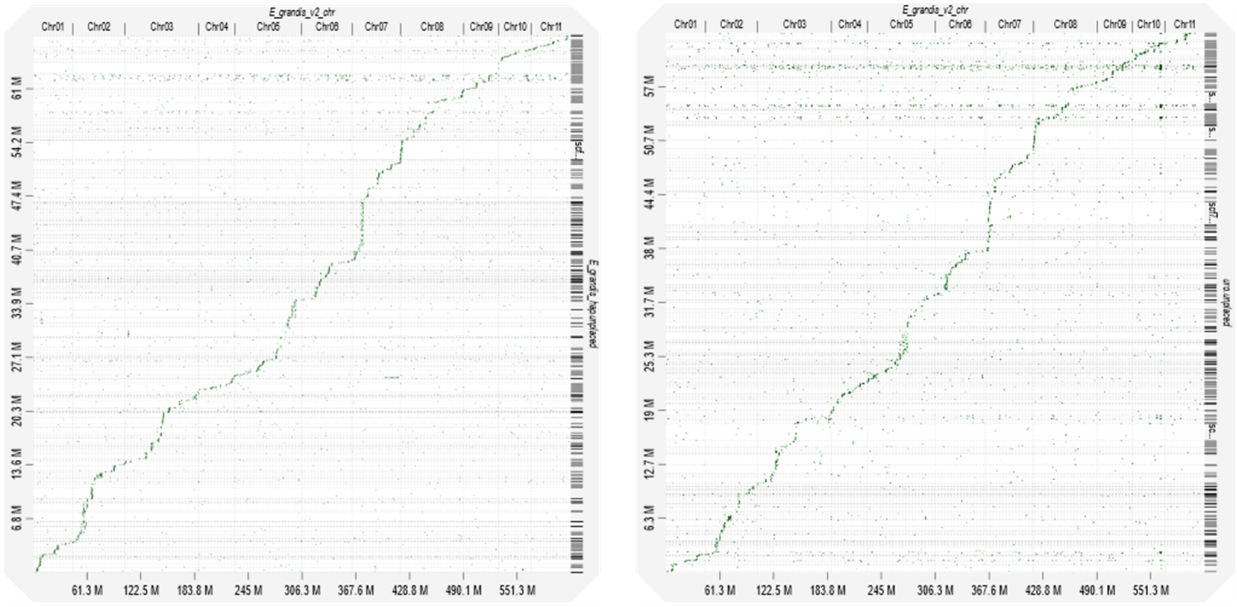
Supplementary Figure 2.5 Pseudochromosomes of *E. grandis* haplogenome, reconstructed from two genetic linkage input maps – gra.allmap and uro.allmap, with unequal weights (2 and 1 respectively). The left-hand panels for each chromosome represent CMAP-style presentation with lines connecting physical positions on the reconstructed chromosomes and genetic map positions of SNP markers used. Boxes of alternating shades represent alternating scaffolds within the reconstructed chromosomes and mark scaffold boundaries. The right-hand panel has a set of two scatter plots, where dots on the x-axis represent the physical position on the chromosomes and the y-axis the map location for *E. urophylla* (blue) and *E. grandis* (green). Pearson's correlation coefficient is also indicated (ρ-value), and values range from -1 to 1 (where values closer to -1 and 1 indicates near-perfect collinearity).



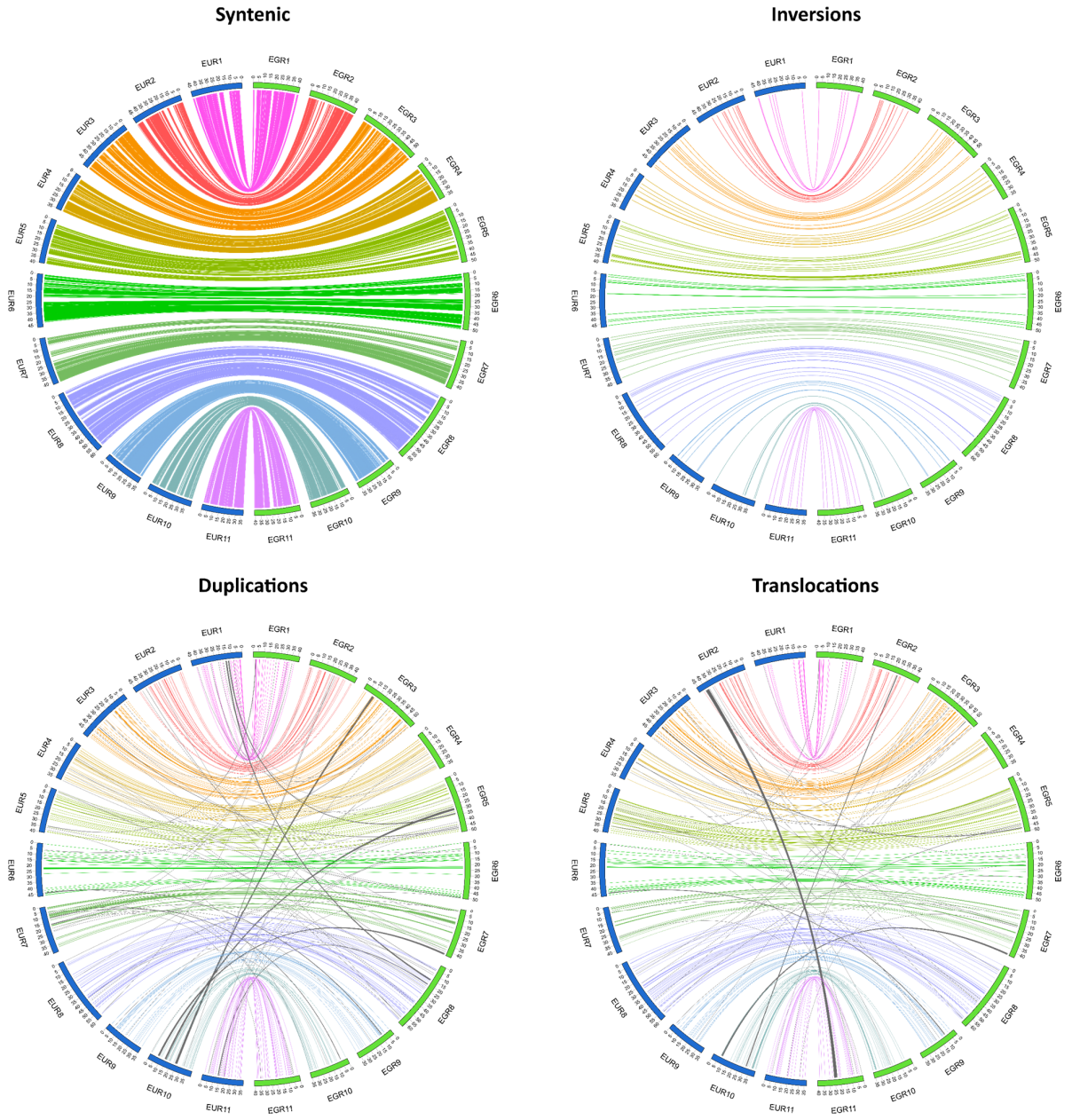
Supplementary Figure 2.6 Corrected pseudochromosomes five and six of the *E. grandis* haplogenome, reconstructed from two genetic linkage input maps – gra.allmap and uro.allmap, with unequal weights (2 and 1 respectively). The left-hand panels for both chromosomes represent CMAP-style presentation with lines connecting physical positions on the reconstructed chromosomes and genetic map positions of SNP markers used. The right-hand panel has a set of two scatter plots, where dots on the x-axis represent the physical position on the chromosomes and the y-axis the map location. The red block indicates the position of the broken contig. Boxes of alternating shades represent alternating scaffolds within the reconstructed chromosomes and mark scaffold boundaries. Pearson's correlation coefficient is also indicated with the ρ-value, and values range from -1 to 1 (values closer to -1 and 1 indicate greater collinearity).



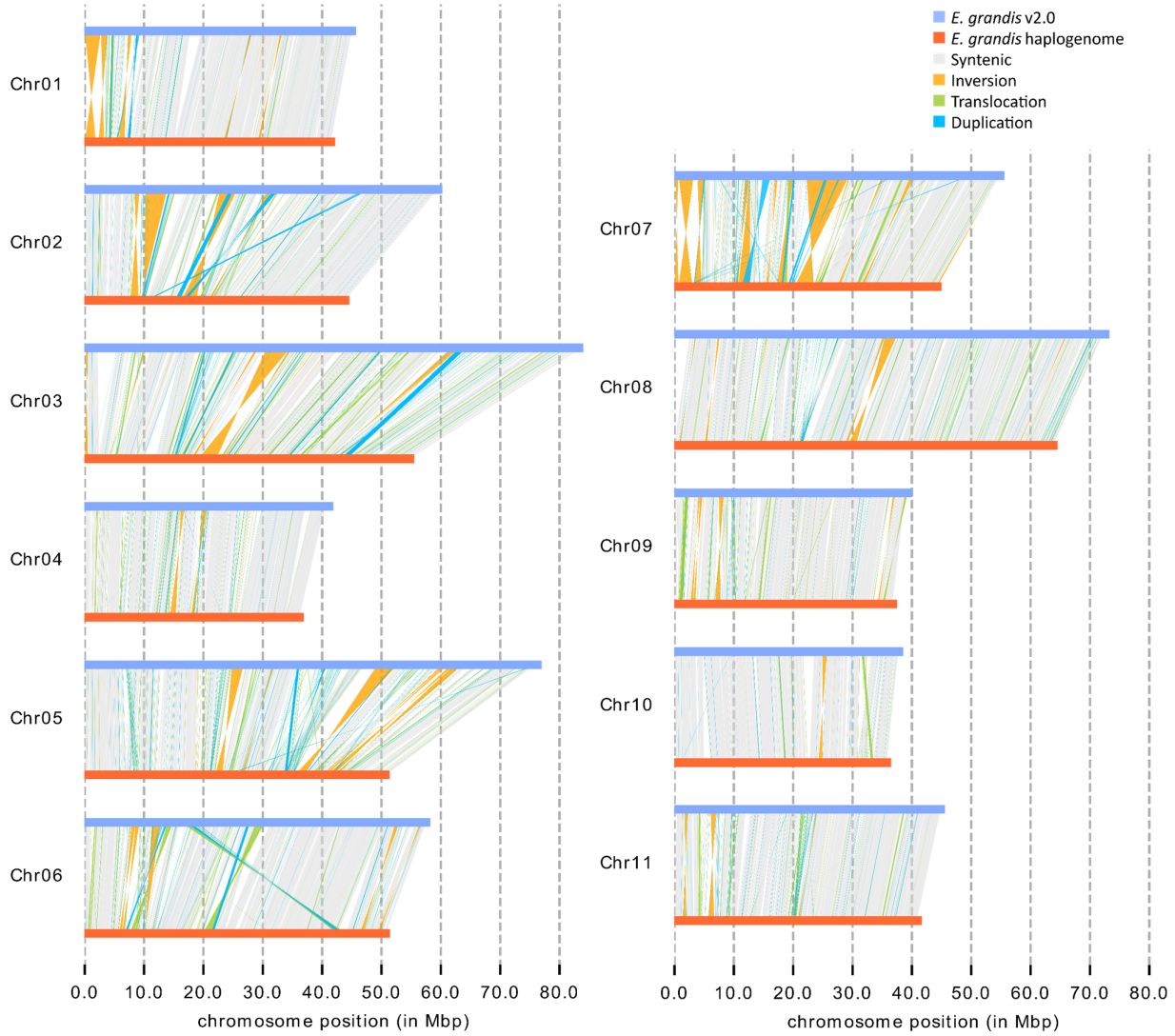
Supplementary Figure 2.7 Scaffolded chromosome sizes of the *E. grandis* v2.0 and the scaffolded *E. grandis* and *E. urophylla* haplogenome assemblies.



Supplementary Figure 2.8 Alignment of unplaced *E. grandis* and *E. urophylla* haplogenome scaffolds to the *E. grandis* v2.0 reference genome. Alignments are shown for unplaced *E. grandis* scaffolds (y-axis) against the *E. grandis* v2.0 reference genome (x-axis) on the left and unplaced *E. urophylla* scaffolds (y-axis) against the *E. grandis* v2.0 reference genome (x-axis) on the right. Alignments are arranged by chromosome (from one to eleven) and alignment size is measured in megabases (M).



Supplementary Figure 2.9 Distribution of syntenic regions and structural variants between the *E. grandis* and *E. urophylla* haplogenome assemblies. The top left-hand plot shows syntenic regions, top right-hand shows inversions, bottom left-hand plot shows duplications, and the bottom right-hand plot shows translocations between the *E. urophylla* (EUR) and *E. grandis* (EGR) chromosomes. Only variants of greater than 10 kilobases are shown as identified by SyRI. Grey links indicate rearrangements between non-homologous chromosomes, while coloured links are rearrangements between homologous chromosomes.



Supplementary Figure 2.10 Syntenic and rearranged regions between the *E. grandis* v2.0 and *E. grandis* haplogenome for all eleven chromosomes. Regions were identified using the Synteny and Rearrangement Identifier (SyRI), with a minimum rearrangement size of 100 base pairs. Chromosome number is indicated on the y-axis, while chromosome position is shown on the x-axis in megabase-pairs (Mbp). The reference genome is the *E. grandis* v2.0 reference genome (blue) and the query genome is the *E. grandis* haplogenome (orange). Syntenic regions are indicated by grey lines, whereas insertions and deletions appear as white gaps inbetween syntenic regions on the reference or query genome side, respectively. Green areas indicate translocations, yellow-orange indicate inversions and light blue translocations.

2.12. Supplementary Notes

Supplementary Note 2.1: Variation in genome size based on k-mer analyses in *E. grandis*, *E. urophylla* and *E. dunnii*

Our results showed that genome size estimates based on k-mer analysis of the F_1 hybrid and the *E. urophylla* and *E. grandis* parents (477.76 Mb, 443.19 Mb and 482.27 Mb, respectively) were smaller than previous size estimates of 650 Mb based on flow cytometry (Grattapaglia & Bradshaw Jr, 1994). To further investigate whether the estimated genome size based on k-mer analysis was within the size range for *E. grandis*, *E. urophylla* and *E. grandis* x *E. urophylla* (GU) F_1 hybrids, we performed k-mer based genome size estimation using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) for 21-mers and visualized with GenomeScope1.0 (Vurture *et al.*, 2017) for a number of Illumina whole genome sequencing datasets produced previously in our laboratory for individuals of key eucalypt tree species, both from unimproved and improved material (unpublished results). We compared estimates for 25 *E. grandis*, 19 *E. dunnii*, three *E. urophylla*, four GU F_1 hybrid and one F_1 GU x *E. urophylla* (GU x U) backcross individual. Most of these samples' sequencing data consisted of 100 bp (PE100) Illumina sequencing reads and had only 19 – 40x genome coverage, whereas those from this study and other individuals within this NAM population had a read length of 150 bp (PE150) and at least 124x coverage (Supplementary Table 2.9). As GenomeScope2.0 requires at least 15x coverage per set of homologous chromosomes to accurately infer diploid genome size (Ranallo-Benavidez *et al.*, 2020), and GenomeScope1.0 allows the user to select the read length while assuming a diploid organism (which is not possible in GenomeScope2.0), we used GenomeScope1.0 to compare genome size and heterozygosity estimates for all the above individuals with all other parameters unchanged.

Genome size estimates ranged from 428.67 Mb to 559.73 Mb for *E. grandis* (Supplementary Table 2.9 and Supplementary Figure 2.11A). Our *E. grandis* parent (FK1758) had the lowest size estimate (428.67 Mb, Supplementary Table 2.9 and Supplementary Figure 2.11A). In comparison, genome size estimates of *E. urophylla* ranged from 412.29 Mb to 485.06 Mb, with our parent (FK1756) again having the lowest estimate at 412.29 Mb (Supplementary Table 2.9 and Supplementary Figure 2.11A), but this could be

due to the high genome coverage as all samples with high coverage had lower genome size estimates. Lastly, genome size for *E. dunnii* ranged from 457.34 Mb to 497.11 Mb, with an average of 476.29 Mb (Supplementary Table 2.9 and Supplementary Figure 2.11A). This is unexpected as the estimated genome size of *E. dunnii* based on flow cytometry is 530 Mb, which is 120 Mb smaller than that of *E. urophylla* at 650 Mb and *E. grandis* at 640 Mb (Grattapaglia & Bradshaw Jr., 1994). The size difference between k-mer based and flow-cytometry based genome size estimates may be partially explained by the fact that repetitive genome content is not fully represented in k-mer based estimates as the maximum k-mer coverage parameter is set to 1,000, which will result in k-mer exclusion of highly repetitive elements when genome size is estimated. The 20 Mb difference seen between the average haploid genome size estimates of *E. urophylla* and *E. grandis* vs *E. dunnii* consists mostly of the repeat length portion of the total haploid genome size estimates (Supplementary Table 2.9 and Supplementary Figure 2.11B). This is as expected as previous studies on the cause of genome size variation in *Eucalyptus* has been attributed to a difference in the repetitive content between species (Myburg *et al.*, 2014). In addition, GenomeScope genome size estimates are known to be affected by the repeat content of the genome (Vurture *et al.*, 2017). As *Eucalyptus* genomes have high repeat content (44.5 – 49%, Table 2.1 and Table 2.3, Wang *et al.*, 2020), it is very likely that the repetitive portion of the genome is underrepresented in k-mer based genome size estimates and this may contribute to the lower-than-expected genome size estimates.

As all k-mer based estimates for samples with more than 100x genome coverage were smaller than those with 19 – 40x genome coverage (the overall average haploid size estimate for full set of sequencing data at max k-mer coverage = 1,000 was 427.39 Mb and 475.16 Mb max k-mer coverage of 10,000, Supplementary Table 2.9 and Supplementary Figure 2.11A), we further investigated whether sequencing depth influences k-mer genome size estimates produced by GenomeScope. To test the effect of sequencing depth on k-mer based genome size estimates, we performed GenomeScope analysis on a subset of 25 Gb of the total sequencing data, relating to 38.5x genome coverage per sample. Using the

seqtk v1.2 (seqtk, Toolkit for processing sequences in FASTA/Q formats), a subset of 25 Gb of paired reads were created and used to estimate genome size based on k-mers. GenomeScope results only converged for five of the eight samples for which a subset of reads was used (sample names which have k-mer based genome size estimates are shown with a _sub in Supplementary Table 2.9 and Supplementary Figure 2.11C, average haploid genome size estimate per species is given in the legend of Supplementary Figure 2.11C). The overall average haploid size estimate for the five converged samples was 426.95 Mb and 475.11 Mb (at max k-mer coverage 1,000 and 10,000) compared to 501.56 Mb and 540.76 Mb respectively, indicating that a lower amount of genome coverage increases the genome size estimates.

Genome heterozygosity estimates ranged from 1.62% to 2.38% for *E. grandis* (average 2.02%), 1.88% to 2.06% for *E. dunnii* (average 1.96%) and 2.41% to 2.72% for *E. urophylla* (average 2.62%, Supplementary Table 2.9, Supplementary Figure 2.11B) at max k-mer coverage of 1,000. Increasing the max k-mer coverage parameter to 10,000 has no effect on the heterozygosity estimates, however the subset data have a slightly lower estimated heterozygosity (Supplementary Table 2.9 and Supplementary Figure 2.11D). The higher average heterozygosity estimates observed for *E. urophylla* could be a result of cryptic hybridization that has occurred between *E. urophylla* and *E. alba* within their natural range before *E. urophylla* selections were made (Dvorak *et al.*, 2008) and likely reflects the hybrid nature of *E. urophylla* itself. The estimated heterozygosity of the *E. grandis* (2.14%, FK1758) and *E. urophylla* (2.72%, FK1756) parents used for trio-binning and genome assembly falls within the observed range of heterozygosity for both species (Supplementary Figure 2.1, Supplementary Table 2.9 and Supplementary Figure 2.11). As expected, all F_1 and F_2 hybrids had higher heterozygosity estimated compared to pure species. This supports that estimated genome heterozygosity falls within the expected range for all sample used in this study (FK1756, FK1758 and FK118).

Supplementary Table 2.9 GenomeScope1.0 analysis of genome size and heterozygosity. A tabular summary is given for the minimum (min) and maximum (ma) and average (avg) estimated heterozygosity (htz). Haploid genome size estimates, as well as the repeat and unique component of each haploid genome size estimate in base pairs (bp) are shown per sample. The species represented are *E. grandis* (GRA), *E. dunnii* (DUN), *E. urophylla* (URO), F_1 *E. grandis* x *E. urophylla* hybrids (GU) or F_2 GU x *E. urophylla* (GUxU) hybrids. The amount of sequencing data is given in gigabases (Gb), followed by the estimated genome coverage and whether the data used for genome size estimates was 100 or 150 bp PE Illumina sequencing data. All results are shown for when the max k-mer coverage parameter was set to 1,000 and 10,000.

Sample[a]	Min Htz	Max Htz	Avg Htz[b]	Min Repeat Length	Max Repeat Length[b]	Min Unique Length	Max Unique Length[b]	Min Haploid Length	Max Haploid Length	Species[c]	Gb Sequencing Data	Coverage[d]	100 or 150 bp PE
Max k-mer coverage = 1,000													
AP928	1.98	2.01	2.00	178,530,441	179,526,157	288,708,550	290,318,760	467,238,992	469,844,917	GRA	12.58	19.35	100
AP929	1.95	1.97	1.96	176,807,887	177,622,483	292,590,543	293,938,577	469,398,430	471,561,060	GRA	13.23	20.35	100
AP923	1.98	2.01	2.00	181,474,488	182,629,403	290,119,013	291,965,349	471,593,501	474,594,752	GRA	12.54	19.29	100
AP924	1.91	1.94	1.93	180,795,937	181,665,066	293,703,297	295,115,198	474,499,234	476,780,264	GRA	13.14	20.22	100
AP932	2.09	2.12	2.11	183,272,604	184,342,504	296,432,554	298,163,054	479,705,157	482,505,557	GRA	12.61	19.40	100
AP926	2.18	2.22	2.20	188,252,303	189,867,411	291,343,629	293,843,207	479,595,932	483,710,618	GRA	13.18	20.28	100
AP921	2.03	2.07	2.05	187,281,436	188,854,217	295,482,626	297,964,074	482,764,062	486,818,291	GRA	12.68	19.51	100
AP927	2.20	2.25	2.23	191,663,235	193,569,948	291,640,802	294,542,116	483,304,036	488,112,064	GRA	13.04	20.06	100
AP931	1.84	1.86	1.85	185,700,604	186,558,524	304,364,426	305,770,561	490,065,030	492,329,085	GRA	13.40	20.62	100
AP922	1.98	2.01	2.00	189,719,722	191,083,320	300,311,786	302,470,257	490,031,509	493,553,577	GRA	12.89	19.83	100
AP925	2.19	2.24	2.22	190,809,144	192,692,517	298,720,330	301,668,835	489,529,474	494,361,351	GRA	12.89	19.83	100
AP962	2.11	2.14	2.13	193,762,556	194,945,739	308,052,071	309,933,146	501,814,627	504,878,885	GRA	16.00	24.62	100
AP966	1.97	2.00	1.99	195,779,968	196,803,140	309,369,865	310,986,673	505,149,833	507,789,813	GRA	16.00	24.62	100
AP967	1.95	1.98	1.97	197,942,269	199,137,543	311,133,338	313,012,116	509,075,607	512,149,659	GRA	16.00	24.62	100
AP939	2.18	2.28	2.23	181,537,599	184,531,989	322,342,105	327,659,008	503,879,704	512,190,997	GRA	13.19	20.29	100
AP959	1.66	1.68	1.67	194,571,440	195,492,002	318,398,240	319,904,655	512,969,680	515,396,657	GRA	16.00	24.62	100
AP968	2.00	2.03	2.02	201,219,079	202,587,245	311,398,110	313,515,427	512,617,189	516,102,672	GRA	16.00	24.62	100
AP965	1.93	1.96	1.95	196,862,541	198,271,470	317,941,130	320,216,608	514,803,671	518,488,077	GRA	16.00	24.62	100
AP964	2.02	2.06	2.04	199,263,655	200,983,187	314,967,194	317,685,182	514,230,849	518,668,369	GRA	16.00	24.62	100
AP960	1.90	1.93	1.92	197,709,891	198,907,333	320,620,644	322,562,503	518,330,534	521,469,836	GRA	16.00	24.62	100

AP930	1.97	2.09	2.03	229,384,353	235,118,103	316,693,025	324,609,164	546,077,378	559,727,267	GRA	13.24	20.37	100
H1701	1.61	1.63	1.62	186,256,306	186,965,570	305,089,118	306,250,897	491,345,424	493,216,467	GRA	16.00	24.62	100
P1381	1.90	1.93	1.92	192,551,941	193,741,384	304,404,458	306,284,843	496,956,399	500,026,227	GRA	15.00	23.08	100
FK1752	2.36	2.39	2.38	165,086,760	166,062,241	283,057,089	284,729,646	448,143,849	450,791,887	GRA	95.00	146.15	150
FK1752_Sub	2.05	2.11	2.08	512,242,422	518,899,072	206,828,540	209,516,301	305,413,882	309,382,771	GRA	25.00	38.46	150
FK1758	2.13	2.14	2.14	140,308,820	140,620,254	287,415,212	288,053,168	427,724,031	428,673,422	GRA	141.60	217.85	150
A0380	3.08	3.13	3.11	196,431,062	198,074,773	297,681,938	300,172,903	494,113,000	498,247,676	GU	15.00	23.08	100
FK1753	2.57	2.59	2.58	145,811,888	146,335,947	282,507,168	283,522,521	428,319,056	429,858,467	GU	81.00	124.62	150
FK1753_Sub	2.30	2.37	2.34	501,672,693	508,422,702	199,406,060	202,089,070	302,266,633	306,333,632	GU	25.00	37.59	150
FK118	3.53	3.61	3.57	135,172,642	135,846,324	287,922,031	289,356,996	423,094,672	425,203,321	GU	116.10	178.62	150
FK118_Sub	3.17	3.35	3.26	514,502,664	530,931,902	206,814,590	213,418,650	307,688,074	317,513,253	GU	25.00	38.46	150
NN2868	4.45	4.58	4.52	156,444,738	157,221,647	269,724,745	271,064,206	426,169,482	428,285,853	GU	81.00	124.62	150
NN0784	4.37	4.54	4.46	136,912,452	137,891,758	275,187,764	277,156,125	412,100,216	415,047,884	GUxU	79.00	131.67	150
M1459	2.38	2.44	2.41	168,701,733	170,398,844	311,525,699	314,659,596	480,227,431	485,058,440	URO	16.00	24.62	100
FK1755	2.69	2.74	2.72	116,366,332	116,869,135	298,438,458	299,727,970	414,804,790	416,597,105	URO	150.80	232.00	150
FK1755_Sub	2.60	2.67	2.64	470,189,670	475,232,117	175,877,876	177,764,040	294,311,794	297,468,077	URO	25.00	38.46	150
FK1756	2.67	2.76	2.72	130,530,346	131,830,497	277,698,523	280,464,547	408,228,868	412,295,044	URO	127.50	196.15	150
FK1756_Sub	2.38	2.42	2.40	470,401,921	474,331,214	179,548,652	181,048,431	290,853,269	293,282,782	URO	25.00	38.46	150
BV174	1.97	1.99	1.98	161,332,242	161,947,419	296,011,432	297,140,155	457,343,673	459,087,574	DUN	22.00	41.51	100
BV143	2.04	2.07	2.06	166,159,841	166,858,313	293,539,579	294,773,506	459,699,420	461,631,819	DUN	18.00	33.96	100
BV164	2.00	2.03	2.02	164,786,880	165,584,602	299,149,025	300,597,184	463,935,905	466,181,786	DUN	19.00	35.85	100
BV157	2.03	2.05	2.04	165,155,949	165,833,648	301,403,086	302,639,861	466,559,035	468,473,509	DUN	20.00	37.74	100
BV100	1.97	1.99	1.98	169,088,050	169,751,445	298,005,431	299,174,617	467,093,481	468,926,062	DUN	19.00	35.85	100
BV139	1.90	1.93	1.92	170,088,615	170,825,411	298,151,430	299,442,973	468,240,045	470,268,384	DUN	17.00	32.08	100
BV170	1.91	1.95	1.93	170,845,007	171,592,469	299,895,147	301,207,215	470,740,154	472,799,684	DUN	16.00	30.19	100
BV175	1.93	1.95	1.94	173,583,891	174,314,807	297,415,001	298,667,338	470,998,892	472,982,145	DUN	16.00	30.19	100
BV138	1.87	1.89	1.88	175,119,058	175,935,820	299,766,255	301,164,376	474,885,313	477,100,197	DUN	17.00	32.08	100
BH1697	1.96	1.98	1.97	175,853,495	176,643,245	300,622,956	301,973,041	476,476,451	478,616,287	DUN	16.00	30.19	100
BH1477	1.94	1.96	1.95	174,352,710	175,178,023	305,945,891	307,394,113	480,298,601	482,572,136	DUN	16.00	30.19	100
BV155	1.98	2.01	2.00	173,808,004	174,785,873	306,218,954	307,941,785	480,026,959	482,727,658	DUN	20.00	37.74	100
BH528	1.87	1.89	1.88	178,367,761	179,130,639	308,247,268	309,565,640	486,615,029	488,696,279	DUN	15.00	28.30	100
BH762	1.92	1.94	1.93	184,922,908	185,827,631	306,798,518	308,299,508	491,721,426	494,127,138	DUN	15.00	28.30	100

BH840	1.93	1.96	1.95	185,670,731	186,816,397	311,440,037	313,361,752	497,110,768	500,178,149	DUN	15.00	28.30	100
Max k-mer coverage = 10,000													
AP928	1.99	2.00	2.00	215,873,340	216,261,375	289,251,637	289,771,571	505,124,977	506,032,946	GRA	12.58	19.35	100
AP929	1.96	1.97	1.97	222,844,114	223,170,212	293,048,752	293,477,584	515,892,865	516,647,796	GRA	13.23	20.35	100
AP923	1.99	2.00	2.00	227,735,079	228,197,322	290,744,481	291,334,617	518,479,560	519,531,938	GRA	12.54	19.29	100
AP924	1.92	1.93	1.93	212,325,815	212,651,393	294,182,183	294,633,278	506,507,998	507,284,671	GRA	13.14	20.22	100
AP932	2.10	2.11	2.11	230,257,508	230,684,423	297,020,187	297,570,884	527,277,695	528,255,307	GRA	12.61	19.40	100
AP926	2.19	2.21	2.20	239,950,656	240,603,139	292,191,356	292,985,895	532,142,012	533,589,034	GRA	13.18	20.28	100
AP921	2.04	2.06	2.05	231,275,256	231,894,090	296,322,255	297,115,138	527,597,511	529,009,227	GRA	12.68	19.51	100
AP927	2.22	2.24	2.23	231,937,564	232,672,009	292,621,711	293,548,316	524,559,275	526,220,325	GRA	13.04	20.06	100
AP931	1.84	1.85	1.85	229,035,061	229,372,666	304,841,364	305,290,710	533,876,425	534,663,375	GRA	13.40	20.62	100
AP922	1.99	2.00	2.00	235,137,409	235,676,889	301,042,215	301,732,901	536,179,625	537,409,790	GRA	12.89	19.83	100
AP925	2.21	2.22	2.22	220,612,280	221,310,232	299,713,986	300,662,191	520,326,266	521,972,423	GRA	12.89	19.83	100
AP962	2.12	2.13	2.13	258,409,826	258,905,598	308,693,908	309,286,152	567,103,734	568,191,749	GRA	16.00	24.62	100
AP966	1.98	1.99	1.99	255,594,426	256,018,147	309,919,480	310,433,261	565,513,906	566,451,408	GRA	16.00	24.62	100
AP967	1.96	1.97	1.97	266,021,735	266,525,994	311,774,686	312,365,673	577,796,421	578,891,667	GRA	16.00	24.62	100
AP939	2.21	2.24	2.23	233,619,559	234,839,641	324,134,613	325,827,412	557,754,172	560,667,054	GRA	13.19	20.29	100
AP959	1.67	1.68	1.68	263,092,800	263,483,698	318,912,924	319,386,759	582,005,725	582,870,457	GRA	16.00	24.62	100
AP968	2.01	2.02	2.02	268,936,911	269,510,725	312,120,552	312,786,504	581,057,462	582,297,230	GRA	16.00	24.62	100
AP965	1.94	1.95	1.95	266,159,937	266,758,786	318,716,659	319,433,759	584,876,596	586,192,545	GRA	16.00	24.62	100
AP964	2.03	2.05	2.04	252,885,709	253,585,723	315,883,753	316,758,152	568,769,462	570,343,876	GRA	16.00	24.62	100
AP960	1.91	1.92	1.92	251,760,864	252,241,181	321,282,463	321,895,414	573,043,327	574,136,595	GRA	16.00	24.62	100
AP930	2.01	2.05	2.03	268,107,423	270,239,610	319,337,465	321,877,070	587,444,888	592,116,680	GRA	13.24	20.37	100
H1701	1.61	1.62	1.62	236,117,036	236,403,979	305,483,402	305,854,643	541,600,438	542,258,621	GRA	16.00	24.62	100
P1381	1.91	1.92	1.92	248,124,388	248,611,302	305,042,747	305,641,356	553,167,135	554,252,658	GRA	15.00	23.08	100
FK1752	2.37	2.38	2.38	208,484,320	208,870,524	283,628,459	284,153,862	492,112,779	493,024,386	GRA	95.00	146.15	150
FK1752_Sub	2.08	2.10	2.09	251,224,976	252,228,924	307,556,281	308,785,340	558,781,257	561,014,264	GRA	25.00	38.46	150
FK1758	2.13	2.14	2.14	193,253,234	193,386,985	287,634,331	287,833,404	480,887,566	481,220,389	GRA	141.60	217.85	150
A0380	3.10	3.11	3.11	246,943,582	247,598,398	298,526,956	299,318,555	545,470,538	546,916,953	GU	15.00	23.08	100
FK1753	2.58	2.58	2.58	199,691,198	199,915,690	282,855,065	283,173,050	482,546,262	483,088,740	GU	81.00	124.62	150
FK1753_Sub	2.34	2.35	2.35	247,002,906	247,954,946	304,138,168	305,310,429	551,141,075	553,265,375	GU	25.00	37.59	150
FK118	3.55	3.58	3.57	185,997,154	186,286,636	288,413,463	288,862,343	474,410,617	475,148,979	GU	116.10	178.62	150

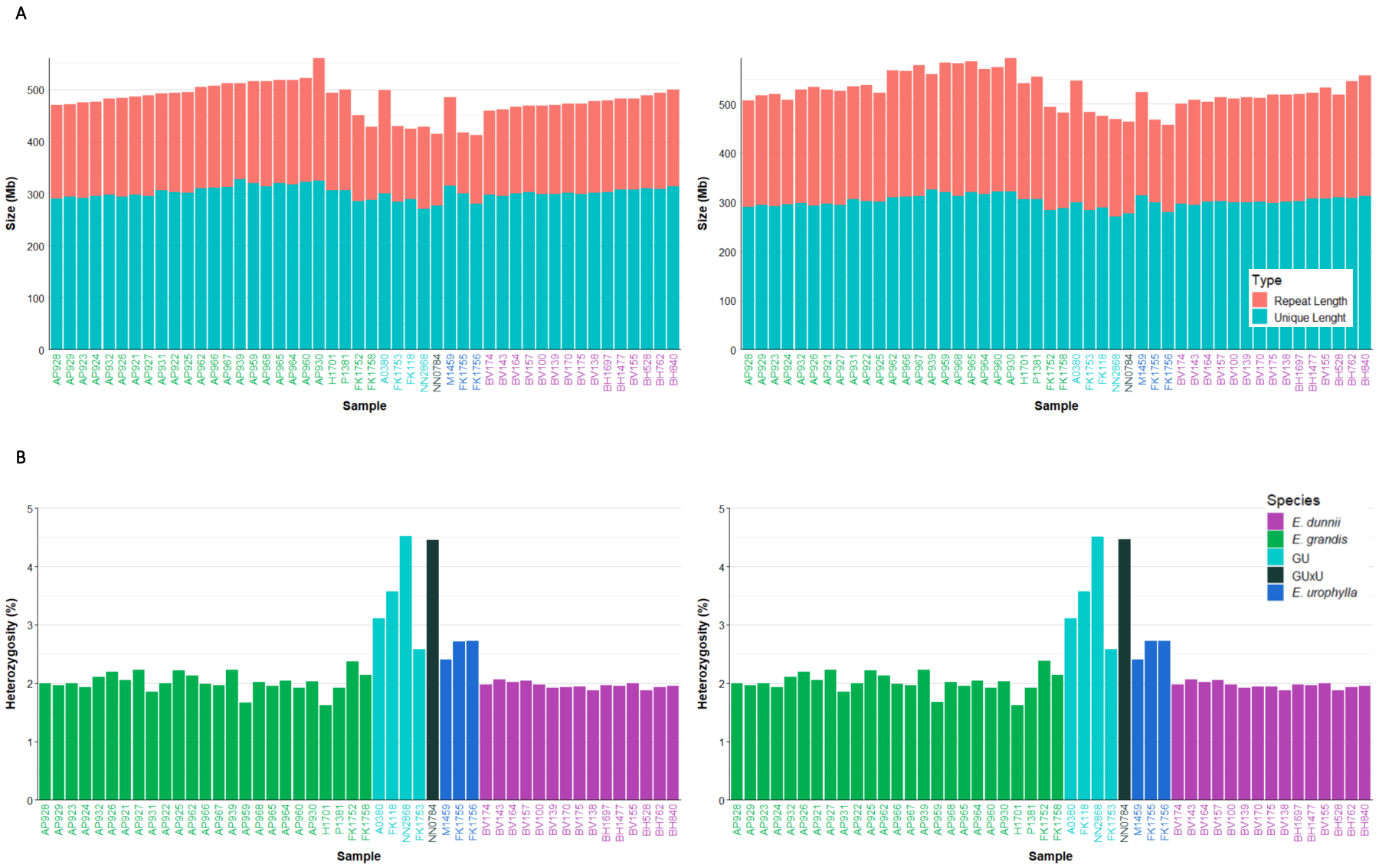
FK118_Sub	3.43	3.46	3.45	239,961,238	240,593,122	307,153,894	307,962,715	547,115,131	548,555,837	GU	25.00	38.46	150
NN2868	4.49	4.53	4.51	198,403,547	198,712,380	270,182,659	270,603,221	468,586,206	469,315,601	GU	81.00	124.62	150
NN0784	4.43	4.48	4.46	186,758,388	187,176,570	275,859,770	276,477,465	462,618,158	463,654,036	GUxU	79.00	131.67	150
M1459	2.40	2.42	2.41	209,652,672	210,323,242	312,585,618	313,585,416	522,238,290	523,908,657	URO	16.00	24.62	100
FK1755	2.71	2.72	2.72	167,922,428	168,148,268	298,881,055	299,283,021	466,803,482	467,431,289	URO	150.80	232.00	150
FK1755_Sub	2.63	2.65	2.64	223,769,777	224,385,769	295,502,968	296,316,427	519,272,746	520,702,196	URO	25.00	38.46	150
FK1756	2.70	2.73	2.72	176,801,589	177,350,821	278,642,552	279,508,152	455,444,141	456,858,973	URO	127.50	196.15	150
FK1756_Sub	2.39	2.40	2.40	226,898,732	227,424,843	292,152,497	292,829,913	519,051,229	520,254,756	URO	25.00	38.46	150
BV174	1.98	1.98	1.98	203,210,459	203,454,828	296,396,614	296,753,043	499,607,073	500,207,871	DUN	22.00	41.51	100
BV143	2.05	2.06	2.06	213,706,814	213,988,642	293,961,549	294,349,215	507,668,363	508,337,857	DUN	18.00	33.96	100
BV164	2.01	2.02	2.02	203,543,222	203,857,244	299,640,400	300,102,679	503,183,622	503,959,923	DUN	19.00	35.85	100
BV157	2.04	2.05	2.05	209,943,985	210,215,714	301,825,009	302,215,659	511,768,994	512,431,373	DUN	20.00	37.74	100
BV100	1.97	1.98	1.98	210,919,606	211,182,215	298,403,225	298,774,758	509,322,831	509,956,973	DUN	19.00	35.85	100
BV139	1.91	1.92	1.92	213,787,979	214,080,359	298,591,677	299,000,037	512,379,656	513,080,397	DUN	17.00	32.08	100
BV170	1.93	1.94	1.94	211,359,254	211,652,963	300,341,154	300,758,515	511,700,407	512,411,478	DUN	16.00	30.19	100
BV175	1.94	1.94	1.94	219,345,907	219,639,078	297,840,948	298,239,032	517,186,855	517,878,110	DUN	16.00	30.19	100
BV138	1.88	1.88	1.88	216,654,640	216,976,510	300,240,837	300,686,885	516,895,477	517,663,394	DUN	17.00	32.08	100
BH1697	1.97	1.98	1.98	217,796,724	218,138,596	301,222,834	301,695,659	519,019,557	519,834,254	DUN	16.00	30.19	100
BH1477	1.95	1.96	1.96	215,220,224	215,546,086	306,436,534	306,900,506	521,656,758	522,446,592	DUN	16.00	30.19	100
BV155	1.99	2.00	2.00	225,305,406	225,703,296	306,807,283	307,349,106	532,112,689	533,052,402	DUN	20.00	37.74	100
BH528	1.88	1.88	1.88	209,288,826	209,575,710	308,693,616	309,116,759	517,982,442	518,692,468	DUN	15.00	28.30	100
BH762	1.93	1.93	1.93	237,038,050	237,406,090	307,308,790	307,785,938	544,346,840	545,192,028	DUN	15.00	28.30	100
BH840	1.94	1.95	1.95	243,812,585	244,285,176	312,095,762	312,700,710	555,908,347	556,985,886	DUN	15.00	28.30	100

[a] subset of the sequencing data was used for all samples with the Sub prefix

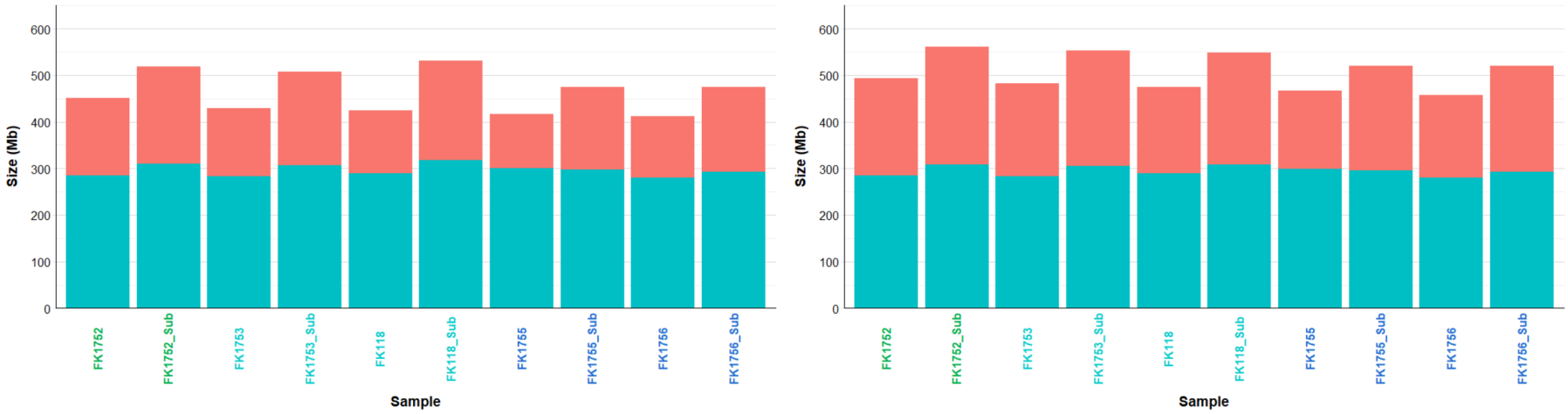
[b] Values used for Supplementary Figure 2.11

[c] GRA – *E. grandis,* DUN - *E. dunnii,* URO - *E. urophylla,* GU - F_1 *E. grandis* x *E. urophylla* hybrids or GUxU - F2 GU x *E. urophylla* hybrids

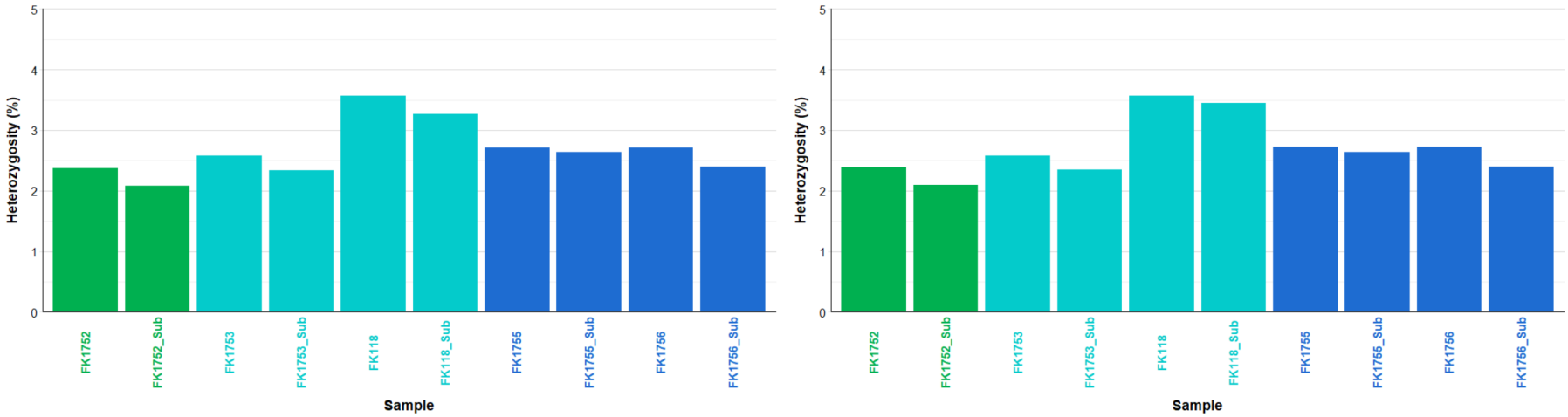
[d] Coverage is estimated based on flow cytometry genome size estimates of 530 Mb for *E. dunnii* and 650 Mb for all other species (*E. grandis, E. urophylla* and hybrids thereof)


A
Size (Mb)
Sample
Type
Repeat Length
Unique Lenght
B
Heterozygosity (%)
Species
E. dunnii
E. grandis
GU
GUxU
E. urophylla

C


Size (Mb)
0
100
200
300
400
500
600
FK1752
FK1752_Sub
FK1753
FK1753_Sub
FK118
FK118_Sub
FK1755
FK1755_Sub
FK1756
FK1756_Sub
Sample

D


Heterozygosity (%)
0
1
2
3
4
5
FK1752
FK1752_Sub
FK1753
FK1753_Sub
FK118
FK118_Sub
FK1755
FK1755_Sub
FK1756
FK1756_Sub
Sample

Supplementary Figure 2.11 K-mer based estimates of genome heterozygosity and genome size. All estimates are shown for max k-mer coverage at 1,000 (left) and 10,000 (right). **(A)** Estimated haploid genome size per species. The x-axis shows sample names, which are coloured according to species (100 bp PE Illumina data is shown first, followed by samples with 150 bp PE data if any is available for the species), and the y-axis indicates the estimated haploid genome size in megabase (Mb). Bars are split into the unique (teal) and repeat (coral) components of the haploid genome size estimate. At the maximum k-mer coverage of 1,000 (left), the average haploid genome size estimates for *E. grandis, E. dunnii* and *E. urophylla* are 494.95 Mb (305.25 Mb unique), 476.29 Mb (302.89 Mb unique) and 437.98 Mb (298.28 Mb unique), respectively. In comparison, with a max k-mer coverage of 10,000 (right) the average unique/haploid genome size estimates are 304.47/543.74 Mb for *E. grandis*, 302.43/519.48 Mb for *E. dunnii* and 297.46/482.73 Mb for *E. urophylla*. **(B)** Estimated genome heterozygosity per species. The x-axis shows the sample names, grouped by species and data type (PE100 Illumina data is shown first, followed by samples with PE150 data if any is available for the species) and the y-axis indicates the percentage of heterozygosity. Bar colour indicates the species or hybrid cross. The average heterozygosity for *E. grandis, E. dunnii* and *E. urophylla* is 2.02%, 1.96% and 2.62%, respectively at max k-mer coverage of 1,000 (left) and stays the same a max k-mer coverage of 10,000 (right). **(C)** Estimated haploid genome size for PE150 samples using a subset of 25 gigabases (Gb) of randomly selected paired reads of the total sequencing data. The x-axis shows samples for which genome size estimates could be made with the original size estimate followed by that of the subset data (denoted with a Sub), which are coloured according to species, and the y-axis indicates the estimated haploid genome size in megabase (Mb). Bars are split into the unique (teal) and repeat (coral) components of the haploid genome size estimate. The average haploid genome size estimates per species for the subset data in the case of max k-mer coverage = 1,000 are 309.38/518.90 Mb (unique/total haploid genome size estimate), 295.38/474.78 Mb and 311.92/519.42 Mb for *E. grandis*, *E. urophylla* and F_1 GU respectively compared to 308.79/561.01 Mb, 294.57/520.48 Mb and 306.64/550.91 Mb when the max k-mer coverage was 10,000. **(D)** Estimated genome heterozygosity for a subset of 25 Gb of PE150 sequencing data. The x-axis shows the samples with the original heterozygosity estimate with the total amount of sequencing data followed by the estimate for the subset data (denoted as Sub), grouped by species and the y-axis indicates the percentage of heterozygosity. Bar colour indicates the species or hybrid cross. The average percentage of heterozygosity for subset data is 2.08%, 2.51% and 2.79% for *E. grandis*, *E. urophylla* and F_1 GU,

respectively for max k-mer coverage of 1,000, and was 2.09%, 2.52% and 2.9% when max k-mer coverage was 10,000. In this figure F_1 *E. grandis* x *E. urophylla* and *E. urophylla* x *E. grandis* hybrids are considered GU. All estimates are based on 21-mer analysis with GenomeScope1.0.

Supplementary Note 2.2: Hap-mer based phasing completeness assessment.

We found that separation of long-reads into haplotype bins before genome assembly resulted in splitting of the long-reads almost equally into *E. urophylla* and *E. grandis* haplotype bins (Supplementary Table 2.3). To further validate that long-reads were separated into the correct haplotype bins, and that the haplogenome assemblies contained mostly a single haplotype, we performed independent assessment of the haplotype specific k-mers contained within each haplogenome assembly and whether those correspond to the parent specific k-mers identified prior to trio-binning which was used for separation of long-reads into haplotype bins by Canu. We used Merqury v1.1 (Rhie *et al.*, 2020) to further validate whether separation of long-reads into *E. urophylla* and *E. grandis* haplotype bins was successful. Using the parental haplo-genome assemblies we could estimate the inherited hap-mers for the child (i.e. haplotype specific k-mers present in the F_1 haplogenome bins) to assess how well phased the assembled haplogenome assemblies are (Rhie *et al.*, 2020). Using the parent specific hap-mers, we determined phase blocks (a consistent set of markers originating from a single haplotype) based on observed haplotype markers within the haplogenome assemblies with Merqury. We observed a block N50 and average block size of 42.45 Mb and 491.75 kb for the *E. urophylla* haplogenome assembly (Supplementary Table 2.10). In addition, using a maximum of 100 consecutive haplotype marker switches per phase-block window of 20 kb, we showed that the *E. urophylla* haplogenome assembly had a low switch error rate of 0.033% per block (Supplementary Table 2.10). In comparison, the *E. grandis* haplogenome assembly had a slightly larger block N50 size (43.82 Mb) and a smaller average block size (432.93 kb) with a lower switch error rate of 0.028% (Supplementary Table 2.10). As some short-range switches may be missed when allowing 100 consecutive switches per 20 kb phase block, we also tested phase block continuity by setting a more stringent parameter of only allowing ten switched per 20 kb block window. This resulted in an even lower phase switch error rate (0.025% and 0.020% for *E. urophylla* and *E. grandis,* respectively), even though the phase block sizes were smaller (average block N50 of 1.65 Mb and 2.37 Mb, respectively). These results further confirm that the long reads were

separated into the correct haplotype bins and that there are few switch errors within our haplogenome assemblies.

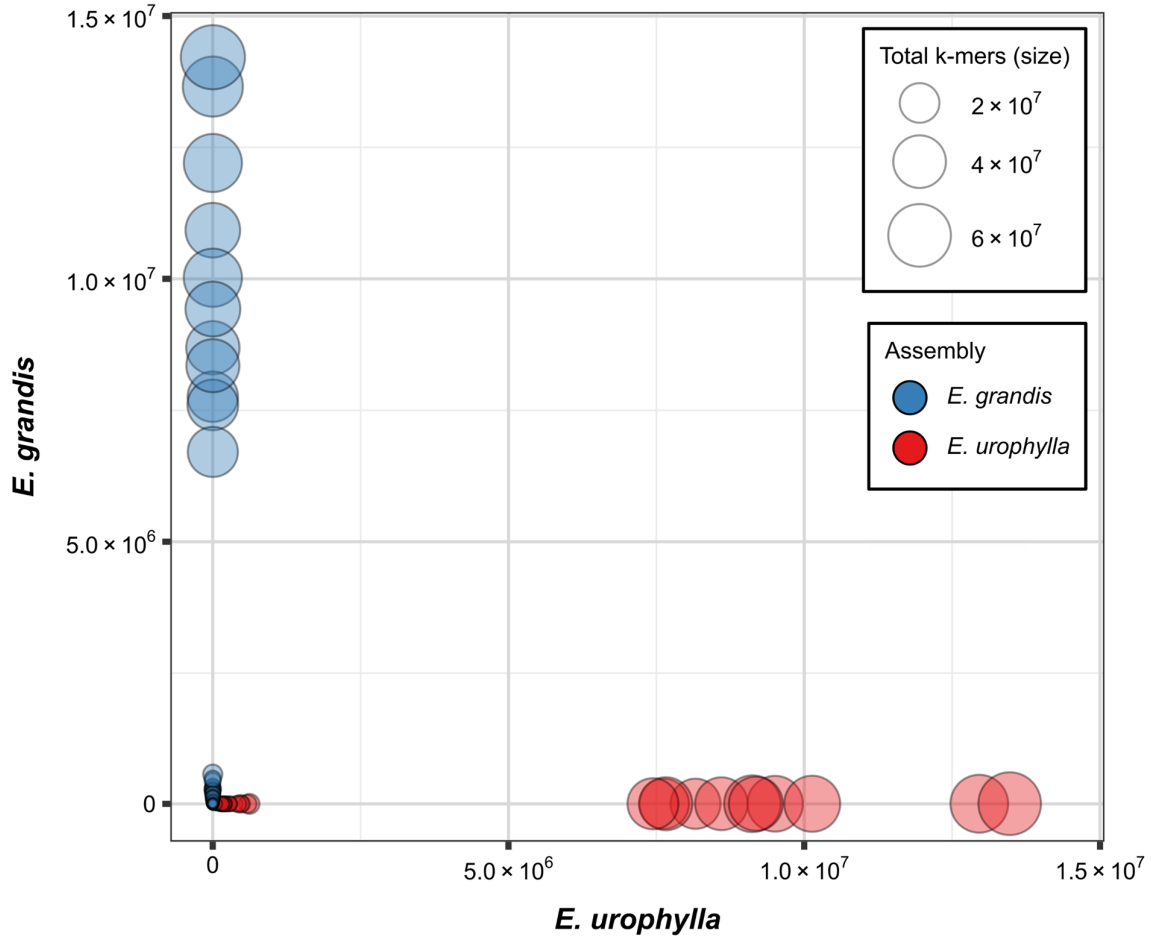
E. urophylla hap-mer markers found in the *E. urophylla* haplogenome assembly and *E. grandis* hap-mer markers found in the *E. grandis* haplogenome assembly and few contaminating markers from the alternative haplogenome (Supplementary Table 2.10). This is reflected in the blob plot (Supplementary Figure 2.12), where the blob represents contigs/scaffolds, the blob size the size of the contig/scaffold, blob colour represents the parental hap-mer to which the blob belongs and how close the blob is to the x- or y-axis represents the assembly in which the hap-mer was found (Rhie *et al.*, 2020). As expected, almost all blobs are close to one of the axes, with the colours matching that of the haplogenome it belongs to (red blobs of *E. urophylla* are close to the *E. urophylla* haplogenome axis, and blue *E. grandis* blobs are close to the *E. grandis* haplogenome axis; Supplementary Figure 2.12). This is expected due to the high level of heterozygosity within *Eucalyptus* (estimated with GenomeScope to be within a range of 1.62% to 3.6%, Supplementary Figure 2.11), which means that most k-mers in the offspring are actually parental hap-mers (Rhie *et al.*, 2020). The high heterozygosity estimates enhance haplotype separation based on trio-binning with Canu, and, together with the results from Merqury, confirms that haplotype separation was highly successful and accurate.

Successful haplotype separation is further evidenced by Supplementary Figure 2.13A, where phase blocks that originate from the wrong haplogenome assembly cannot be seen when 100 and ten hap-mer marker switches are allowed per 20 kb block window. This supports that contigs likely contain markers from only one haplotype. In addition, when plotting the size of the phased blocks and contigs together, phase blocks were larger than contigs and, when plotting the size of phased blocks and scaffolds together, phase blocks were the same size as scaffolds showing good phasing performance (Supplementary Figure 2.13C and D) when 100 switches are allowed per 20 kb block. In comparison, when allowing only 10 switches per block, phase blocks have sizes similar to those of the contigs, indicating phase continuity

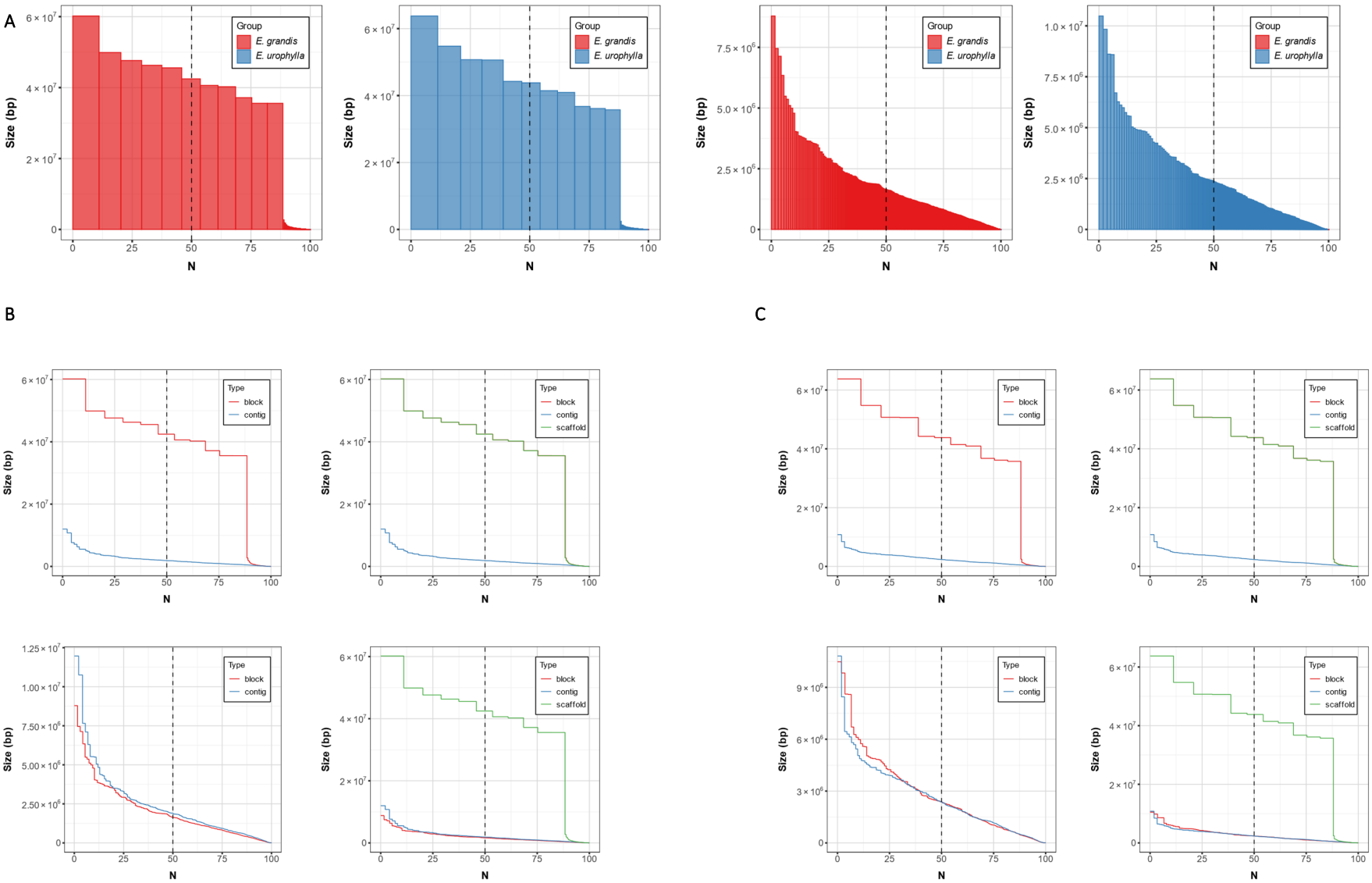
within contigs. Together, these results suggest that Trio-binning with Canu was successful and have resulted in a highly phased haplogenome assembly for *E. grandis* and *E. urophylla*.

Supplementary Table 2.10 Phase block statistics of the *E. grandis* and *E. urophylla* haplo-genome assemblies. Switch error rates are also shown. The number of switch errors per 20 kb are indicated in the phase block column (10 or 100 errors per 20 kb).

Phase blocks	Num. of blocks	Block sum (assembly size, bp)	Smallest block size (bp)	Avg. block size (bp)	Block N50 size (bp)	Longest block size (bp)	Num. of parent specific k-mers from the other haplotype	Total num. of parent specific k-mers in blocks	Switch (%)
E. grandis haplogenome 100/20 kb	1,308	566,274,335	21	432,931	43,818,674	63,773,254	34,846	126,230,267	0.027%
E. urophylla haplogenome 100/20 kb	1,106	543,874,933	21	491,750	42,454,739	60,186,461	39,881	119,331,039	0.033%
E. grandis haplogenome 10/20 kb	2,300	566,082,275	21	246,123	2,374,034	10,482,862	25,382	126,230,267	0.020%
E. urophylla haplogenome 10/20 kb	2,420	543,681,397	21	224,662	1,646,325	8,790,625	30,006	119,331,037	0.025%



Supplementary Figure 2.12 Hap-mer blob plot of the *E. grandis* and *E. urophylla* haplogenome assemblies. All hap-mer information was generated with Merqury v1.1 (Rhie *et al.*, 2020). *E. urophylla* haplogenome contigs are represented by red blobs and *E. grandis* haplogenome contigs are represented by blue blobs. Blob size and contig size are proportional. The number of *E. grandis* (y-axis) and *E. urophylla* (x-axis) hap-mers are plotted per blob/contig. There are almost no *E. grandis* specific k-mers found in the *E. urophylla* assembly, while *E. urophylla* specific k-mers are found in the *E. urophylla* haplogenome assembly.


A
B
C
Group
E. grandis
E. urophylla
Type
block
contig
scaffold
Size (bp)
N

Supplementary Figure 2.13 Evaluation of haplotype phase blocks. All hap-mer information was generated with Merqury v1.1 (Rhie *et al.*, 2020). **(A)** Size sorted phase block N plots of the *E. urophylla* (red) and *E. grandis* (blue) haplogenome assemblies for 100 (left) and 10 (right) switch errors per 20 kb phase block. N shows the percentage of genome size covered by phase blocks of this size and larger are indicated on the x-axis, where the y-axis gives the block size. Blocks from the wrong haplotype are very small and are absent (too small to be seen). (**B** and **C**) Phase block N plots show the continuity of the *E. urophylla* **(B)** and *E. grandis* **(C)** haplogenome assemblies (100 and 10 switch errors allowed per 20 kb on the top and bottom respectively).

Supplementary Note 2.3: Read and assembly alignment and validation of high peak content

To validate that the lower assembly sizes observed in our study were not due to genomic regions that were missing due to low or no long-read sequence coverage, we aligned 150 bp PE short-read sequencing data of the *E. grandis* (FK1758) and *E. urophylla* (FK1756) parents, binned long-read sequencing data and our haplogenome assemblies (contigs) to the *E. grandis* v2.0 reference genome (coverage was calculated as normalized reads per kilobase per million mapped reads and visualised in bins of 100 kb, Supplementary Figure 2.14A). Short-read sequencing data had a mapping rate of 94.34%, 95.07% and 94.78% for the *E. urophylla* and *E. grandis* parent as well as the F_1 hybrid (Supplementary Table 2.1). Long-read sequencing data had a mapping rate of 99.45% (Supplementary Table 2.2), showing that almost all short- and long-reads mapped to the *E. grandis* v2.0 reference genome. There were no bins with zero sequence coverage, suggesting that the entire genome was sequenced, and that the smaller assembly size was not due to low or no long-read sequence coverage. We noted that there were some bins that had very high sequence coverage, and that some of these regions included mitochondrial and chloroplast sequences, which is expected as there is organellar sequence introgression into the nuclear genome of *Eucalyptus* (Pinard *et al.*, 2019). However, we cannot distinguish between reads mapping to introgressed regions and those that are derived from organellar genomes. Another potential cause for bins with high sequence coverage could be repeat elements within that region. To further evaluate the nature of the sequences within high coverage bins, we extracted the bin sequences from the *E. grandis* v2.0 reference genome where genome coverage was above the total bin average within high coverage bins. We used the extracted sequences to identify their origin as either organellar or repetitive, by searching them against a blast database created from the mitochondrial and chloroplast genomes (Pinard *et al.*, 2019), or by performing repeat element identification with RepeatModeler and RepeatMasker as previously described in the methods and materials.

We found that organellar introgression is indeed responsible for some of the high coverage bins (Supplementary Table 2.11). This was expected as Pinard *et al.* (2019) showed that there is significant

introgression of organellar DNA into many regions of the nuclear genome of *E. grandis.* In particular, the high coverage bin on chromosome 9 is due to organellar DNA introgression in *E. grandis* and *E. urophylla* (Supplementary Table 2.11 and Supplementary Figure 2.15), however whether their origin is ancestral or shared still needs to be explored. This is as expected as Pinard *et al.* (2019) also found multiple introgression events in chromosome 9. Although introgression of organellar DNA explains some of the high coverage bins we have observed, the majority of identified high coverage bins contained repetitive elements, many of which are rRNA elements from the rnd-2 repeat class (Supplementary Table 2.11 and Supplementary Figure 2.15). There were also only two repeat family classes on a single chromosome (Supplementary Table 2.11 and Supplementary Figure 2.15). In conclusion, bins with high coverage are mostly the result of the high number of repeat elements found within then, with the exception of chromosome 9, which is due to organellar introgression.

Supplementary Table 2.11 *E. grandis* and *E. urophylla* high coverage bin content. A summary of the blast and RepeatMasker results is given for the genomic sequences in high *E. grandis* v2.0 genome coverage bins. The genomic regions are given (chromosome followed by the sequence position) as Query seqid, and the repeat element or mitochondrial or chloroplast sequence as the Subject seqid. Results are sorted by chromosome followed by the sequence positions.

Query seqid	Query start	Query end	Subject seqid[a]	Type[b]	Subject start	Subject end
			E. grandis			
Chr01:27900113-27901812	818	945	NC_040010.1	mitochondrial	1528	1400
Chr01:27904946-27909398	2	4450	rnd-2_family-40	Unknown	1	4456
Chr01:27905124-27906277	1	1153	rnd-2_family-40	Unknown	179	1341
Chr01:27906500-27907809	1	1309	rnd-2_family-40	Unknown	1561	2869
Chr01:27907885-27909052	1	1167	rnd-2_family-40	Unknown	2947	4112
Chr01:27910186-27912824	306	1078	rnd-2_family-40	Unknown	1851	3272
Chr01:27910186-27912824	1063	2572	rnd-2_family-40	Unknown	2358	3409
Chr01:27910756-27910861	1	105	rnd-2_family-40	Unknown	2100	2202
Chr01:27910945-27911345	1	298	rnd-2_family-40	Unknown	2932	3243
Chr01:27910945-27911345	304	400	rnd-2_family-40	Unknown	2262	2357
Chr01:27911384-27912795	5	1374	rnd-2_family-40	Unknown	2358	3409
Chr01:27912844-27915201	67	161	rnd-2_family-40	Unknown	572	662
Chr01:27912844-27916498	67	161	rnd-2_family-40	Unknown	572	662
Chr01:27929727-27932852	1590	3123	rnd-2_family-40	Unknown	1	1544
Chr01:27930207-27932411	1110	2202	rnd-2_family-40	Unknown	1	1101
Chr01:27932461-27933315	1	583	rnd-2_family-40	Unknown	1154	1745
Chr01:27932461-27933315	582	854	NC_040010.1	mitochondrial	152314	152582
Chr01:27932894-27933012	1	118	rnd-2_family-40	Unknown	1590	1707
Chr01:27933390-27936405	356	1235	rnd-2_family-40	Unknown	1713	3237
Chr01:27933390-27936405	1242	1354	rnd-2_family-40	Unknown	2260	2373
Chr01:27933390-27936405	1354	2854	rnd-2_family-40	Unknown	2358	3546
Chr01:27933390-27936405	2911	3014	rnd-2_family-40	Unknown	4352	4456
Chr01:27933787-27935996	1	838	rnd-2_family-40	Unknown	1755	3237
Chr01:27933787-27935996	845	957	rnd-2_family-40	Unknown	2260	2373
Chr01:27933787-27935996	957	2209	rnd-2_family-40	Unknown	2358	3298
Chr01:27936076-27936202	1	126	rnd-2_family-40	Unknown	3379	3504
Chr01:27936428-27939805	814	958	rnd-2_family-40	Unknown	560	702
Chr01:27936428-27939805	3244	3361	rnd-2_family-40	Unknown	2396	2513
Chr01:27936999-27937661	243	387	rnd-2_family-40	Unknown	560	702
Chr01:27939234-27939742	438	508	rnd-2_family-40	Unknown	2396	2466
Chr01:27939817-27944583	1	1842	rnd-2_family-40	Unknown	2605	4456
Chr01:27939817-27944583	2899	3668	rnd-2_family-40	Unknown	1851	3272
Chr01:27939817-27944583	3653	4365	rnd-2_family-40	Unknown	6	685
Chr01:27939817-27944583	4413	4632	NC_040010.1	mitochondrial	152280	152500
Chr01:27940076-27942426	1	1583	rnd-2_family-40	Unknown	2872	4456
Chr01:27940076-27942426	935	1013	NC_040010.1	mitochondrial	436100	436179
Chr01:27942542-27944575	174	923	rnd-2_family-40	Unknown	1851	3244
Chr01:27942542-27944575	928	1640	rnd-2_family-40	Unknown	6	685
Chr01:27942542-27944575	1688	1907	NC_040010.1	mitochondrial	152280	152500

Chr01:27944607-27956058	711	1391	rnd-2_family-40	Unknown	2654	3356
Chr01:27944607-27956058	1603	1709	rnd-2_family-40	Unknown	560	662
Chr01:27945169-27947611	149	829	rnd-2_family-40	Unknown	2654	3356
Chr01:27945169-27947611	1041	1147	rnd-2_family-40	Unknown	560	662
Chr01:27956077-27960535	672	3240	rnd-2_family-40	Unknown	14	3240
Chr01:27956077-27960535	3239	3351	rnd-2_family-40	Unknown	2260	2373
Chr01:27956077-27960535	3351	3600	rnd-2_family-40	Unknown	86	330
Chr01:27956118-27957607	631	1483	rnd-2_family-40	Unknown	14	867
Chr01:27957619-27960426	1	1698	rnd-2_family-40	Unknown	873	3240
Chr01:27957619-27960426	1697	1809	rnd-2_family-40	Unknown	2260	2373
Chr01:27957619-27960426	1809	2058	rnd-2_family-40	Unknown	86	330
Chr01:27960576-27965778	2645	2965	rnd-2_family-40	Unknown	2358	2607
Chr01:27962629-27963653	592	912	rnd-2_family-40	Unknown	2358	2607
Chr01:27965815-27967195	278	913	rnd-2_family-40	Unknown	2684	3352
Chr01:27965815-27967195	1134	1279	rnd-2_family-40	Unknown	560	702
Chr01:27966095-27966990	1	633	rnd-2_family-40	Unknown	2687	3352
Chr01:27967713-27975626	6977	7913	rnd-2_family-40	Unknown	1	942
Chr01:27973686-27974796	1004	1110	rnd-2_family-40	Unknown	1	108
Chr01:27975718-27976943	1	1225	rnd-2_family-40	Unknown	1036	2259
Chr01:27975965-27976224	1	259	rnd-2_family-40	Unknown	1281	1540
Chr01:27976675-27977201	1	526	rnd-2_family-40	Unknown	1990	2264
Chr01:27977261-27977344	1	83	rnd-2_family-40	Unknown	2275	2357
Chr01:27977755-27977887	4	132	rnd-2_family-40	Unknown	2444	2572
Chr01:27978212-27978291	1	79	rnd-2_family-40	Unknown	2907	2985
Chr01:27978614-27978675	1	61	rnd-2_family-40	Unknown	3310	3370
Chr01:27979737-27980327	189	334	rnd-2_family-40	Unknown	560	702
Chr01:27992448-27993267	158	819	rnd-2_family-40	Unknown	1	663
Chr01:27993781-27993980	1	199	rnd-2_family-40	Unknown	1280	1479
Chr01:27994158-27994259	1	101	rnd-2_family-40	Unknown	1648	1748
Chr01:27998719-27998772	1	53	rnd-2_family-40	Unknown	264	316
Chr02:22324794-22324881	1	81	rnd-2_family-5	rRNA	-235	899
Chr02:22324894-22325061	2	167	rnd-2_family-5	rRNA	-346	788
Chr02:22325073-22325232	3	159	rnd-2_family-5	rRNA	-477	657
Chr02:22331577-22331666	1	89	rnd-2_family-5	rRNA	-413	721
Chr02:22351050-22351099	1	49	rnd-2_family-5	rRNA	-236	898
Chr02:22354417-22354496	1	79	rnd-2_family-5	rRNA	-500	634
Chr02:22361244-22361395	1	151	rnd-2_family-7	rRNA	-1	2497
Chr02:22361450-22361633	1	183	rnd-2_family-7	rRNA	-207	2291
Chr02:22361657-22361886	1	229	rnd-2_family-7	rRNA	-415	2083
Chr02:22363675-22364328	265	529	rnd-2_family-5	rRNA	-680	454
Chr02:22383597-22383625	1	28	rnd-2_family-5	rRNA	-66	1068
Chr02:22383636-22387348	843	956	rnd-2_family-5	rRNA	-253	881
Chr02:22383636-22387348	1402	1666	rnd-2_family-5	rRNA	99	454
Chr02:22383636-22387348	2140	2201	rnd-2_family-5	rRNA	-992	142
Chr02:22383636-22387348	2193	2420	rnd-2_family-5	rRNA	1	227
Chr02:22383636-22387348	2202	3014	rnd-2_family-5	rRNA	-360	774
Chr02:22383636-22387348	2486	3700	rnd-2_family-5	rRNA	2	1134
Chr02:22384542-22385844	1	50	rnd-2_family-5	rRNA	-326	808
Chr02:22384542-22385844	496	760	rnd-2_family-5	rRNA	99	454
Chr02:22384542-22385844	1234	1302	rnd-2_family-5	rRNA	-992	142
Chr02:22385034-22385154	4	120	rnd-2_family-5	rRNA	217	331

Chr02:22385238-22385372	1	39	rnd-2_family-5	rRNA	416	454
Chr02:22385549-22385883	227	331	rnd-2_family-5	rRNA	-992	142
Chr02:22386100-22387076	22	976	rnd-2_family-5	rRNA	2	899
Chr02:22387015-22387072	1	57	rnd-2_family-5	rRNA	839	895
Chr02:22387099-22387240	1	141	rnd-2_family-5	rRNA	924	1064
Chr02:22387208-22387258	1	50	rnd-2_family-5	rRNA	1033	1082
Chr02:22387272-22387338	1	64	rnd-2_family-5	rRNA	1066	1134
Chr02:22387371-22391423	142	2637	rnd-2_family-7	rRNA	0	2498
Chr02:22387510-22388473	3	963	rnd-2_family-7	rRNA	0	2498
Chr02:22387513-22388150	1	637	rnd-2_family-7	rRNA	-1	2497
Chr02:22388484-22388627	1	143	rnd-2_family-7	rRNA	-973	1525
Chr02:22388639-22388943	1	304	rnd-2_family-7	rRNA	-1128	1370
Chr02:22388984-22389313	1	329	rnd-2_family-7	rRNA	-1473	1025
Chr02:22389335-22389521	1	186	rnd-2_family-7	rRNA	-1824	674
Chr02:22389538-22389836	1	298	rnd-2_family-7	rRNA	-2027	471
Chr02:22398421-22401537	9	48	rnd-2_family-5	rRNA	-63	1071
Chr02:22398421-22401537	212	300	rnd-2_family-5	rRNA	144	228
Chr02:22398421-22401537	893	1006	rnd-2_family-5	rRNA	-253	881
Chr02:22398421-22401537	1452	1716	rnd-2_family-5	rRNA	99	454
Chr02:22398421-22401537	2194	2258	rnd-2_family-5	rRNA	-992	142
Chr02:22398421-22401537	2247	3116	rnd-2_family-5	rRNA	1	1009
Chr02:22399335-22400258	1	92	rnd-2_family-5	rRNA	-275	859
Chr02:22399335-22400258	538	802	rnd-2_family-5	rRNA	99	454
Chr02:22400399-22401537	216	280	rnd-2_family-5	rRNA	-992	142
Chr02:22400399-22401537	269	1138	rnd-2_family-5	rRNA	1	1009
Chr02:22404442-22404583	101	141	rnd-2_family-5	rRNA	756	799
Chr02:22412412-22413943	1	1531	rnd-2_family-7	rRNA	97	1627
Chr02:22442538-22451781	1	1335	rnd-2_family-7	rRNA	1038	2367
Chr02:22442538-22451781	1333	1463	rnd-2_family-7	rRNA	2327	2460
Chr02:22442538-22451781	1449	1716	rnd-2_family-7	rRNA	2254	2498
Chr02:22442538-22451781	1822	2884	rnd-2_family-5	rRNA	0	1134
Chr02:22442538-22451781	3602	3715	rnd-2_family-5	rRNA	756	881
Chr02:22442538-22451781	4308	4396	rnd-2_family-5	rRNA	-906	228
Chr02:22442538-22451781	4560	4599	rnd-2_family-5	rRNA	1032	1071
Chr02:22442538-22451781	7252	9243	rnd-2_family-7	rRNA	2	1983
Chr02:22443302-22443506	1	204	rnd-2_family-7	rRNA	1802	2005
Chr02:22443529-22444113	1	344	rnd-2_family-7	rRNA	2029	2367
Chr02:22443529-22444113	342	472	rnd-2_family-7	rRNA	2327	2460
Chr02:22443529-22444113	458	583	rnd-2_family-7	rRNA	2254	2369
Chr02:22444127-22444507	1	127	rnd-2_family-7	rRNA	2372	2498
Chr02:22444127-22444507	233	380	rnd-2_family-5	rRNA	0	1134
Chr02:22444583-22445042	1	442	rnd-2_family-5	rRNA	-237	897
Chr02:22445098-22445297	1	188	rnd-2_family-5	rRNA	-362	772
Chr02:22445329-22445373	2	44	rnd-2_family-5	rRNA	-529	605
Chr02:22446018-22449703	122	235	rnd-2_family-5	rRNA	756	881
Chr02:22446018-22449703	828	916	rnd-2_family-5	rRNA	-906	228
Chr02:22446018-22449703	1080	1119	rnd-2_family-5	rRNA	1032	1071
Chr02:22449788-22450114	2	326	rnd-2_family-7	rRNA	2	326
Chr02:22450125-22450236	1	106	rnd-2_family-5	rRNA	-624	510
Chr02:22450165-22450229	1	64	rnd-2_family-5	rRNA	-622	512
Chr02:22450244-22450360	10	116	rnd-2_family-7	rRNA	455	561

Chr02:22450371-22450534	1	163	rnd-2_family-7	rRNA	573	735
Chr02:22450545-22450593	1	48	rnd-2_family-7	rRNA	747	794
Chr02:22450610-22450962	1	352	rnd-2_family-7	rRNA	812	1163
Chr02:22450974-22451577	1	603	rnd-2_family-7	rRNA	1176	1778
Chr02:22451596-22451781	1	185	rnd-2_family-7	rRNA	1798	1983
Chr02:22454416-22455567	6	217	rnd-2_family-5	rRNA	-903	231
Chr02:22454416-22455567	125	270	rnd-2_family-5	rRNA	12	142
Chr02:22454416-22455567	743	1007	rnd-2_family-5	rRNA	-680	454
Chr02:22461592-22462381	1	376	rnd-2_family-7	rRNA	2096	2460
Chr02:22461592-22462381	362	612	rnd-2_family-7	rRNA	2254	2498
Chr02:22461592-22462381	627	696	rnd-2_family-5	rRNA	-449	685
Chr02:22461592-22462381	692	780	rnd-2_family-5	rRNA	1038	1117
Chr02:22461592-22462381	738	789	rnd-2_family-5	rRNA	0	1134
Chr04:1305-5155	1	3850	rnd-2_family-2	Unknown	1	703
Chr07:28766528-28775042	2103	3671	NC_040010.1	mitochondrial	1577	1
Chr07:28767636-28767757	1	121	NC_040010.1	mitochondrial	1296	1175
Chr07:28767777-28770565	854	2422	NC_040010.1	mitochondrial	1577	1
Chr07:28770856-28775023	2104	3670	NC_040010.1	mitochondrial	1577	1
Chr07:28775105-28779796	2513	4080	NC_040010.1	mitochondrial	1577	1
Chr07:28779940-28780380	1	440	NC_040010.1	mitochondrial	530	91
Chr08:3021-5525	2	2504	rnd-2_family-2	Unknown	6	696
Chr08:436-3002	3	2566	rnd-2_family-2	Unknown	6	703
Chr08:5539-8320	52	2778	rnd-2_family-2	Unknown	6	703
Chr09:28439306-28439677	1	371	NC_040010.1	mitochondrial	269327	269697
Chr09:28444416-28445106	1	690	NC_040010.1	mitochondrial	274439	275128
Chr09:28446035-28446150	1	115	NC_040010.1	mitochondrial	276058	276172
Chr09:28446331-28447025	1	694	NC_040010.1	mitochondrial	276354	277045
Chr09:28447432-28447515	1	83	NC_040010.1	mitochondrial	277453	277535
Chr09:28447697-28447937	1	240	NC_040010.1	mitochondrial	277718	277957
Chr09:28455592-28455658	1	66	NC_040010.1	mitochondrial	285623	285688
Chr09:28455675-28455819	1	144	NC_040010.1	mitochondrial	285706	285849
Chr09:28456407-28456820	1	413	NC_040010.1	mitochondrial	286438	286850
Chr09:28458096-28458505	1	409	NC_040010.1	mitochondrial	288128	288536
Chr09:28470448-28471709	1	1261	MG925369.1	chloroplast	100652	101912
Chr09:28472440-28472983	1	543	MG925369.1	chloroplast	102643	103185
Chr09:28473031-28473130	1	99	MG925369.1	chloroplast	103234	103332
Chr09:28473262-28473993	1	731	MG925369.1	chloroplast	103465	104195
Chr09:28473262-28473993	569	731	NC_040010.1	mitochondrial	476331	476493
Chr09:28474042-28474263	1	221	MG925369.1	chloroplast	104245	104465
Chr09:28474042-28474263	1	221	NC_040010.1	mitochondrial	476543	476763
Chr09:28474377-28475057	1	680	MG925369.1	chloroplast	104580	105259
Chr09:28474377-28475057	1	680	NC_040010.1	mitochondrial	476878	477557
Chr09:28475081-28475313	1	232	MG925369.1	chloroplast	105284	105515
Chr09:28475081-28475313	8	232	NC_040010.1	mitochondrial	477690	477914
Chr09:28477909-28478023	1	114	MG925369.1	chloroplast	107058	107171
Chr09:28477909-28478023	1	114	NC_040010.1	mitochondrial	123515	123628
Chr09:28478200-28479134	321	934	NC_040010.1	mitochondrial	291070	291683
Chr09:28484219-28484629	1	410	NC_040010.1	mitochondrial	296770	297179
Chr09:28485755-28485968	1	213	NC_040010.1	mitochondrial	298369	298581
Chr09:28486279-28486454	1	175	NC_040010.1	mitochondrial	298893	299067
Chr09:28488520-28488772	1	252	NC_040010.1	mitochondrial	301135	301386

Chr09:28489954-28490880	1	926	NC_040010.1	mitochondrial	302571	303496
Chr09:28491392-28491687	1	295	NC_040010.1	mitochondrial	304057	304351
Chr09:28494819-28495465	1	646	NC_040010.1	mitochondrial	307394	308039
Chr09:28495963-28496584	1	621	NC_040010.1	mitochondrial	308538	309158
Chr09:28497280-28497483	1	203	NC_040010.1	mitochondrial	309855	310057
Chr09:28497548-28497837	67	289	NC_040010.1	mitochondrial	310248	310470
Chr09:28498452-28498808	1	356	NC_040010.1	mitochondrial	311086	311441
Chr09:28498846-28499532	1	686	NC_040010.1	mitochondrial	311480	312165
Chr10:32730112-32730311	1	199	MG925369.1	chloroplast	100704	100506
Chr10:32730391-32730696	1	305	MG925369.1	chloroplast	100424	100120
Chr10:32730776-32730963	1	187	MG925369.1	chloroplast	99702	99888
Chr10:32732160-32732296	4	136	MG925369.1	chloroplast	32644	32787
Chr10:32732160-32732296	4	136	NC_040010.1	mitochondrial	159194	159051
Chr10:32745461-32745819	1	358	MG925369.1	chloroplast	89122	89479
Chr10:32745461-32745819	1	358	NC_040010.1	mitochondrial	23775	23418
Chr10:32749232-32749415	1	183	MG925369.1	chloroplast	57977	58159
Chr10:32753395-32753617	1	222	MG925369.1	chloroplast	27696	27917
Chr10:32765140-32765194	1	54	MG925369.1	chloroplast	23078	23025
Chr10:32769304-32769430	1	36	MG925369.1	chloroplast	90646	90681
Chr10:32769304-32769430	1	36	NC_040010.1	mitochondrial	22251	22216
Chr10:51819-51891	1	72	rnd-2_family-2	Unknown	73	144
Chr10:52214-52263	1	49	rnd-2_family-2	Unknown	102	150
Chr10:52488-52632	1	144	rnd-2_family-2	Unknown	216	360
Chr10:54026-54861	3	835	rnd-2_family-2	Unknown	1	703
Chr10:55576-56233	2	651	rnd-2_family-2	Unknown	52	702
Chr10:61671-78127	1	4864	rnd-2_family-2	Unknown	1	703
Chr10:61671-78127	4969	16452	rnd-2_family-2	Unknown	1	703
Chr10:61671-78129	1	4864	rnd-2_family-2	Unknown	1	703
Chr10:61671-78129	4969	16452	rnd-2_family-2	Unknown	1	703
Chr10:61761-61924	1	163	rnd-2_family-2	Unknown	539	701
Chr10:61961-62286	1	325	rnd-2_family-2	Unknown	189	514
Chr10:62325-62681	2	356	rnd-2_family-2	Unknown	189	544
Chr10:62739-63028	1	287	rnd-2_family-2	Unknown	236	524
Chr10:63056-63212	1	156	rnd-2_family-2	Unknown	387	542
Chr10:63281-63590	1	309	rnd-2_family-2	Unknown	260	569
Chr10:63638-63731	1	93	rnd-2_family-2	Unknown	68	160
Chr10:63795-64151	1	356	rnd-2_family-2	Unknown	42	395
Chr10:64193-64272	1	79	rnd-2_family-2	Unknown	71	149
Chr10:64300-64402	1	102	rnd-2_family-2	Unknown	362	463
Chr10:64433-64922	1	489	rnd-2_family-2	Unknown	129	618
Chr10:64946-65011	1	65	rnd-2_family-2	Unknown	460	524
Chr10:65050-65379	1	323	rnd-2_family-2	Unknown	381	703
Chr10:65401-65564	1	163	rnd-2_family-2	Unknown	365	527
Chr10:65663-66438	1	775	rnd-2_family-2	Unknown	1	703
Chr10:66672-66755	1	83	rnd-2_family-2	Unknown	75	157
Chr10:66786-67110	1	324	rnd-2_family-2	Unknown	204	528
Chr10:67174-67323	1	149	rnd-2_family-2	Unknown	43	191
Chr10:67387-67645	1	258	rnd-2_family-2	Unknown	440	697
Chr10:67679-67795	1	116	rnd-2_family-2	Unknown	1	116
Chr10:67813-68074	1	260	rnd-2_family-2	Unknown	443	703
Chr10:68184-68273	1	89	rnd-2_family-2	Unknown	95	182

Chr10:68419-68890	1	471	rnd-2_family-2	Unknown	146	618
Chr10:69132-69213	1	80	rnd-2_family-2	Unknown	253	332
Chr10:69232-69296	1	64	rnd-2_family-2	Unknown	168	231
Chr10:69322-69459	1	137	rnd-2_family-2	Unknown	75	211
Chr10:69504-69720	2	216	rnd-2_family-2	Unknown	75	289
Chr10:69895-70035	1	140	rnd-2_family-2	Unknown	99	238
Chr10:70366-70526	1	160	rnd-2_family-2	Unknown	75	234
Chr10:70689-71002	1	313	rnd-2_family-2	Unknown	215	528
Chr10:71096-71396	1	300	rnd-2_family-2	Unknown	75	375
Chr10:71429-71626	1	197	rnd-2_family-2	Unknown	409	605
Chr10:71835-71993	1	158	rnd-2_family-2	Unknown	455	612
Chr10:72056-72107	1	51	rnd-2_family-2	Unknown	126	176
Chr10:72214-72365	1	151	rnd-2_family-2	Unknown	102	252
Chr10:72387-72522	1	135	rnd-2_family-2	Unknown	92	227
Chr10:72621-72806	1	185	rnd-2_family-2	Unknown	328	512
Chr10:73043-73341	1	297	rnd-2_family-2	Unknown	384	680
Chr10:73419-73542	1	122	rnd-2_family-2	Unknown	577	698
Chr10:73575-74282	2	703	rnd-2_family-2	Unknown	1	703
Chr10:74443-75237	1	794	rnd-2_family-2	Unknown	1	703
Chr10:75306-76368	1	1062	rnd-2_family-2	Unknown	1	703
Chr10:76394-76479	1	85	rnd-2_family-2	Unknown	75	158
Chr10:76579-77019	1	440	rnd-2_family-2	Unknown	75	515
Chr10:77079-77366	1	287	rnd-2_family-2	Unknown	393	679
Chr10:77408-77805	1	397	rnd-2_family-2	Unknown	172	568
Chr10:77892-77945	1	53	rnd-2_family-2	Unknown	106	158
Chr10:77996-78035	1	39	rnd-2_family-2	Unknown	210	248
Chr11:24024757-24026068	1146	1251	rnd-2_family-45	Unknown	379	484
Chr11:24024757-24026068	1269	1310	rnd-2_family-45	Unknown	474	515
Chr11:24028609-24032432	2	41	rnd-2_family-45	Unknown	445	484
Chr11:24028609-24032432	48	577	rnd-2_family-45	Unknown	761	1379
Chr11:24042370-24042566	146	196	rnd-2_family-45	Unknown	1	51
Chr11:24042680-24043121	1	441	rnd-2_family-45	Unknown	166	609
Chr11:24043199-24047383	1	694	rnd-2_family-45	Unknown	688	1382
Chr11:24047445-24049188	1619	1743	rnd-2_family-45	Unknown	1	125
Chr11:24049219-24049802	1	583	rnd-2_family-45	Unknown	199	781
Chr11:24049856-24051712	1	538	rnd-2_family-45	Unknown	840	1382
Chr11:24070224-24071609	2	1382	rnd-2_family-45	Unknown	1	1382
Chr11:24080542-24083870	1	522	rnd-2_family-45	Unknown	412	967
Chr11:24080542-24083870	540	870	rnd-2_family-45	Unknown	1054	1376
Chr11:24084573-24085756	358	444	rnd-2_family-45	Unknown	395	484
Chr11:24084573-24085756	451	980	rnd-2_family-45	Unknown	761	1379
Chr11:24096117-24098112	863	1995	rnd-2_family-45	Unknown	1	1128
E. urophylla						
Chr02:22324906-22325221	3	314	rnd-2_family-5	rRNA	-87	4605
Chr02:22350942-22351100	1	131	rnd-2_family-5	rRNA	4236	4364
Chr02:22354435-22354497	1	61	rnd-2_family-5	rRNA	4236	4296
Chr02:22355414-22355852	4	73	rnd-2_family-5	rRNA	1205	1280
Chr02:22361242-22361405	1	163	rnd-2_family-5	rRNA	-780	3912
Chr02:22361452-22361929	1	477	rnd-2_family-5	rRNA	-996	3696
Chr02:22363675-22364401	305	358	rnd-2_family-5	rRNA	1220	1279
Chr02:22363675-22364401	362	431	rnd-2_family-5	rRNA	1205	1280

Chr02:22383595-22387348	1	31	rnd-2_family-5	rRNA	4178	4208
Chr02:22383595-22387348	1541	1610	rnd-2_family-5	rRNA	-3412	1280
Chr02:22383595-22387348	1614	1667	rnd-2_family-5	rRNA	-3413	1279
Chr02:22383595-22387348	2224	2352	rnd-2_family-5	rRNA	4571	4691
Chr02:22383595-22387348	2269	2723	rnd-2_family-5	rRNA	0	4692
Chr02:22383595-22387348	2684	3413	rnd-2_family-5	rRNA	4236	4686
Chr02:22383595-22387348	3103	3753	rnd-2_family-5	rRNA	0	4692
Chr02:22383826-22385890	1310	1379	rnd-2_family-5	rRNA	-3412	1280
Chr02:22383826-22385890	1383	1436	rnd-2_family-5	rRNA	-3413	1279
Chr02:22383826-22385890	1995	2064	rnd-2_family-5	rRNA	-250	4442
Chr02:22384628-22385451	508	577	rnd-2_family-5	rRNA	-3412	1280
Chr02:22384628-22385451	581	634	rnd-2_family-5	rRNA	-3413	1279
Chr02:22385911-22391416	1	407	rnd-2_family-5	rRNA	-28	4664
Chr02:22385911-22391416	52	483	rnd-2_family-5	rRNA	4267	4685
Chr02:22385911-22391416	408	740	rnd-2_family-5	rRNA	-109	4583
Chr02:22385911-22391416	549	1097	rnd-2_family-5	rRNA	4236	4686
Chr02:22385911-22391416	787	5505	rnd-2_family-5	rRNA	0	4692
Chr02:22385920-22386235	1	315	rnd-2_family-5	rRNA	-37	4655
Chr02:22386247-22386308	2	61	rnd-2_family-5	rRNA	-388	4304
Chr02:22386344-22387232	9	664	rnd-2_family-5	rRNA	4236	4686
Chr02:22386344-22387232	354	888	rnd-2_family-5	rRNA	0	4692
Chr02:22387266-22387953	1	687	rnd-2_family-5	rRNA	-535	4157
Chr02:22387371-22391423	1	4045	rnd-2_family-5	rRNA	-640	4052
Chr02:22387970-22388203	1	233	rnd-2_family-5	rRNA	-1245	3447
Chr02:22398432-22401443	1	37	rnd-2_family-5	rRNA	4180	4216
Chr02:22398432-22401443	1539	1608	rnd-2_family-5	rRNA	-3412	1280
Chr02:22398432-22401443	1612	1665	rnd-2_family-5	rRNA	-3413	1279
Chr02:22398432-22401443	2228	2405	rnd-2_family-5	rRNA	-84	4608
Chr02:22398432-22401443	2234	2450	rnd-2_family-5	rRNA	4465	4692
Chr02:22398432-22401443	2406	2835	rnd-2_family-5	rRNA	0	4692
Chr02:22398432-22401443	2453	2886	rnd-2_family-5	rRNA	4233	4685
Chr02:22398432-22401443	2837	3011	rnd-2_family-5	rRNA	-185	4507
Chr02:22399334-22400324	637	706	rnd-2_family-5	rRNA	-3412	1280
Chr02:22399334-22400324	710	763	rnd-2_family-5	rRNA	-3413	1279
Chr02:22399342-22400314	629	698	rnd-2_family-5	rRNA	-3412	1280
Chr02:22399342-22400314	702	755	rnd-2_family-5	rRNA	-3413	1279
Chr02:22400403-22401476	255	636	rnd-2_family-5	rRNA	4316	4692
Chr02:22400403-22401476	435	864	rnd-2_family-5	rRNA	0	4692
Chr02:22400403-22401476	661	1073	rnd-2_family-5	rRNA	4233	4663
Chr02:22400573-22400895	87	322	rnd-2_family-5	rRNA	-84	4608
Chr02:22400940-22401001	2	61	rnd-2_family-5	rRNA	-388	4304
Chr02:22409506-22411840	1404	2334	rnd-2_family-5	rRNA	1	931
Chr02:22409514-22413388	1396	3874	rnd-2_family-5	rRNA	1	2479
Chr02:22410204-22411039	706	835	rnd-2_family-5	rRNA	1	130
Chr02:22411851-22412147	1	296	rnd-2_family-5	rRNA	943	1236
Chr02:22412162-22412788	3	626	rnd-2_family-5	rRNA	1256	1879
Chr02:22412804-22412991	1	187	rnd-2_family-5	rRNA	1896	2082
Chr02:22413004-22413066	1	62	rnd-2_family-5	rRNA	2096	2157
Chr02:22413408-22413477	1	69	rnd-2_family-5	rRNA	2500	2568
Chr02:22413489-22413592	1	103	rnd-2_family-5	rRNA	2581	2683
Chr02:22442537-22443877	1	1336	rnd-2_family-5	rRNA	2444	3780

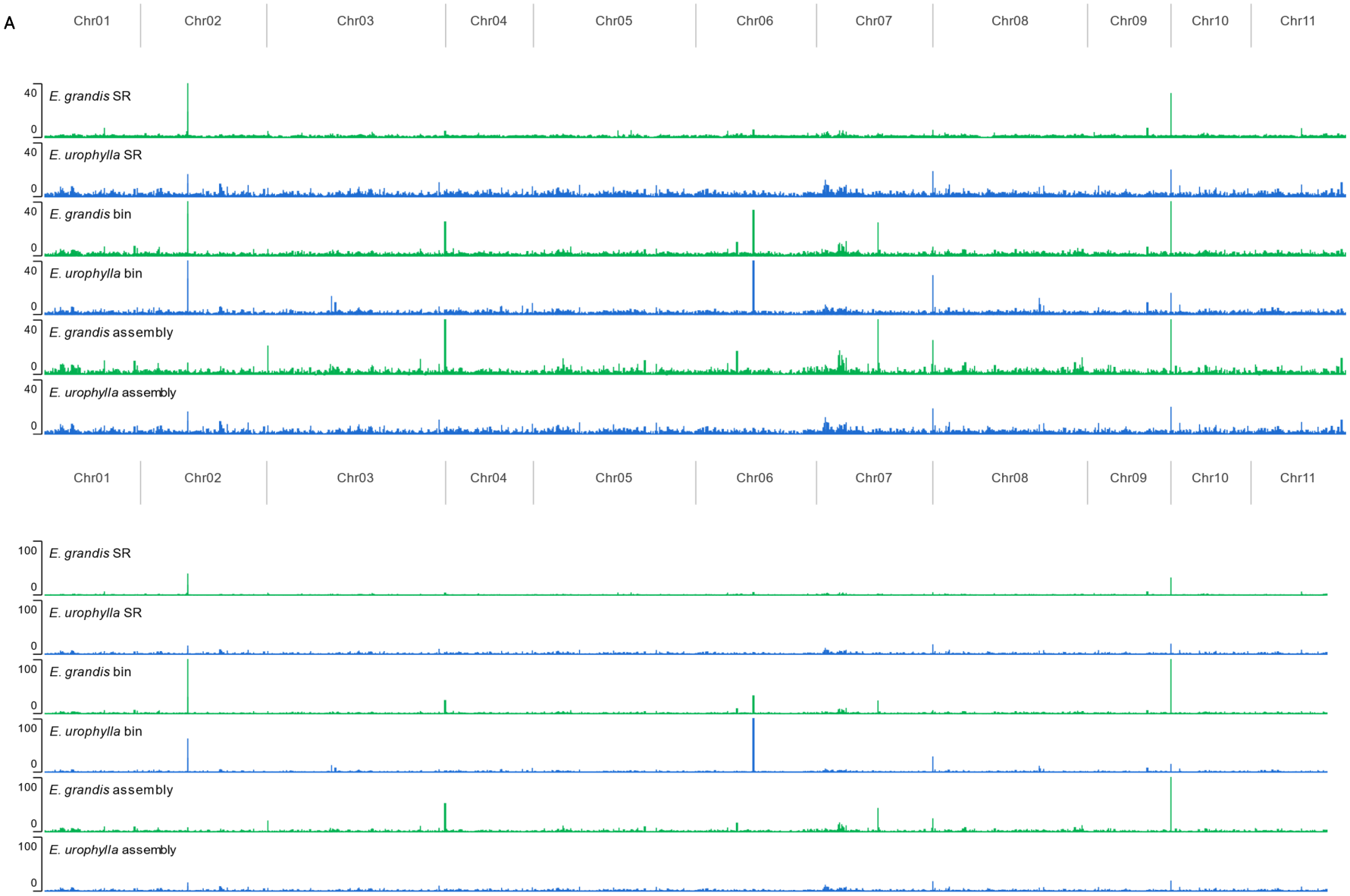
Chr02:22443304-22443516	1	212	rnd-2_family-5	rRNA	3211	3422
Chr02:22443319-22443861	1	542	rnd-2_family-5	rRNA	3226	3768
Chr02:22443533-22443873	1	340	rnd-2_family-5	rRNA	3440	3780
Chr02:22443987-22444113	1	125	rnd-2_family-5	rRNA	3662	3782
Chr02:22444038-22451781	1	921	rnd-2_family-5	rRNA	3713	4692
Chr02:22444038-22451781	738	1220	rnd-2_family-5	rRNA	-6	4686
Chr02:22444038-22451781	1062	1392	rnd-2_family-5	rRNA	4236	4608
Chr02:22444038-22451781	3060	3099	rnd-2_family-5	rRNA	-476	4216
Chr02:22444038-22451781	4489	5751	rnd-2_family-5	rRNA	1	1265
Chr02:22444038-22451781	5750	7743	rnd-2_family-5	rRNA	1407	3390
Chr02:22444043-22444113	1	69	rnd-2_family-5	rRNA	3718	3782
Chr02:22444127-22444170	1	42	rnd-2_family-5	rRNA	3785	3826
Chr02:22444127-22449758	1	832	rnd-2_family-5	rRNA	3785	4692
Chr02:22444127-22449758	649	1131	rnd-2_family-5	rRNA	-6	4686
Chr02:22444127-22449758	973	1303	rnd-2_family-5	rRNA	4236	4608
Chr02:22444127-22449758	2971	3010	rnd-2_family-5	rRNA	-476	4216
Chr02:22444127-22449758	4400	5631	rnd-2_family-5	rRNA	1	1231
Chr02:22444363-22445065	1	596	rnd-2_family-5	rRNA	4092	4692
Chr02:22444363-22445065	287	702	rnd-2_family-5	rRNA	-6	4686
Chr02:22445099-22445378	1	279	rnd-2_family-5	rRNA	4236	4538
Chr02:22448026-22448678	501	652	rnd-2_family-5	rRNA	1	152
Chr02:22449122-22449268	1	146	rnd-2_family-5	rRNA	597	742
Chr02:22449812-22451781	1	1969	rnd-2_family-5	rRNA	1432	3390
Chr02:22450121-22450235	1	114	rnd-2_family-5	rRNA	4412	4538
Chr02:22454414-22455498	2	227	rnd-2_family-5	rRNA	4370	4692
Chr02:22454414-22455498	785	838	rnd-2_family-5	rRNA	1220	1279
Chr02:22454414-22455498	842	911	rnd-2_family-5	rRNA	1205	1280
Chr02:22454416-22461960	9	225	rnd-2_family-5	rRNA	4370	4608
Chr02:22454416-22461960	783	836	rnd-2_family-5	rRNA	1220	1279
Chr02:22454416-22461960	840	909	rnd-2_family-5	rRNA	1205	1280
Chr02:22454416-22461960	2373	2411	rnd-2_family-5	rRNA	-476	4216
Chr02:22454416-22461960	3766	7544	rnd-2_family-5	rRNA	1	3865
Chr02:22461623-22461823	1	200	rnd-2_family-5	rRNA	3534	3735
Chr02:22462028-22462360	1	297	rnd-2_family-5	rRNA	3720	4045
Chr08:17700-29944	9	10033	rnd-2_family-8	Unknown	45	715
Chr08:17700-29944	10097	12152	rnd-2_family-8	Unknown	59	715
Chr08:17700-29991	9	10033	rnd-2_family-8	Unknown	45	715
Chr08:17700-29991	10097	12291	rnd-2_family-8	Unknown	57	715
Chr09:28437603-28437722	1	119	NC_040010.1	Mitochondrial	267646	267764
Chr09:28439381-28439435	1	54	NC_040010.1	Mitochondrial	269402	269455
Chr09:28439504-28439703	1	199	NC_040010.1	Mitochondrial	269525	269723
Chr09:28439892-28440021	1	129	NC_040010.1	Mitochondrial	269913	270041
Chr09:28444447-28445247	1	800	NC_040010.1	Mitochondrial	274470	275269
Chr09:28445380-28445475	1	95	NC_040010.1	Mitochondrial	275403	275497
Chr09:28446057-28446129	1	72	NC_040010.1	Mitochondrial	276080	276151
Chr09:28446336-28446999	1	663	NC_040010.1	Mitochondrial	276359	277019
Chr09:28447412-28448014	1	602	NC_040010.1	Mitochondrial	277433	278034
Chr09:28448187-28448338	1	151	NC_040010.1	Mitochondrial	278208	278364
Chr09:28451258-28451561	1	303	NC_040010.1	Mitochondrial	281286	281588
Chr09:28456459-28456580	1	121	NC_040010.1	Mitochondrial	286490	286610
Chr09:28456671-28456785	1	114	NC_040010.1	Mitochondrial	286702	286815

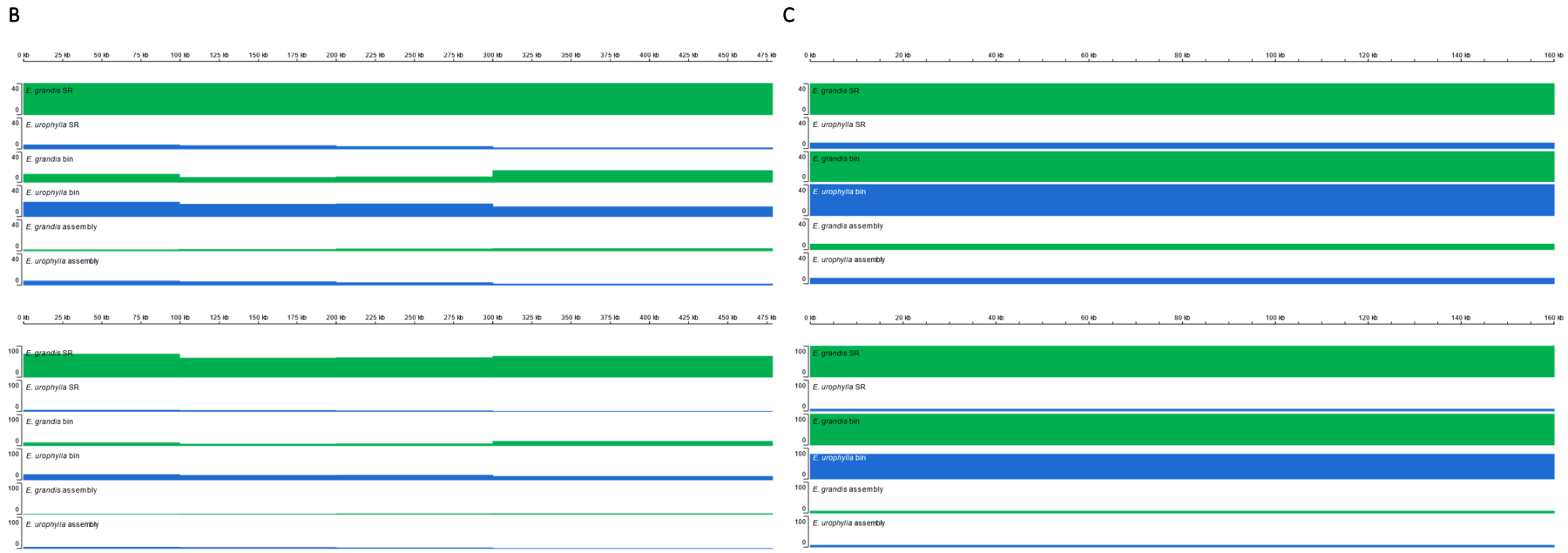
Chr09:28457605-28457827	1	222	NC_040010.1	Mitochondrial	287636	287857
Chr09:28457906-28458525	1	619	NC_040010.1	Mitochondrial	287937	288556
Chr09:28458599-28458877	1	278	NC_040010.1	Mitochondrial	288627	288909
Chr09:28458927-28459128	1	201	NC_040010.1	Mitochondrial	288960	289160
Chr09:28459527-28459559	1	32	NC_040010.1	Mitochondrial	19954	19923
Chr09:28459575-28459874	1	299	NC_040010.1	Mitochondrial	289603	289901
Chr09:28468142-28471705	2281	3563	MG925369.1	chloroplast	100626	101908
Chr09:28468142-28471705	1	38	NC_040010.1	Mitochondrial	7407	7370
Chr09:28468142-28475684	2281	7542	MG925369.1	chloroplast	100626	105890
Chr09:28468142-28475684	5689	7048	NC_040010.1	Mitochondrial	476331	477690
Chr09:28472443-28473000	1	557	MG925369.1	chloroplast	102646	103202
Chr09:28473263-28473981	1	718	MG925369.1	chloroplast	103466	104183
Chr09:28473263-28473981	568	718	NC_040010.1	Mitochondrial	476331	476481
Chr09:28474046-28474262	1	216	MG925369.1	chloroplast	104249	104464
Chr09:28474046-28474262	1	216	NC_040010.1	Mitochondrial	476547	476762
Chr09:28474377-28475065	1	688	MG925369.1	chloroplast	104580	105267
Chr09:28474377-28475065	1	688	NC_040010.1	Mitochondrial	476878	477565
Chr09:28475083-28475314	1	231	MG925369.1	chloroplast	105286	105516
Chr09:28475083-28475314	6	231	NC_040010.1	Mitochondrial	477690	477915
Chr09:28477624-28479611	1	495	MG925369.1	chloroplast	106773	107267
Chr09:28477624-28479611	897	1987	NC_040010.1	Mitochondrial	291070	292160
Chr09:28477862-28479115	1	257	MG925369.1	chloroplast	107011	107267
Chr09:28477862-28479115	659	1253	NC_040010.1	Mitochondrial	291070	291664
Chr09:28479734-28479938	1	204	NC_040010.1	Mitochondrial	292284	292487
Chr09:28479971-28480041	1	70	NC_040010.1	Mitochondrial	292521	292590
Chr09:28480066-28480228	1	162	NC_040010.1	Mitochondrial	292616	292777
Chr09:28484235-28484573	1	338	NC_040010.1	Mitochondrial	296786	297123
Chr09:28488566-28488768	1	202	NC_040010.1	Mitochondrial	301181	301382
Chr09:28488888-28488998	1	110	NC_040010.1	Mitochondrial	301503	301612
Chr09:28489249-28489422	1	173	NC_040010.1	Mitochondrial	301864	302036
Chr09:28489454-28489672	1	218	NC_040010.1	Mitochondrial	302069	302286
Chr09:28489961-28490878	1	917	NC_040010.1	Mitochondrial	302578	303494
Chr09:28491456-28491654	1	198	NC_040010.1	Mitochondrial	304121	304318
Chr09:28492672-28492783	1	111	NC_040010.1	Mitochondrial	305244	305354
Chr09:28494838-28495447	1	609	NC_040010.1	Mitochondrial	307413	308021
Chr09:28495983-28496570	1	587	NC_040010.1	Mitochondrial	308558	309144
Chr09:28498517-28498785	1	268	NC_040010.1	Mitochondrial	311151	311418
Chr09:28498872-28499502	1	630	NC_040010.1	Mitochondrial	311506	312135
Chr10:54026-54853	3	827	rnd-2_family-8	Unknown	1	715
Chr10:55590-56226	1	636	rnd-2_family-8	Unknown	1	715
Chr10:61671-78118	1	4864	rnd-2_family-8	Unknown	1	715
Chr10:61671-78118	4969	7441	rnd-2_family-8	Unknown	1	715
Chr10:61671-78118	7451	8400	rnd-2_family-8	Unknown	1	715
Chr10:61671-78118	8544	16447	rnd-2_family-8	Unknown	1	715
Chr10:61677-78127	1	4858	rnd-2_family-8	Unknown	1	715
Chr10:61677-78127	4963	7435	rnd-2_family-8	Unknown	1	715
Chr10:61677-78127	7445	8394	rnd-2_family-8	Unknown	1	715
Chr10:61677-78127	8538	16450	rnd-2_family-8	Unknown	1	715
Chr10:61806-62087	2	281	rnd-2_family-8	Unknown	93	372
Chr10:62327-62714	1	387	rnd-2_family-8	Unknown	65	451
Chr10:62787-63009	1	220	rnd-2_family-8	Unknown	159	379

Chr10:63088-63199	1	111	rnd-2_family-8	Unknown	110	220
Chr10:63300-63579	1	279	rnd-2_family-8	Unknown	154	432
Chr10:63662-63721	1	59	rnd-2_family-8	Unknown	150	208
Chr10:63847-64145	1	298	rnd-2_family-8	Unknown	149	446
Chr10:64221-64253	1	32	rnd-2_family-8	Unknown	157	188
Chr10:64302-64401	1	99	rnd-2_family-8	Unknown	238	336
Chr10:64491-64923	1	431	rnd-2_family-8	Unknown	62	492
Chr10:65069-65252	1	182	rnd-2_family-8	Unknown	91	272
Chr10:65311-65350	1	39	rnd-2_family-8	Unknown	150	188
Chr10:65413-65573	1	160	rnd-2_family-8	Unknown	251	410
Chr10:65666-65848	1	182	rnd-2_family-8	Unknown	153	334
Chr10:66040-66344	1	304	rnd-2_family-8	Unknown	162	465
Chr10:66693-66733	1	40	rnd-2_family-8	Unknown	154	193
Chr10:66796-67079	1	283	rnd-2_family-8	Unknown	89	371
Chr10:67214-67267	1	53	rnd-2_family-8	Unknown	141	193
Chr10:67370-67645	2	275	rnd-2_family-8	Unknown	115	388
Chr10:67687-67788	1	101	rnd-2_family-8	Unknown	67	167
Chr10:67864-68055	1	191	rnd-2_family-8	Unknown	185	375
Chr10:68418-68855	1	437	rnd-2_family-8	Unknown	20	457
Chr10:68972-69041	1	69	rnd-2_family-8	Unknown	26	94
Chr10:69522-69720	1	198	rnd-2_family-8	Unknown	150	347
Chr10:69893-70035	1	142	rnd-2_family-8	Unknown	155	296
Chr10:70710-70917	1	207	rnd-2_family-8	Unknown	477	683
Chr10:71187-71258	1	71	rnd-2_family-8	Unknown	590	660
Chr10:71331-71404	1	73	rnd-2_family-8	Unknown	2	74
Chr10:71425-71631	1	206	rnd-2_family-8	Unknown	462	667
Chr10:71848-71961	1	113	rnd-2_family-8	Unknown	159	271
Chr10:72056-72114	1	58	rnd-2_family-8	Unknown	1	58
Chr10:72553-72589	1	36	rnd-2_family-8	Unknown	500	535
Chr10:72621-72787	1	166	rnd-2_family-8	Unknown	202	367
Chr10:73091-73319	1	228	rnd-2_family-8	Unknown	123	350
Chr10:73563-74054	1	491	rnd-2_family-8	Unknown	46	535
Chr10:74094-74274	1	179	rnd-2_family-8	Unknown	28	206
Chr10:74463-74500	1	37	rnd-2_family-8	Unknown	33	69
Chr10:74527-74758	1	231	rnd-2_family-8	Unknown	463	693
Chr10:74780-75700	18	915	rnd-2_family-8	Unknown	1	715
Chr10:75712-75794	2	82	rnd-2_family-8	Unknown	1	81
Chr10:76052-76369	1	316	rnd-2_family-8	Unknown	340	655
Chr10:76628-77345	3	717	rnd-2_family-8	Unknown	1	715
Chr10:77430-77804	1	374	rnd-2_family-8	Unknown	69	441
Chr10:77822-77855	1	33	rnd-2_family-8	Unknown	277	309
Chr11:24028678-24032420	1	523	rnd-2_family-9	Unknown	581	1192
Chr11:24042713-24042789	5	76	rnd-2_family-9	Unknown	1	72
Chr11:24042902-24043101	1	199	rnd-2_family-9	Unknown	189	387
Chr11:24043204-24046225	1	698	rnd-2_family-9	Unknown	491	1190
Chr11:24049224-24049797	1	573	rnd-2_family-9	Unknown	2	574
Chr11:24049861-24051475	1	545	rnd-2_family-9	Unknown	644	1193
Chr11:24070256-24071621	172	1362	rnd-2_family-9	Unknown	1	1193
Chr11:24080480-24083129	30	584	rnd-2_family-9	Unknown	177	766
Chr11:24080480-24083129	602	950	rnd-2_family-9	Unknown	853	1192
Chr11:24084573-24085764	358	444	rnd-2_family-9	Unknown	193	282

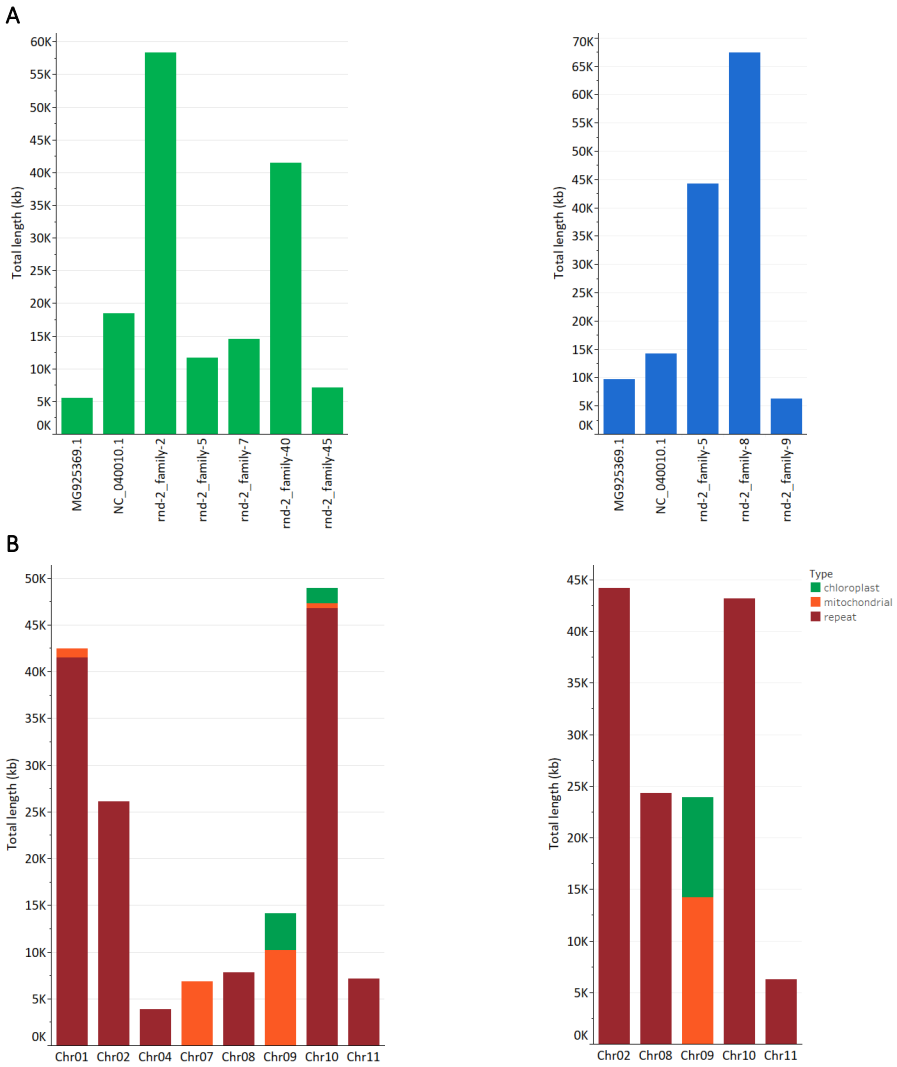
Chr11:24084573-24085764	451	995	rnd-2_family-9	Unknown	559	1192
Chr11:24096497-24098093	685	1596	rnd-2_family-9	Unknown	1	908
Chr11:24098339-24098495	1	34	rnd-2_family-9	Unknown	1160	1193

[a] NC_040010.1 is mitochondrial genome sequences and MG925369.1 are chloroplast genome sequences.
[b] Mitochondrial, chloroplast, repeat family/class


A
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
40
0
E. grandis SR
E. urophylla SR
E. grandis bin
E. urophylla bin
E. grandis assembly
E. urophylla assembly
100



Supplementary Figure 2.14 Genome coverage of the *E. grandis* v2.0 nuclear reference and plastid genomes. (A) Alignment of *E. grandis* (FK1758, green) and *E. urophylla* (FK1756, blue) parental short-read (SR), binned long-read sequencing data and haplogenome assemblies (contigs) to the *E. grandis* v2.0 reference genome (Myburg *et al.*, 2014; Bartholome *et al.*, 2015). Coverage is shown on the y-axis, with max coverage parameters set to 40X (top panel) and 100X (bottom panel), along the eleven *Eucalyptus* chromosomes in bins of 100 kb shown on the x-axis. Alignment of the same sequencing data and assemblies to the *E. grandis* **(B)** mitochondrial (478.8 kb) and **(C)** chloroplast (160.1 kb) genomes (Pinard *et al.*, 2019), at 40X (top panel) and 100X (bottom panel) maximum coverage. All alignments were viewed in the IGV browser and bins were 100 kb in size



Supplementary Figure 2.15 Summary of the total size and type of elements found in high genome coverage bins. Organellar introgression was identified through BLAST analysis to the *E. grandis* plastid genomes (Pinard *et al.*, 2019), while repeat elements were identified with RepeatMasker. **(A)** The total size of different type of elements found in high coverage bins (see Supplementary Figure 2.14) for the *E. grandis* (green) and *E. urophylla* (blue) alignments. The element type is indicated on the x-axis as either mitochondrial (NC040010.1), chloroplast (MG925369.1) or repeat elements (rnd, the repeat family/class is given) and the total length the element contributes to all high coverage bins is given in kb (kilobases) on the y-axis. **(B)** The total length of different types of elements contributed per chromosome within high coverage bins. The chromosomes are indicated on the x-axis for *E. grandis* (left) and *E. urophylla* (right) and the total length contributed by each element is given on the y-axis. Contributions are either repetitive elements in red, mitochondrial introgression in orange or chloroplast introgression in green. Note that in both cases chromosome 9 only has organellar introgression, whereas the majority of other chromosomes have mostly repeat elements.