



De-novo assembly of a reference genome for a critically threatened Aotearoa New Zealand tree species, *Syzygium maire* (Myrtaceae)

Colan G. Balkwill¹ · Julie R. Deslippe¹ · Paul Horton² · Charles David³ · Chen Wu^{4,5} · Emily Koot^{4,6} · Peter Ritchie^{1,4} · Wayne Blissett⁷ · David Chagné^{4,6}

Received: 26 March 2024 / Revised: 8 July 2024 / Accepted: 17 July 2024 / Published online: 6 August 2024
© The Author(s) 2024

Abstract

Aotearoa New Zealand's swamp forests have experienced significant habitat loss in fewer than two hundred years. Many of the country's tree species are endemic with sparse to no genetic information available to underpin conservation strategies. *Syzygium maire*, Aotearoa's only endemic *Syzygium* species, is a culturally and ecologically important component of swamp forest habitats. Unfortunately, populations of *S. maire* have been greatly reduced, heavily fragmented and are susceptible to the emergent pathogen *Austropuccinia psidii* (myrtle rust), posing eminent danger of a further decline of the species. We sought to develop genomic resources to inform conservation management of *S. maire*. To this end, we used long read, high accuracy sequencing technology to produce a highly complete reference quality genome for *S. maire*. The genome sequence was named 'Ngā Hua o te Ia Whenua' by the local Māori tribe where the tree used for genome sequencing grows. We assess whether genome-level divergence with other Myrtaceae may have followed geographic isolation of the species. We detect conservation of large scale synteny between three *Syzygium* species and *Eucalyptus grandis*, providing support for the stability of *Syzygium* genomes across evolutionary time. We annotate genes implicated in fungal pathogen defence, identifying several hundred putative NLR genes, including putative homologs of previously identified *Austropuccinia psidii* resistance genes. Finally, we evaluate the genetic relationships of individuals of a small, isolated population of trees. We find evidence of high levels of kinship and inbreeding within small and isolated *S. maire* populations, informing local-scale conservation strategies for the species. Our findings enable practical conservation actions and provide resources for larger scale studies of *S. maire* and other *Syzygium* species in the future.

Keywords Myrtaceae · Myrtle rust · Maori · Genome assembly · Indigenous genomics

Communicated by L. Bianco

✉ Wayne Blissett
wayne@rantgitaane.iwi.nz

✉ David Chagné
David.Chagne@plantandfood.co.nz

¹ School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand

² Te Ao Turoa Environmental Centre, Rangitāne o Manawātū, Palmerston North (Papaioea), New Zealand

³ The New Zealand Institute for Plant and Food Research Limited (Plant & Food Research), Lincoln, New Zealand

⁴ Genomics Aotearoa, Dunedin, New Zealand

⁵ Plant & Food Research, Auckland, New Zealand

⁶ Plant & Food Research, Palmerston North, Manawātū, New Zealand

⁷ Rangitāne o Manawātū, Palmerston North (Papaioea), New Zealand

Introduction

The global decline of biodiversity is one of the most pressing issues facing humanity. Trees represent approximately 20% of plant species (Beech et al. 2017), and forests contribute significantly to biodiversity (Betts et al. 2017) by providing habitat and food for a complex web of life. Forest trees also support multiple provisioning and regulating services to humankind (Assessment 2005, Seidl et al. 2016). The combined pressures of shifting land use, deforestation, habitat fragmentation, rapidly changing climate and increased pathogen load, however, are destabilizing forest ecosystems, putting the continuity of forest stands and individual tree species at risk (Sala et al. 2000; Potter et al. 2017; Curtis et al. 2018).

Genomics analyses may support the improved conservation and restoration of trees by assessing genetic risk of extinction and non-adaptability and identifying adaptive

genetic variation, which can inform assisted breeding or artificial migration of species. These analyses have become a powerful tool in demographic assessments, enabling variation associated with environmental adaptation to be identified (Hodgins and Moore 2016; García et al. 2018). A major stumbling block in the application of genome-wide studies in conservation, however, has been the difficulty of constructing reference genomes. Previously, hybrid approaches using high-accuracy, short read sequencing combined with low-accuracy, ultra-long read technologies and chromatin-capture (e.g., Hi-C) were required to produce chromosome-spanning assemblies or genetic maps. These approaches are often prohibitively expensive and technically challenging, especially given the typically scant resources of conservation research groups studying rare or endangered organisms. Genomic data is therefore lacking for many species of conservation concern, hampering the use of genomic data to support their conservation and restoration (García et al. 2018).

The rapid construction of near complete reference genomes, however, is becoming more feasible and affordable, thanks to highly accurate, long read sequencing technologies (Hon et al. 2020; Lang et al. 2020). One such technology is PacBio High-Fidelity (HiFi) sequencing, which delivers long sequencing reads (~20 KB) with accuracies that compare favourably with short-read sequencing technologies (Hon et al. 2020). This effectively reduces the input data and downstream processing required for the assembly of accurate and highly complete genomes, reducing the cost barrier to *de novo* assemblies, and making genomic tools more accessible in threatened species conservation.

In Aotearoa New Zealand (hereafter 'Aotearoa'), over 70% of what was once nearly ubiquitous forest cover has been transformed to agricultural and urban uses (Ewers et al. 2006). Some of the starkest examples of degradation are wetland forests (i.e., swamps), where some regions have lost close to 99% of their natural cover (Tomscha et al. 2019), largely due to drainage for pasture. Wetlands are an extremely important ecosystem type across Aotearoa, providing diverse ecosystem services (Tomscha et al. 2021, 2023; Bentley et al. 2022), and are habitat for a wide variety of lifeforms (Gibbs 2000). While restoration of ecosystems has received considerable interest in Aotearoa over the past few decades (Clarkson et al. 2013), there has been limited genetic information to inform these plans (Jamieson et al. 2008). Aotearoa's wetland forests now exist as highly modified and fragmented systems and, as a result, are likely to experience significant genetic erosion (Schlaepfer et al. 2018). These forests are also likely to be more susceptible to the negative genetic effects of fragmentation, as they are often surrounded by non-forest environments (such as pasture) through which forest adapted animal and insect dispersers are less likely to move (Schlaepfer et al. 2018).

Syzygium maire (A. Gunn.) Sykes et Garn.-Jones (swamp maire, maire tawake, waiwaka), is a structurally important mid-canopy tree of Aotearoa's swamp forests, and its fruit are an important food source for native birds (Van Der Walt et al. 2020). Specialised pneumatophores enable it to grow well in perpetually waterlogged soils (Bettoni et al. 2022), making it an important umbrella species of swamp forest habitats. Despite *Syzygium* being the most species rich genus of trees (Low et al. 2022), *S. maire* is the only species endemic to Aotearoa (Dawson et al. 2011). *S. maire* is an important species to Māori (the indigenous people of Aotearoa). Māori share deep cultural and intergenerational connections with the tree through the concept of whakapapa (common ancestry), a duty of guardianship for the species through the concept of kaitiakitanga, and a right of sovereignty within its habitats through the concept of tino rangatiratanga. Further discussion on these concepts and *S. maire* also provided for the day to day health and wellbeing of Māori as a rongoā (traditional medicine), a perfume or air freshener, and a weapons timber for self-defence in close combat as maripa (shark tooth knife), wahaika (hand club) and in fortified Pā defence as kooii (a 3-5 m long slender pointed spear used by two people).

Syzygium maire once grew across much of Te Ika-a-Māui (North Island) and the north eastern tip of Te Waipounamu (South Island) (McCarthy et al. 2019), but is now nearly absent in much of its former range due to habitat conversion (Van Der Walt et al. 2020). Extant populations are small, fragmented and surrounded primarily by extensive pasture (Clarkson 2014). Furthermore, the seeds are highly sensitive to desiccation, do not appear to persist in seedbanks, and are recalcitrant to artificial preservation, including cryopreservation (Van Der Walt et al. 2020, 2023). These factors place *S. maire* in significant danger of inbreeding depression and genetic loss, reducing the species' capacity to adapt to emerging diseases and other consequences of global change. Another major threat to persistence of *S. maire* is a rust disease of Myrtaceae caused by *Austropuccinia psidii*, i.e., myrtle rust. Field observations suggest that young leaves, fruit and flowers are all susceptible to the pathogen (Toome-Heller et al. 2020). Many *S. maire* populations across Aotearoa are already showing signs of infection (pers. obs.) with total reproductive failure possible in some small and fragmented populations, potentially rendering them functionally extinct. Similar observation have been made for *Syzygium* species in Australia (Pegg et al. 2014). As a result of these pressures, the tree is currently listed as nationally critically threatened (De Lange et al. 2018) the highest classification for extinction risk issued by the New Zealand Department of Conservation (Townsend et al. 2008).

Identifying genetic variation that is associated with myrtle rust resistance has the potential to play an important role

in conservation and restoration planning for the species. Previous work on *A. psidii* resistance in Eucalypts (Christie et al. 2016; Santos et al. 2020; Yong et al. 2021), has identified a number of nucleotide-binding leucine-rich repeat (NLR) proteins potentially affecting susceptibility to myrtle rust. NLR genes are primarily involved in pathogen detection and signalling, and are characterised by the presence of nucleotide-binding adaptor shared by APAF-1, resistance proteins, and CED-4 (NB-ARC) domains (McHale et al. 2006). The family can be further split in to two major subfamilies – those containing coiled-coil (CC) or Toll/interleukin-1 receptor (TIR) domains (McHale et al. 2006). A number of other domains involved in cellular signalling and other downstream processes are also present. Many of these genes are characterised by duplication and recombination events, which have been hypothesized to be a driving factor behind evolution of resistance to novel pathogen threats by allowing for diversifying selection to take place (Tobias and Guest 2014). In *Eucalyptus* specifically, five *Ppr* (*Puccinia psidii* resistance 1–5) QTL loci containing 60 genes have been identified (Christie et al. 2016; Santos et al. 2020; Yong et al. 2021). Genes in the *Ppr1* locus have also been implicated in myrtle rust resistance in *Melaleuca* (Hsieh et al. 2018) and *Syzygium* (Tobias et al. 2018). Identifying these genes in *S. maire* may therefore provide information through which myrtle rust resistance could be investigated.

Furthermore, optimal strategies for genetic conservation and restoration of *S. maire* and swamp forests in Aotearoa are currently unexplored. This is despite significant interest in replanting programmes for the species, and availability of both seed and vegetative based propagation (Bettoni et al. 2022). Information on *S. maire*'s genetic diversity would shed light on how the species may respond to, for example, emergent disease and climate change, and could underpin an array of conservation and restoration strategies to protect it from extinction, or at the very least allow diversity-informed preservation.

Here, we aim to establish new genomic resources for a rare and understudied tree species in Aotearoa through constructing an annotated reference genome based on PacBio® HiFi sequencing. We go on to conduct a comparative genome analysis to evaluate the completeness of our assembly and question whether isolation of the species may have resulted in any large-scale genome variation when compared with other Myrtaceae. Given the demonstrable role of NLR and *Ppr* genes specifically in rust pathogen defence among other Myrtaceae, we detect putative resistance genes in the *S. maire* genome, including putative homologs of previously identified *Austropuccinia psidii* resistance. Finally, we evaluate the genetic diversity of a small, isolated population of trees as a preliminary analysis of local scale variation within the species.

Methods

Sample preparation and sequencing

Young leaf tissue was harvested in November 2020 from a mature *S. maire* tree, named Ngā Hua o te Ia Whenua, at a location in the waka kotahi New Zealand Transport Agency Te Apati 'Te Ahuaturanga Manawatū Tararua Highway Project near Otangaki (Ashhurst), 15 km from Papaioea/Palmerston North, Aotearoa within the rohe (traditional lands) of Rangitāne o Manawatū. This site is referred to as Ecoviaduct hereafter (Fig. 1). Leaf tissue was immediately placed on dry ice until samples were transferred later that day to storage at -80 °C. High molecular weight DNA was extracted from nuclei obtained from 20 g of leaf tissue using the methods of Hilario (2018) at Plant and Food Research (Palmerston North, New Zealand). DNA was shipped to Neogen (<https://www.neogen.com/>) USA, where library preparation took place. PacBio® HiFi sequencing was performed on a single SMRT cell, using Sequel 2.0 chemistry. HiFi reads underwent quality control and were assembled at the sequencing facility.

Genome assembly and quality assessment

Pre-processing and raw-read quality assessment

Genome assembly quality can be affected by HiFi adapter sequence contamination (Sim et al. 2022), so we assessed adapter contamination of our data through a BLASTn search against raw HiFi reads for adapter sequences (<https://www.neogen.com/>), with percent sequence identity set to 100 and without masking low-complexity regions (*-dust no*).

We utilized k-mer analysis to evaluate potential sequence contamination from exogenous organisms and estimate genome size, ploidy and heterozygosity of the sequenced individual. Optimal k-mer size was estimated using KmerGenie v1.7051 (Chikhi and Medvedev 2014). Statistics for all sizes between 21 and 1021 were calculated, and that with the maximum k-mers chosen. We computed further k-mer statistics using Jellyfish 2.3.0 with canonical enabled (*-C*) (Marçais and Kingsford 2011) and GenomeScope (version 25/09/2022) (Vurture et al. 2017). Values of *k* = 21 (default) were used as *k* = 601 (maximum identified by KmerGenie) proved too memory intensive and did not result in a markedly different estimation of total genome size.

Genome assembly and quality assessment

We conducted *de novo* genome assembly using Flye v2.8.3 (Kolmogorov et al. 2019) and Hifiasm v0.15.4 (Cheng et al.

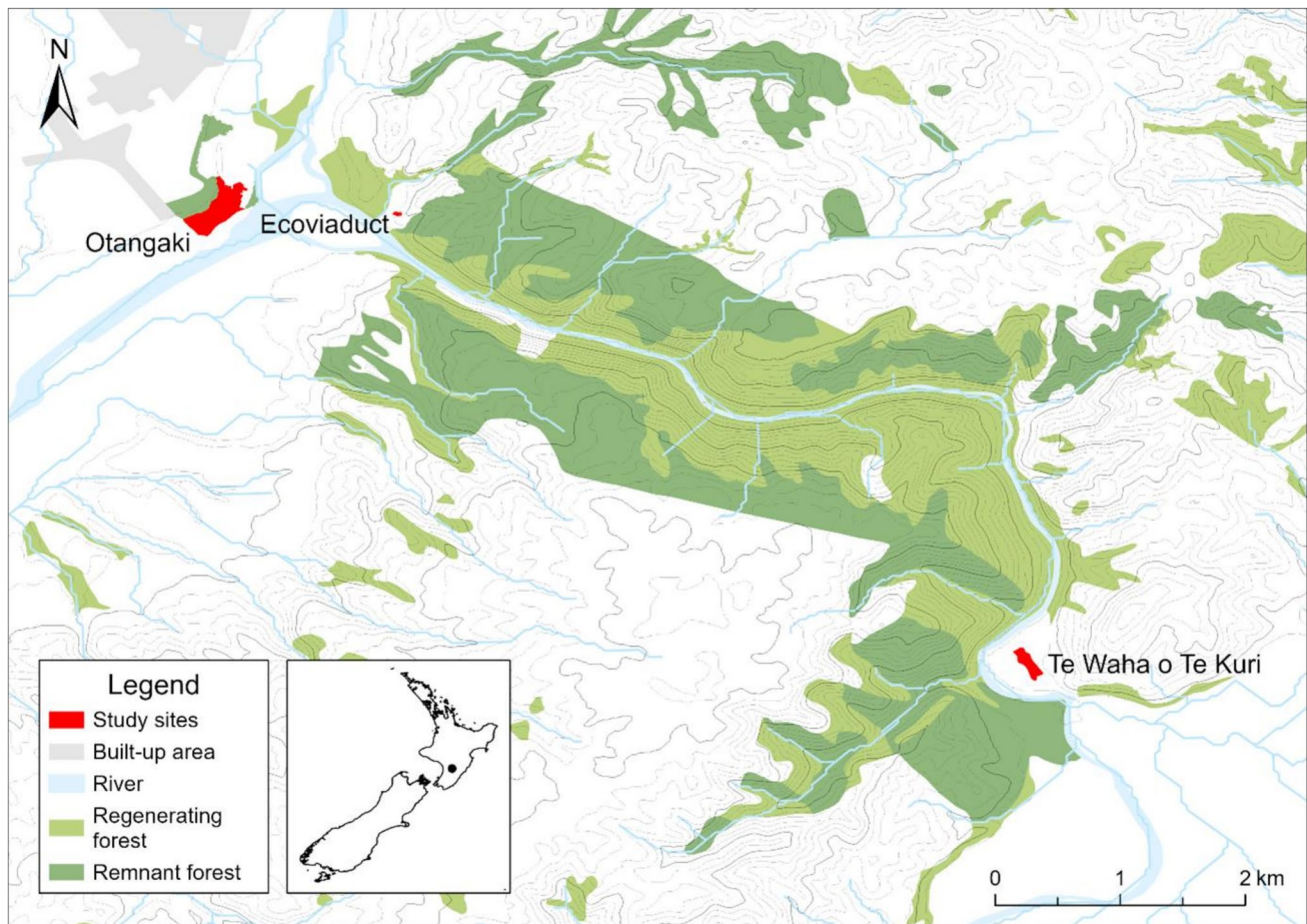


Fig. 1 Distribution and extent of the sampling sites of *S. maire* in the Manawātū region of New Zealand. Leaf samples were collected from three sites (in red), Otangaki (Ashhurst, 7.60 ha), Ecoviaduct (0.16 ha) and Te Waha o Te Kuri (Ferry Reserve, 1.91 ha). Sites flanked a large river gorge (Te Āpiti or Manawātū Gorge) located north-east of Papaioea (Palmerston North, see inset), a major urban centre. The gorge is designated a scenic reserve is forested in roughly

equal proportions of regenerating bush (light green, LCDB5 class 54) and remnant indigenous forest (dark green, LCDB2 class 69). Urban area is depicted in light grey. Contours are spaced at 20 m intervals. Landcover information was derived from the New Zealand Landcover Database version 5 (LRIS), river networks from the River Environment Classification, v2.0 (NIWA) and elevation data at 15 m resolution from NZSoSDEM v1 (Columbus et al. 2011)

2021). The Flye assembly used the settings *-pacbio-hifi* and *-keep-haplotypes* while the Hifiasm assembly was performed using default parameters.

To identify potential contaminating contigs derived from exogenous sources, we used blobtools (Laetsch and Blaxter 2017). HiFi reads were aligned to the Hifiasm assembly using minimap2 v2.24 (Li 2018) *-x map-hifi* (Specify the mapping preset aligning HiFi reads) *-H* (hard clipping alignment mode) and the resulting sam files were converted to bam files with SAMtools v1.9 (Li et al. 2009). We submitted the assembly to megaBLAST (*-max_target_seqs 1 -max_hsps 1 -evalue 1e-10*) to determine the taxonomy for contigs and identified non-plant derived sequences using blobplot. Non-plant sequences were removed from the assembly.

As different assemblers may perform differently depending on sequence characteristics, we chose to explore a hybrid option for our assembly. The hybrid used Flye assembled

contigs to reconcile any gaps in the Hifiasm assembly based on sequence homology. In order to do this, we employed RagTag v2.10 (Alonge et al. 2022) with the patch function, which uses homologous regions of contigs from one assembly to scaffold and fill gaps in another.

Gene content was assessed with BUSCO v5.2.2 (Seppey et al. 2019) using default parameters and lineage set to eudicots_odb10. Completeness was also evaluated by determining the intactness of long-terminal repeat retrotransposons (LTR-RTs), which often lead to regions that are assembled incompletely due to collapse of long repetitive copies (Ou et al. 2018). LTR_FINDER (Xu and Wang 2007) and LTRharvest (Ellinghaus et al. 2008) implemented in GenomTools v1.6.1 (Gremme et al. 2013) were used to detect LTR candidates. We then combined the outputs and passed them to LTR-retriever v2.9.0 (Ou and Jiang 2018) to provide final annotations and calculate the LTR-assembly index (LAI).

Synteny analysis

We conducted a comparative genome analysis to evaluate the completeness of the *S. maire* assembly, identified contigs that may correspond to whole or part of chromosomes, and investigated any potential chromosomal rearrangements. This was accomplished through whole genome comparison of the *S. maire* hybrid assembly to the *Eucalyptus grandis* v2 (Myburg et al. 2014) (downloaded from https://phytozome-next.jgi.doe.gov/info/Egrandis_v2_0), *Syzygium aromaticum* (clove) (Ouadi et al. 2022) (downloaded from <https://zenodo.org/record/6579856>) and *Syzygium grande* (sea apple) (Low et al. 2022) (downloaded from <https://genomevolution.org/coge/GenomeInfo.pl?gid=60239>) genomes using Mashmap v2.0 (Jain et al. 2017). For *E. grandis*, we set the minimum segment length to 5000 bp, percent identity to 95 and kmer length to 16 (length suggested by developers). We used a segment length of 50000 bp for *S. aromaticum* and *S. grande* due to higher sequence similarity, otherwise parameters were identical to those used for *E. grandis*. Dot plots were generated using generatdotplot.pl, a script provided alongside Mashmap, with minor edits.

Based on greater synteny, it was deemed appropriate to scaffold *S. maire* contigs according to the *S. aromaticum* genome assembly. RagTag v2.1.0 scaffold (Alonge et al. 2022) was used to join contigs based on homology with *S. aromaticum*. Gaps were filled with runs of 100 Ns.

Genome annotation

Repeat identification

To identify repeats in the genome assembly we used the Extensive *de novo* TE Annotator (EDTA) v2.1.0 (Ou et al. 2019) with parameters *-anno*, *-sensitive* and *-evaluate*. *S. aromaticum* coding sequences (Ouadi et al. 2022) were provided to the annotator to decrease the likelihood of classifying repetitive coding sequence as transposable elements. In addition, the Telomere Identification Toolkit (tidk) v0.2.2 (<https://github.com/tolkit/telomeric-identifier>) was used to predict telomere repeats (*tidk find -clade plants*) by searching the genome using all known telomeric repeats for plants. These sequences are available in the telomeric repeat database (https://telomerase.asu.edu/sequences_telomere.html), including the most common plant telomeric repeat sequence (TTTAGGG) (Shakirov et al. 2022).

Gene prediction

Gene prediction was performed using MAKER v2.31.9 (Cantarel et al. 2008) with the simple repeat models generated from EDTA. Since there is no transcriptome data available for *S. maire*, we used the *S. aromaticum* transcriptome as

a priori information for MAKER annotation. Splicing-aware gene prediction was informed through provision of predicted proteins from a range of Myrtaceae (*Eucalyptus grandis*, *Psidium guajava*, *Syzygium aromaticum*, *Syzygium grande* and *Syzygium oleosum*). All were downloaded from NCBI during October 2022 barring *S. aromaticum* and *E. grandis* which were obtained from Zenodo (record 6,579,856) and Phytozome 13, respectively. After an initial round of gene prediction, genes were used to train SNAP v2013.11.29 (Korf 2004) and Augustus v3.4.0 (Stanke et al. 2006) using the training pipeline implemented in BUSCO v4.1.4 for the latter. This method is described in more detail at <https://gist.github.com/darencard/bb1001ac1532dd4225b030cf0cd61ce2>. After training, a second round of MAKER was run using predicted genes and trained gene predictors from the first round. A final round of training and MAKER was then run to generate gene models.

Statistics for predicted gene models were calculated using the script *gaas_maker_merge_outputs_from_datastore.pl* (<https://github.com/NBISweden/GAAS>). We analysed the AED distribution and BUSCO completeness for predicted transcripts and selected the prediction round with the highest scores for both metrics for further analysis. InterProScan v5.51–85 (Jones et al. 2014) was used to annotate all predicted proteins according to protein domains.

NLR and *Ppr* gene discovery

We used a pipeline developed by Chen et al. (2023) (<https://github.com/ZhenyanLuo/FindPlantNLRs>) to detect nucleotide-binding leucine-rich repeat (NLR) intracellular immune receptor proteins in the unmasked *S. maire*, *S. aromaticum* and *S. grande* genomes.

In order to understand the evolutionary relationships of NLR genes, the entirety of the NLR gene family in *S. maire* was subject to phylogenetic analysis. We included all identified genes containing NLR domains. Where multiple transcripts had been predicted, the longest one was used. NLR domains (Pfam ID PF00931) were extracted from each protein sequence and aligned using MUSCLE v3.8.31 (Edgar 2004) before undergoing approximate maximum likelihood phylogenetic analysis with FastTree v2.1.11 (Price et al. 2010).

Identification of potential *Ppr* orthologs

Previously identified *Eucalyptus Ppr* genes (Butler et al. 2016; Christie et al. 2016), as compiled by Chakrabarty et al. (2023) were downloaded before performing BLASTp against the *Eucalyptus grandis* v2.0 genome in Phytozome 13 (phytozome-next.jgi.doe.gov) (Goodstein et al. 2012). All hits were downloaded, NLR domains extracted and aligned to the NLR domains for *S. maire* genes identified

above before undergoing approximate maximum likelihood phylogenetic analysis with FastTree v2.1.11 (Price et al. 2010). *S. maire* NLR genes were considered potential orthologs if they were within one common ancestor of an *E. grandis Ppr* gene, occurred on the same chromosome and exhibited similar domain structure.

Local genetic diversity and intra-individual heterozygosity

Low-coverage whole genome resequencing

In order to understand local-scale population genetics for the species, young leaf tissue was harvested from 30 mature *S. maire* trees at three sites near Papaioea (Palmerston North), Aotearoa/New Zealand in the traditional lands of Rangitāne o Manawatū in November 2020. The trees were distributed as follows: 13 trees were located within the same stand as Ngā Hua o te Ia Whenua, the tree from which we derived the reference genome (Ecoviaduct), eight trees were collected from a nearby stand (Otangaki/Ashhurst), and the remaining nine trees were from a third population (Te Waha o Te Kuri/Ferry Reserve) separated from the first two by a large river gorge (Te Āpiti/Manawatū Gorge) (Fig. 1). Upon harvest, the tissue was immediately placed on dry ice and transferred to storage at -80 °C on the same day.

DNA was extracted from young leaf tissue using a modified CTAB protocol. Briefly, we flash-froze the tissue in liquid nitrogen and ground it to a fine powder in a tissue homogeniser (TissueLyser LT). Tissue lysis then took place in 1 ml CTAB extraction buffer (3% CTAB, 3% PVP, 0.1% 2-mercaptoethanol, 100 mM Tris pH 8.0, 25 mM EDTA pH 8.0 and 1.5 M NaCl) at 60 °C for 1 h. We then added one volume of 24:1 chloroform:isoamyl alcohol and gently inverted the tube until homogenized. After centrifugation at 16,000 rcf for 10 min, we carefully removed the supernatant using a wide-bore pipette tip. We then added 4 µl of RNase A and incubated the solution for 30 min at 37 °C. After another round of chloroform:isoamyl alcohol, we used 1 volume of isopropyl alcohol to precipitate DNA. After incubation for 30 min at room temperature, we centrifuged the solution for a further 20 min at 16 rcf to allow formation of a DNA pellet. We washed the pellet twice in 70% ethanol before resuspending it in 100 µl TE buffer.

Dried DNA samples were sent to Neogen (<https://www.neogen.com/>) for library preparation, using the Illumina Nextera Kit chemistry. Sequencing was performed on an Illumina NovaSeq instrument (150 bp, paired end) at an average of 10× coverage per sample. Neogen conducted initial quality filtering so that only samples with a score of Q30 at over 75% of bases were returned.

Variant calling

We assessed read quality using FastQC v0.11.9 (Andrews 2010) and multiQC v1.9 (Ewels et al. 2016) before trimming reads using fastp v0.23.2 (Chen et al. 2018) with adapter detection, overrepresentation analysis, correction and cut right enabled. Next, we mapped the samples using bwa-mem2 (Li 2013) in SAMtools v1.15.1 (Li et al. 2009). The resulting bam files were filtered for a quality score of 20 using SAMtools *view*, mate pairs cleaned with SAMtools *fixmate* and sorted with SAMtools *sort*. To remove PCR duplicates from the mapped reads, we used Picard MarkDuplicates (version 2.7.1; <https://broadinstitute.github.io/picard/>). Finally, the depth and width of mapping coverage were calculated using BEDTools version 2.23.0 (Quinlan and Hall 2010). We performed indel realignment using GATK v3.5 (McKenna et al. 2010) RealignerTarget-Creator to generate target intervals followed by realignment with IndelRealigner.

Variants for each individual were detected using BCFtools v1.15.1 *mpileup* against the unmasked reference genome and called with *bcftools call*. Per contig variant files were concatenated into a single file in *bcftools concat*. Individual depth and missingness statistics were calculated in VCFtools v0.1.15 (Danecek et al. 2011) and individuals with an average coverage of less than 5 or missing information of greater than 10% were removed. Variant calling was then reformed on the dataset minus these individuals. Site level statistics relating to average coverage, missingness, quality and allele frequency were calculated in VCFtools. Sites with depth of less than 5 or more than 12 (twice the mean coverage), missingness greater than 10%, a quality score less than 30 and more than two alleles were also removed. All filtering steps were performed using VCFtools.

Population genetic clustering, summary statistics and kinship

In order to group individuals, K-means clustering was performed using the *find.clusters* function from the *adeigenet* package in R (Jombart 2008). The analysis was performed with the maximum number of clusters (max.n) set at 10 and the number of principal components (n.pca) to the number of individuals in the analysis (30). Choice of the optimal number of principal components clusters (3) was made interactively based on Bayesian information criterion (BIC) score (Supplementary Fig. 4). We also performed principal component analysis to further assess structure in the data. We first replaced missing values with mean allele frequencies. We then used the *dudi.pca* function from *ade4* (Dray and Dufour 2007) without norming for column vectors (scale = FALSE).

Summary statistics were calculated using the *hierfstat* package in R (Goudet 2005). Observed, expected and overall heterozygosity, F_{ST} and F_{IS} were calculated using the *basic.stats* function which uses the methods of Nei (1987). We also calculated Weir and Cockerham's (1984) and Weir and Goudet's (2017) F-statistics as they take sample size into account and we are working with a small dataset with unequal number of individuals per population. These calculations were made with the *hierfstat* function *wc* and *betas* with 1000 bootstraps for the latter. Pairwise F_{ST} values between sites as well as 2.5 and 97.5% confidence intervals (bootstraps set to 100) were calculated with the *genet.dist* function using method WC84 (Weir and Cockerham 1984) and *boot.ppfst* (bootstraps set to 100) respectively, as well as *pairwise.betas* (Weir and Goudet 2017) together with *boot.ppbetas*.

Kinship and inbreeding for each individual was calculated using the *popkin* function in the *popkin* (Ochoa and Storey 2021) package in R with subpopulations set according to collection site. Values in the kinship matrix were scaled through setting the minimum kinship value to zero with the *rescale_popkin* function. Individual-level pairwise F_{ST} was also estimated with the package through the *pwfst* function. Any negative F_{ST} values were set to zero.

Data sovereignty

The *S. maire* specimens used in this study were from the traditional lands of Rangitāne o Manawatū, a mana whenua iwi (tribe) of the lower half of Te Ika-a-Māui/North Island of Aotearoa New Zealand. Therefore, Rangitāne o Manawatū is the rightful guardian of the *S. maire* specimens used in this study, including raw and analysed data derived from them. Rangitāne o Manawatū's authority over their *S. maire* gives effect to the Treaty of Waitangi, a treaty signed in 1840 between Māori chiefs and the British Crown. Internationally, Māori ownership and control of natural resources are supported by the UN Declaration on the Rights of Indigenous Peoples 2007. The Treaty of Waitangi and genomics research ethics in Aotearoa is discussed further by Morgan et al. (2019), Hudson and Russell (2009) and Collier-Robinson et al. (2019). Biocultural (BC) Labels and Notices were submitted by Rangitāne o Manawatū and Plant & Food Research, respectively through the Local Contexts Hub (localcontexts.org) and the Aotearoa Genomics Data Repository under project "Maire Tawake Genome".

Results

Pre-processing and raw read quality assessment

PacBio HiFi sequencing yielded 1,003,855 high accuracy reads of Q29 or above, representing 19.5 GB of data with an

average read length of 19.4 KB. No contamination from HiFi adapters or k-mer peaks, which could suggest contaminating sequence data, were observed. K-mer analyses at $k=21$ estimated the *S. maire* genome size at 351 MB (Supplementary Fig. 1), while that at $k=601$ suggested a size of 362 MB. The coverage for heterozygous regions was 25x. K-mer frequency distribution showed one major peak (at 50x), with a smaller peak at a coverage half that of the major peak (25x) (Supplementary Fig. 1).

Genome sequencing and *de novo* assembly

Both Flye and Hifiasm produced assemblies with high contiguity (Table 1, Fig. 2). The Hifiasm assembly was more contiguous than that generated by Flye, with N50 of 22 Mb and 2.9 Mb and 324 and 764 contigs for Hifiasm and Flye, respectively. The size distribution of contigs from Hifiasm and Flye also indicated a more contiguous assembly using Hifiasm. The Hifiasm assembly L50 was 7 (against 42 for Flye) and the 23 largest contigs (greater than 1 MB) represented 93% (375 MB) of the total (400 MB) assembly.

In order to improve assembly contiguity, a Hifiasm-Flye hybrid assembly was constructed. Significant improvements, including a 46% increase in N50 and assembly of an additional telomere-to-telomere contig (putative chromosome 3) was noted (Table 1, Fig. 1). This assembly was then scaffolded using the *S. aromaticum* genome as reference, resulting in 95% being represented in one of 11 chromosomes. The average number of contigs in scaffolds was 1.2.

Assembly completeness

Completeness measures were similar across all assemblies (Table 1). Gene-based assessment suggested that assemblies were highly complete, each containing over 98% of 2,326 complete Basic Universal Single Copy Ortholog (BUSCO) eudicot genes. Low BUSCO duplication levels were also observed (2.3, 2.4 and 2.2% for Hifiasm, Flye, and the hybrid, scaffolded assembly respectively). The genome wide LTR Assembly Index (LAI) was 11.37 for the Hifiasm

Table 1 Completeness information for two genome assemblers Hifiasm and Flye using a BUSCO and LAI analysis

	Flye	Hifiasm	Scaffolded assembly
Complete BUSCOs (%)	98.1	98.3	98.3
Complete and single-copy BUSCOs (%)	95.7	96.0	96.1
Complete and duplicated BUSCOs (%)	2.4	2.3	2.2
Fragmented BUSCOs (%)	0.6	0.6	0.6
Missing BUSCOs (%)	1.3	1.1	1.1
LAI	12.51	11.37	11.55

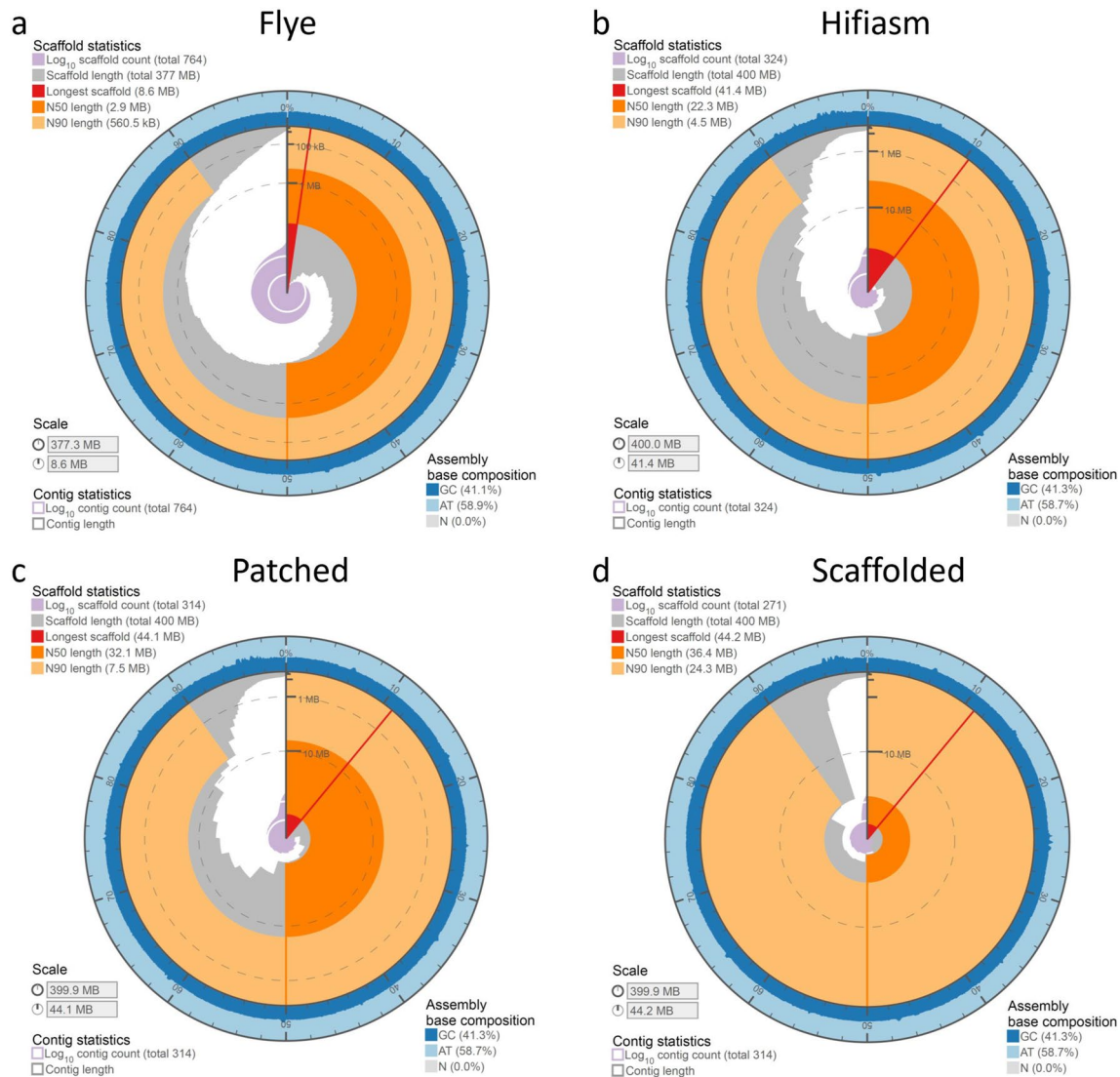


Fig. 2 Summary statistics of the *S. maire* draft assemblies. Snail plot representations for the (a) Flye, (b) Hifiasm, (c) Hifiasm-Flye hybrid and (d) scaffolded genome assemblies. The outer ring (blue) represents the GC content and assembly position (shown as percent of the

total assembly). Contigs are arranged from smallest to largest, and thus the largest contig (in red) represents just over 2% of the flye and 10% of the other three assemblies. Snail plots were generated using assembly-stats (<https://github.com/rjchallis/assembly-stats>)

assembly and 12.51 for Flye. LAI was marginally improved in the hybrid assembly (11.55). Finally, telomeres were detected at both ends of 8 scaffolds, with the remainder having telomeres at one end (Supplementary Table S1).

Naming of the Genome

The final *S. maire* genome has been named by Rangitāne o Manawātū as “Ngā Hua o te Ia Whenua” (translating in English as “The Fruits of Ia Whenua”). The name talks about the characteristics of its habitat, namely the flora and fauna including its vast undergrowth. The phrase ‘ia whenua’ also

speaks of the lay of the land, its waterways, the life force, its sacredness and freedom from restrictions. This describes, explains and locates the uniqueness of maire tawake. The te reo (Māori language) description of the *S. maire* genome name by Rangitāne o Manawātū is: Nā Rangitāne tēnei Maire Tawake i tautapa ki “Ngā Hua o te Ia Whenua”. E kōrero ana tēnei ingoa mō te āhua o te whaitua e tū rā ia, arā a ngāi tipu, a ngāi kīrehe, tae atu ki tōna wao nui. E kōrero hoki ana te ‘ia whenua’ mō te takoto o te whenua, ōna wai, tōna mauri, ōna tapu, ōna noa anō hoki. The assembled chromosomes were named according to the prefix IW for ‘ia whenua’, followed by the chromosome number.

Synteny analysis with Myrtaceae genomes

Syntenic conservation was observed between the *S. maire* assembly and the genomes of *S. aromaticum* (clove), *S. grande* (sea apple) and *E. grandis* (rose gum) (Fig. 3, Supplementary Fig. 2). The greatest number of structural differences existed between *E. grandis* and *Syzygium* spp., as has been previously reported for *S. aromaticum* (Ouadi et al. 2022) (Supplementary Fig. 2). High genome

collinearity and no obvious rearrangements were observed for *S. maire* versus *S. aromaticum*. A small inversion on one end of putative chromosomes 5 was observed between *S. maire*/*S. aromaticum* and *S. grande*. Of the 21 *S. maire* contigs over 1 MB, five align directly to entire chromosome scaffolds in both other *Syzygium* spp. (Supplementary Table S1, Fig. 3, Supplementary Fig. 2). The remaining six chromosomes are represented primarily by large contigs, with no obvious sections missing. Alignment of

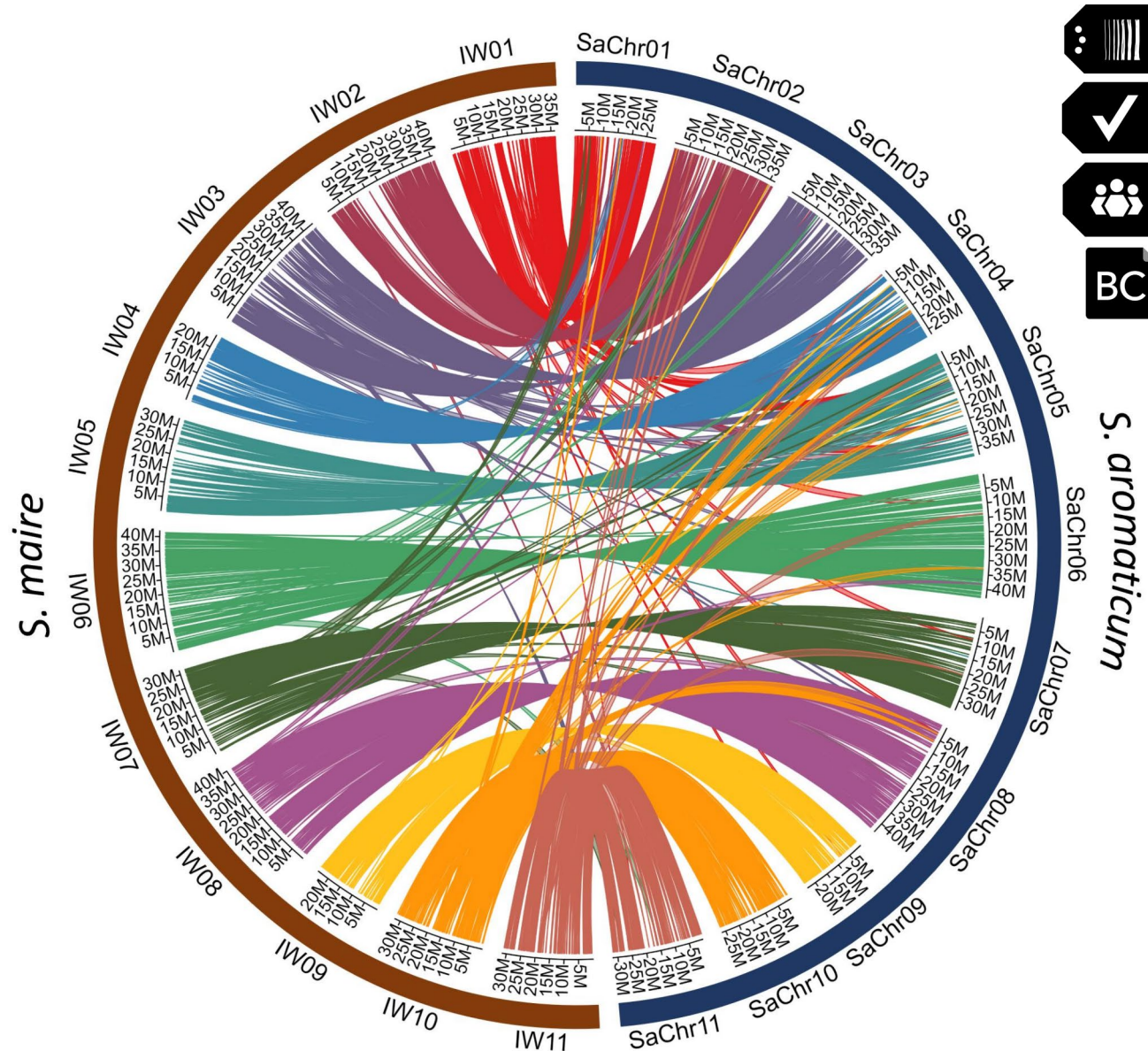


Fig. 3 Circa plot depicting genome-wide synteny between the Rag-Tag scaffolded *S. maire* assembly and the 11 *S. aromaticum* chromosomes. The outermost track represents scaffold size. Each line represents a 50000 bp sequence region with >95% identity. Scaffolds are colour coded and named according to the most similar *S. aromaticum*

chromosome. Circos plots were created in Circa (<http://omgenomics.com/circa>). Biocultural (BC) Labels and Notices were submitted by Rangitāne o Manawātū and Plant & Food Research, respectively. BC Labels: provenance, consent verified, open to collaboration. BC Notice: biocultural

the *S. maire* genome to itself did not identify any regions mapping to multiple parts of the genome, apart from some small regions clustered at the centre or ends of contigs (Supplementary Fig. 2).

The evidence of synteny between *S. maire*, *S. aromaticum* and *E. grandis* was used to ascribe chromosome identities to the *S. maire* contigs. Numbering was according to the system used for *Eucalyptus* (Supplementary Table S1).

Genome annotation

Approximately 183 MB or 46% of the genome was predicted to be repetitive (Supplementary Table S2). Of this, the majority (24.24%) LTR sequence primarily from the Tys3 superfamily of transposable elements (15.94%). Tys3 was previously known as Gypsy, but the name has been reviewed (Wei et al. 2022). The second most abundant

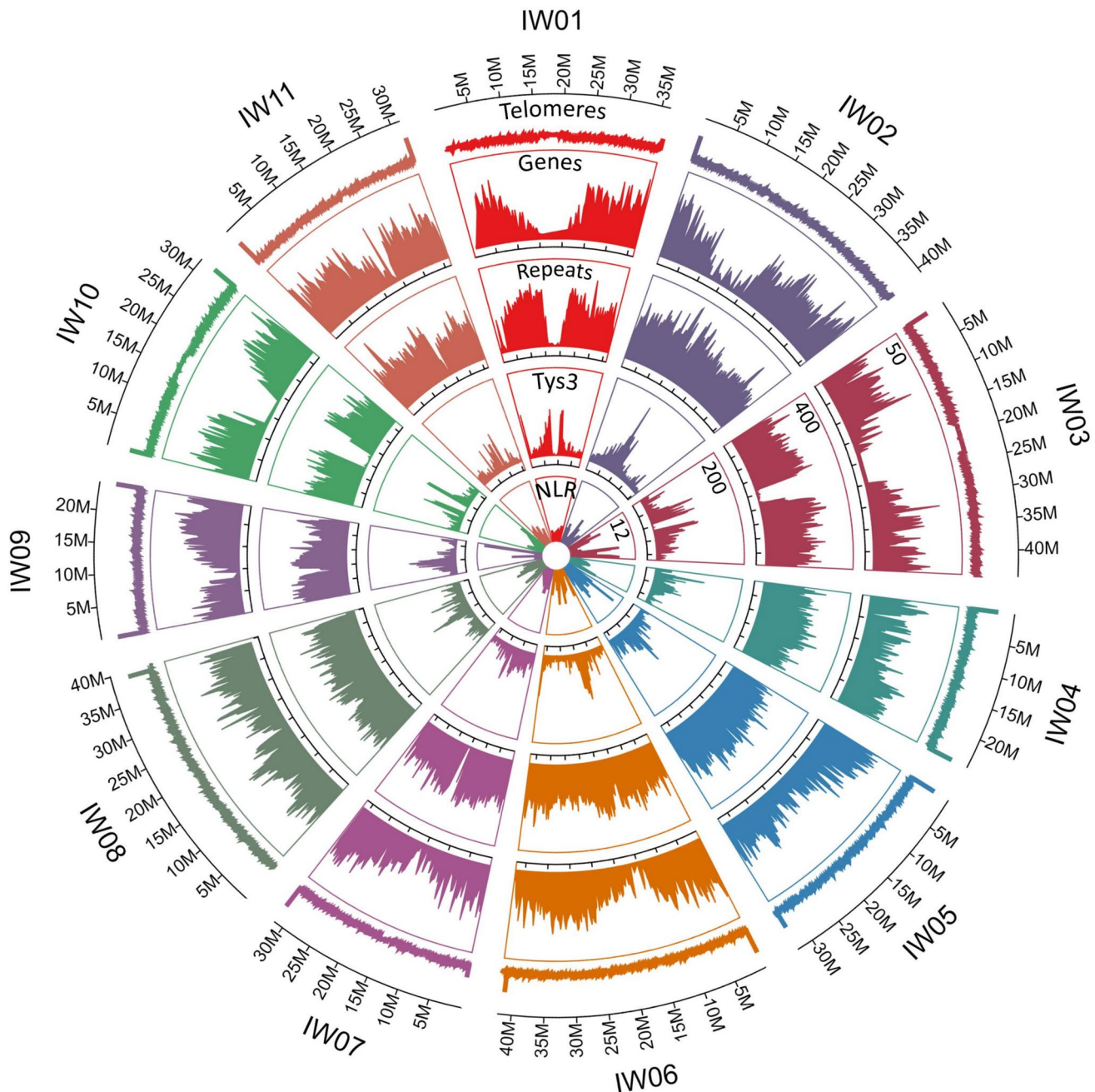


Fig. 4 Circa plot representing genome feature density for the 11 largest scaffolds from the *S. maire* genome assembly. The outermost plot depicts contig size, followed by telomeric repeat sequence density, gene density, overall repeat density, Tys3 class LTR density and

finally NLR gene density. All statistics were calculated on 250,000 bp windows. Y-axis limits for each track are shown in the top left corner of each plot of chromosome three. Circos plots were created in Circa (<http://omgenomics.com/circa>)

repeat class was terminal inverted repeat (TIR) at 14.34%. For the 11 chromosome-level scaffolds, eight showed a high density of telomere repeats at both ends while three had these features at one end (Supplementary Table S1, Fig. 4). Repeats were unevenly distributed across the assembly (Fig. 4), with Tys3 class repeats most abundant near regions of low gene density (Fig. 4).

Gene prediction identified 32,307 putative genes (Supplementary Table S3) representing approximately 29% of the assembly. Coding regions alone accounted for 9.8% of the assembly with an average length of 1202 bp and 5.3 exons per gene. The number of mRNAs was identical to the number of predicted genes. Genes were detected for 161 contigs, with the majority (31,550) contained within the chromosome level scaffolds (Fig. 4). Gene distribution appears to be reasonably uniform, although areas of lower gene density do occur, especially at the centres of IW01, IW02, IW03 and IW10 (Fig. 4).

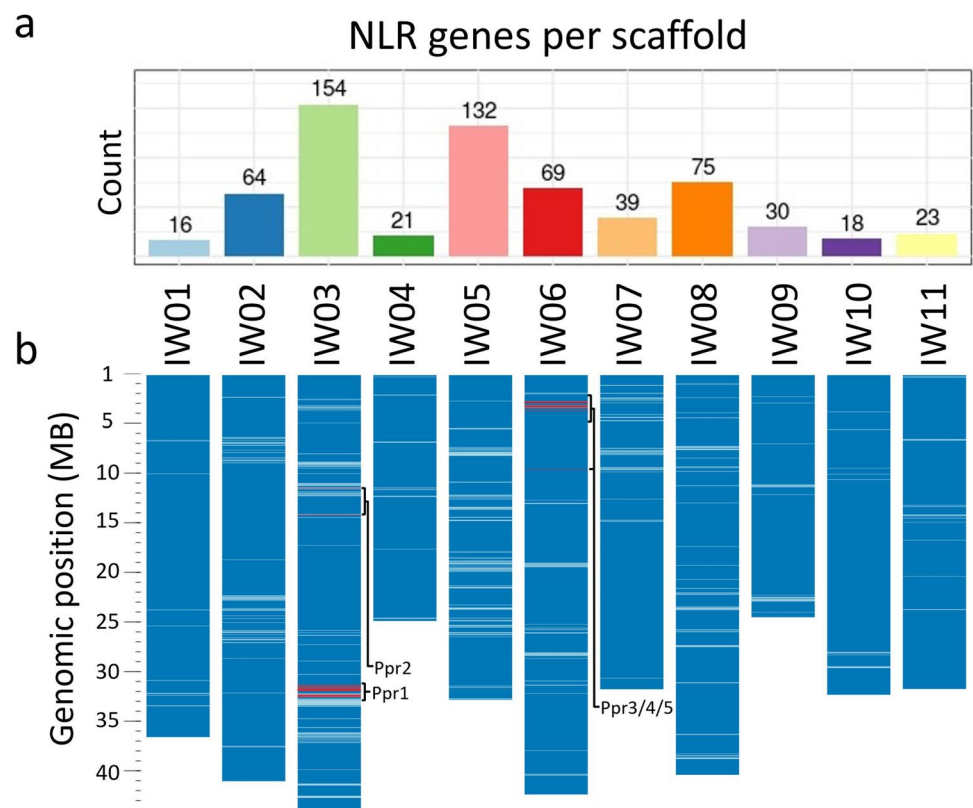
NLR and *Ppr* gene discovery

We identified 642 NLR class genes in *S. maire*, 562 in *S. aromaticum* and 867 in *S. grande* (Supplementary Table S3). NLRs were detected on all 11 putative *S. maire* chromosomes, with the majority on IW03 (154) and IW05 (132) (Fig. 5). IW02, 03, 05, 06 and 09 all had 250 KB regions

with high NLR gene density (Figs. 4 and 5). NLR genes contained a wide range of associated domains (Fig. 6). There were a total of 333, 270 and 402 TIR-type genes for *S. maire*, *S. aromaticum* and *S. grande* respectively, with the remainder representing non-TIR type NLRs. These included Rx N-terminal (PF18052—RxBN/RxNL), and *Arabidopsis* broad-spectrum mildew resistance protein RPW8 (PF05659—RNB/RNL) genes. Genes on IW03 were exclusively of the TNB/TNL class.

Putative orthologs for the majority of *Eucalyptus Puccinia psidii* resistance (*Ppr*) genes, in all five previously associated loci were identified (Supplementary table S4, Supplementary Fig. 3). We identified 18 potential *Ppr1* orthologs, two *Ppr2*, and 12 which are most closely related to *Eucalyptus Ppr3/4/5* genes. In the case of *Ppr1* genes, possible tandem duplication of an *S. maire Ppr1* ortholog for Eucgr.C0278 was identified (Supplementary Fig. 3), suggesting expansion of this clade. Within this cluster, two genes with novel architecture were also present. These contained, in addition to NB-ARC/TIR domains, GAG-polypeptide of LTR copia-type (PF14223), GAG-pre-integrase (PF13976), integrase core (PF00665) and reverse transcriptase domains (PF07727). The cluster contained approximately one gene every 70 KB. Expansion of the *Ppr3/4/5* group does not seem to be the case in *S. maire*, as it does in *Eucalyptus*.

Fig. 5 Nucleotide-binding leucine-rich repeat (NLR) gene number and position per scaffold. **(a)** Predicted NLR gene count per scaffold. **(b)** Genomic position of predicted NLR and *Ppr* genes, each light line represents a single gene. NLR genes are shown in light blue, while *Ppr* genes are in red. Genomic position is shown on the y-axis



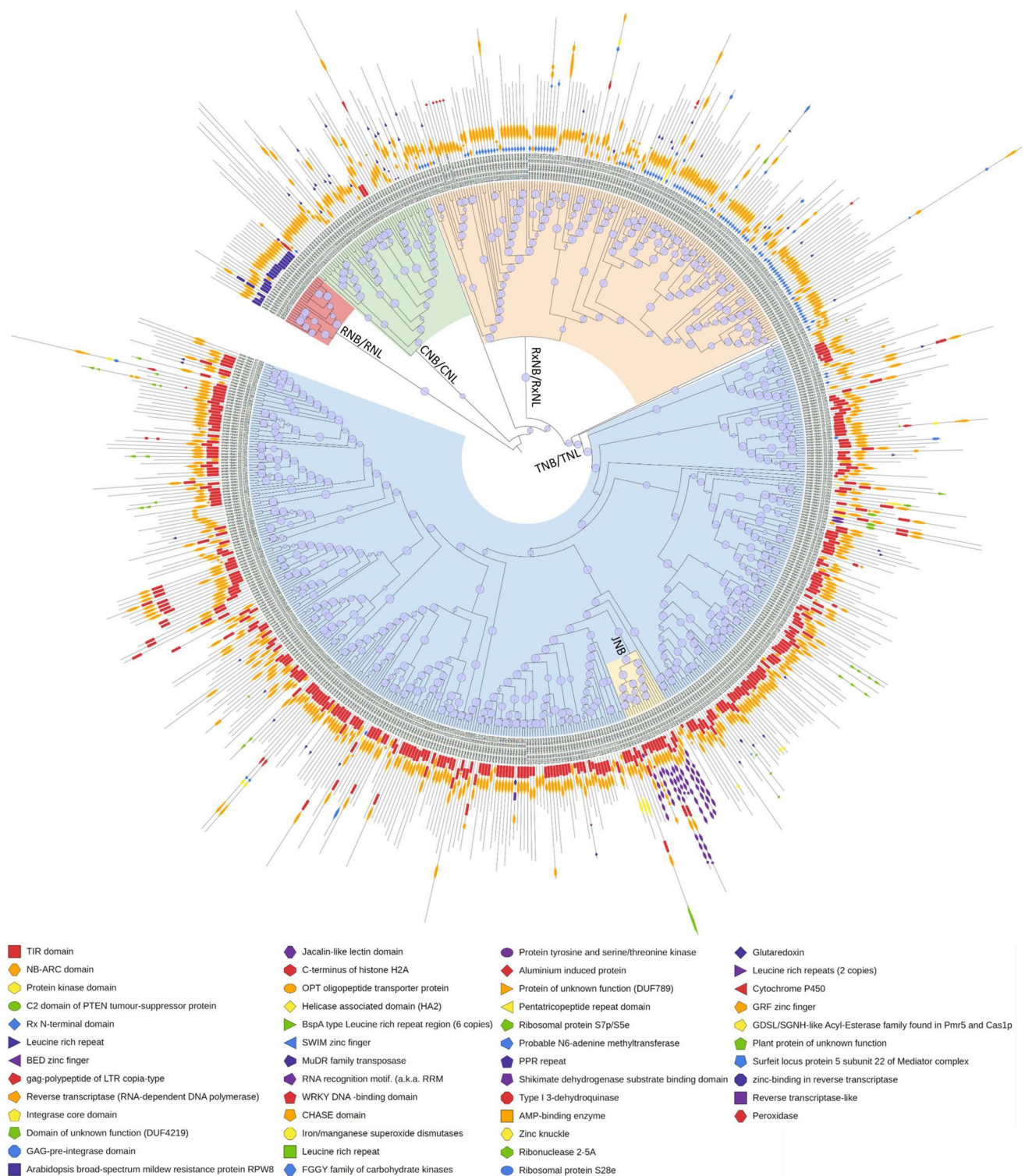


Fig. 6 Phylogenetic tree for the *S. maire* nucleotide-binding leucine-rich repeat (NLR) gene family. The NB-ARC domain for the longest transcripts for all 512 predicted NLR genes were aligned using MUSCLE v3.8.31. The tree was calculated using approximate maximum

likelihood implemented in FastTree. Pfam domain architecture for each gene is depicted in the outer ring. Gene subfamilies are represented by coloured areas on the plot. The figure was created using Interactive Tree Of Life (iTOL) v5 (Letunic and Bork 2021)

Local genetic diversity

We assessed local genetic diversity in the 30 re-sequenced *S. maire* individuals. We identified 1,288,573 raw SNPs from the dataset. After quality filtering, 536,542 SNPs were retained.

Based on BIC decay (Supplementary Fig. 4) and K-means clustering, it was estimated that three clusters exist in the dataset. These clusters roughly corresponded to sampling sites, with cluster 2 comprising solely Te Waha o Te Kuri individuals and cluster 3 Ecoviaduct individuals. However, four Ecoviaduct trees grouped together in cluster 1 with those sampled at Otangaki. Ngā Hua o te Ia Whenua grouped with the Ecoviaduct only cluster (cluster 3). PCA analysis suggested similar groupings (Fig. 7), with trees from each site forming distinct clusters along PC1, which described 20.4% of the total variance. PC2, which described an additional 10.2% of total variance separated the Te Waha o Te Kuri trees from the Otangaki trees, with those from Ecoviaduct occupying an intermediate position along PC2.

Observed, expected and total heterozygosity for the entire dataset were 0.23, 0.24 and 0.29, respectively (Table 2). Overall F_{ST} were similar (0.148 vs 0.198 and 0.060 vs 0.075) for Weir and Cockerham's (1984) and Nei's (1987)

Table 2 Global population genetic statistics for all loci across all individuals

	Ho	Hs	Ht	F_{ST}	F_{IS}
Nei (1987)	0.230	0.239	0.292	0.180	0.040
Weir and Cockerham (1984)	-	-	-	0.181	0.088
Weir and Goudet (2017) ($\pm 97.5, 2.5\%$ confidence interval)				0.176 ± 0.0008	

estimates, while F_{IS} was low for both (0.04 vs 0.09). Weir and Goudet's (2017) F_{ST} was marginally higher at 0.176. Population specific F_{ST} for Ecoviaduct (0.32) was almost three times that of Otangaki (0.11) and Te Waha o Te Kuri (0.10) (Table 3). Pairwise F_{ST} for the subpopulations were highest for Ecoviaduct versus Otangaki at 0.236 than Ecoviaduct-Te Waha o Te Kuri (0.15) or Otangaki-Te Waha o Te Kuri (0.14) (Supplementary table S5).

Demographics of the subpopulations appeared to be reflected by kinship, with higher kinship values observed between individuals of the same subpopulation (Fig. 8a). Mean kinship for the entire population was 0.13. Kinship

Fig. 7 Principal component analysis of 536,542 SNPs across 30 individuals from three sampling sites in the Manawatu region of New Zealand. Sampling sites are depicted according to colours with Ecoviaduct (red), Otangaki (blue) and Te Waha o Te Kuri (green) represented. The inset displays the percent contribution of each eigenvector to differentiation between individuals. The majority of variation is explained by axis 1 (21.5%) followed by axis 2 (10.3%). Biocultural (BC) Labels and Notices were submitted by Rangitāne o Manawātū and Plant & Food Research, respectively. BC Labels: provenance, consent verified, open to collaboration. BC Notice: biocultural

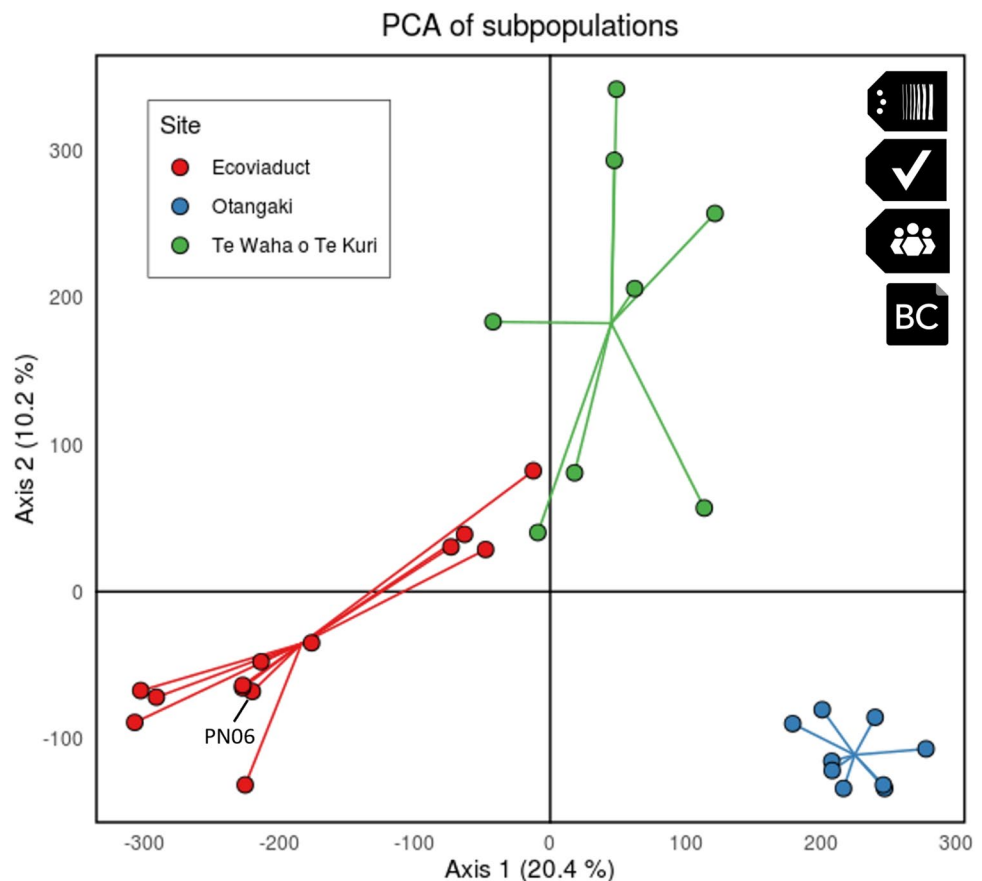


Table 3 Population specific F_{ST} for subpopulations calculated as in Weir and Goudet (2017)

Population	F_{ST}	97.5% CI	2.5% CI
Ecoviaduct	0.317	0.319	0.316
Otangaki	0.107	0.109	0.106
Te Waha o Te Kuri	0.104	0.106	0.103

appeared to be highest among trees in the Ecoviaduct site, with the greatest levels of kinship and inbreeding observed for three trees at this site (PN07, PN08 and PN12). However, some of the lowest levels of inbreeding were also observed in this subpopulation. The four trees (PN03, PN11, PN13 and PN14) that grouped with the Otangaki population in previous analyses showed the lowest kinship with the Ecoviaduct cohort and were not more highly related to each other either. In general, trees in Ecoviaduct and Otangaki appear to be more related to each other than either subpopulation is to those in Te Waha o Te Kuri. Pairwise F_{ST} for pairs of individuals reflected some of these trends (Fig. 8b). Individuals within stands generally shared lower pairwise F_{ST} than individuals in different stands. High values were however observed for some individuals within stands, with certain individuals appearing to be distinct from almost all other individuals in the dataset. This was especially obvious

for the three trees with the highest relative kinship and inbreeding.

Discussion

We have constructed a highly contiguous, near complete reference genome for *Syzygium maire*, a critically threatened swamp forest tree species endemic to Aotearoa. The tree from which the genome was derived, and thus the genome, has been named Ngā Hua o te Ia Whenua by Rangitāne o Manawatū, and this name will be used going forward. It is expected that this resource will be used to enhance conservation efforts by facilitating genomics studies aimed at understanding the distribution of genetic diversity, the evolutionary and adaptive genetic characters of the species, and potential mechanisms to resist emerging pathogens and other agents of global change. We offer two such examples of the use of this genome to inform conservation of the species. Firstly, we annotated NLR resistance genes which may have population-level variation providing resistance to infection by *Austropuccinia psidii* (myrtle rust), a primary threat to survival of the species. Secondly, we describe high levels of kinship and inbreeding within a small remnant population of trees which are isolated and fragmented because of habitat loss. This information will aid restoration efforts by

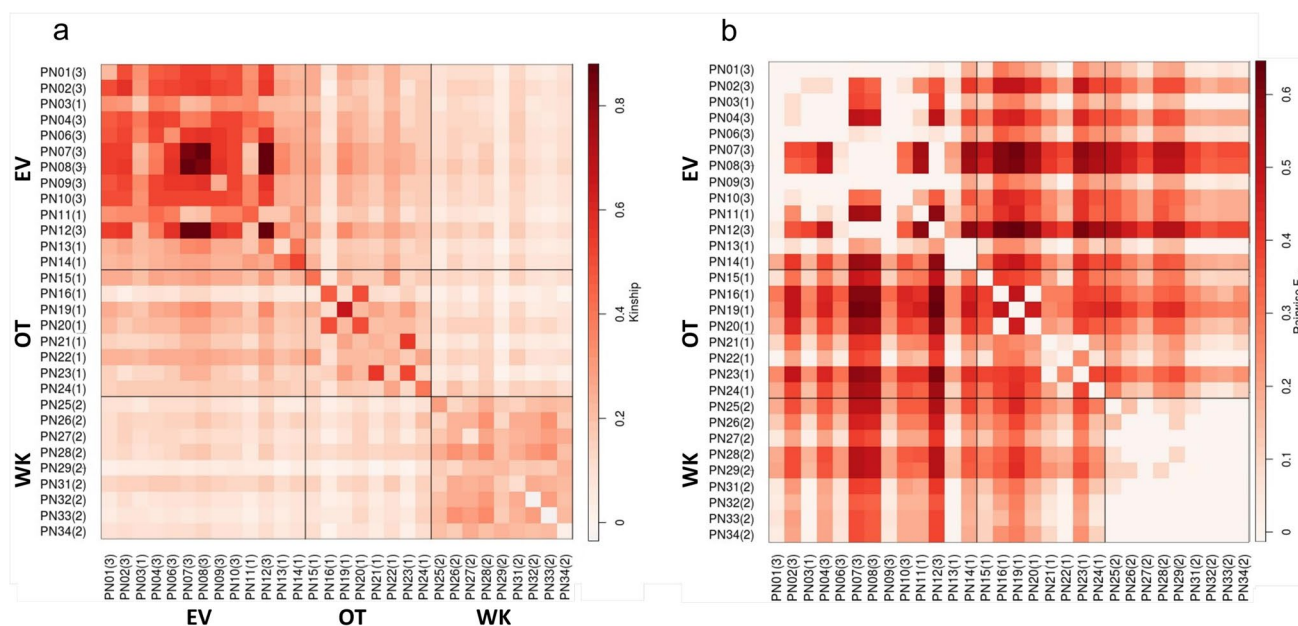


Fig. 8 Heatmap of kinship and pairwise-interindividual F_{ST} for three subpopulations of *S. maire* based on 536,542 SNPs. **(a)** kinship and inbreeding coefficients and **(b)** pairwise-interindividual F_{ST} . Black lines depict the boundaries of each subpopulation, Ecoviaduct (EV), Otangaki (OT) and Te Waha o Te Kuri (WK). Each sample ID is followed by its respective K-means clustering grouping in parentheses.

Kinship values per pairs of individuals are represented at the intersection of two individuals. Values were scaled according to the minimum kinship coefficient calculated. Inbreeding coefficients are displayed on the diagonal (self-vs-self). Pairwise F_{ST} was calculated for each pair of individuals relative to a putative ancestral population. Negative F_{ST} values were set to zero

informing seed sourcing and assisted migration of individuals to these and nearby intact or restoring swamp forests.

High-accuracy, long read sequencing technology (PacBio® HiFi) enabled us to assemble a highly contiguous reference genome at lower cost and with less computational demands than would be required with short-read sequencing technologies. We trialled two assemblers (Hifiasm and Flye) to assess which would produce the best assembly. While the Hifiasm was more contiguous, and therefore chosen as our working assembly, gap filling with the Flye assembly in a hybrid approach resulted in improved contiguity. This demonstrates that exploiting different strengths of multiple assemblers can be a viable strategy for improving genome assemblies. The assembly size (400 MB) fell within the range of previously published *Syzygium* genomes, which vary from 370–1788 MB, the majority at the lower end (Ohri et al. 2004; Matra et al. 2019; Low et al. 2022; Ouadi et al. 2022). The assembly statistics were comparable to similarly sized Myrtaceae genomes assembled using HiFi reads (Driguez et al. 2021; Feng et al. 2021; Chen et al. 2023). Low BUSCO duplication levels suggest that contigs were not duplicated in the assembly, nor was there evidence of any major genome duplication in the lineage. LAI scores fell within what is considered appropriate for reference-level assemblies, where anything below 10 is classified as draft, and over 20 is a “gold standard” (Ou et al. 2018). Our results demonstrate the ability of high accuracy long read sequencing technologies to produce *de novo* reference standard genomes for rare and endangered species with sufficient ease and quality to become a key resource for future genomics studies. This may be especially relevant for Myrtaceae as they generally have moderately sized genomes (Grattapaglia et al. 2012), enabling reference quality assemblies to be obtained from relatively small amounts of data.

As Myrtaceae are generally diploid with 11 chromosomes stable genome structure (Grattapaglia et al. 2012; Butler et al. 2017; Ouadi et al. 2022), synteny between genomes provides complimentary evidence of assembly integrity. Our study species is diploid, with $2n = 22$ (Dawson 1987), and all unscaffolded contigs map exclusively to single regions on one of the 11 pseudochromosomes in *S. aromaticum*, *S. grande* and *E. grandis*, providing no evidence of misassembly or lineage specific duplication events. Congruently, these observations suggest conserved genome structure among these species. *S. maire* seems to conform to the norm for this family, exhibiting a relatively small genome with no novel genome-wide duplications and a similar structural arrangement to other *Syzygium* spp. The combination of conserved genome size, structural arrangement of chromosomes and synteny, suggests there are not significant large-scale differences between the *Syzygium* species genomes assessed here either. Previously it has been hypothesised that genome duplication can provide evolutionary redundancy

among genomic regions, and allow opportunities for adaptation and species divergence (Taylor et al. 2001). The genus *Syzygium*, however, appears, in most cases, to have similarly sized genomes and lacks lineage-specific genome duplication or structural rearrangement, despite being the most speciose tree genus on the planet (Craven and Biffin 2010; Low et al. 2022). A small structural difference relative to *S. grande* was observed for both *S. aromaticum* and *S. maire*. *S. aromaticum* and *S. grande* are more closely related to each other than *S. maire*, suggesting that this may be a rearrangement specific to the *S. grande* lineage. Rearrangements when compared to *E. grandis* appear to match those observed in *S. aromaticum* (Ouadi et al. 2022). These results lend further support to the hypothesis that conservation of synteny amongst the Myrtaceae and *Syzygium* spp. is high (Grattapaglia et al. 2012; Low et al. 2022).

The proportion of repeat sequence was similar to the similarly sized *S. aromaticum* and *P. guajava* assemblies (Feng et al. 2021; Ouadi et al. 2022) but more than the smaller *M. quinquenervia* genome (Chen et al. 2023). Chen et al. (2023) suggest that the lower repeat content in *M. quinquenervia* is likely responsible for its smaller genome size, and given the high genome stability in Myrtaceae, it is not unlikely that much of the difference in genome size in the family is due to variation in repeat content. The number of genes we detected was again within the range of other Myrtaceae, although higher than both clove (27,528), *M. quinquenervia* (Chen et al. 2023) (28,744) and guava (25,601), and similar to *L. scoparium* (31,220) but far less than the 36,000 reported for *E. grandis* (Myburg et al. 2014) or 39,903 for *S. grande* (Low et al. 2022). The absolute gene content in the Myrtaceae therefore varies quite considerably, and seems not to be strictly associated with overall genome size.

Previous studies have observed that Tys3 class repeats may be associated with centromeric regions (Wu et al. 2022), suggesting that regions of high Tys3-class and low gene density observed here may correspond to centromeres. For most contigs, large peaks in Tys3-class repeats and troughs in gene density were observed in a single region. The presence of a low number of these regions and their occurrence at non-telomeric ends suggests that we may have partially assembled putative centromeric regions for the majority of contigs.

We successfully annotated several hundred NLR genes each for *S. maire*, *S. aromaticum* and *S. grande*. *S. grande* appeared to have far more NLR genes than either other species. The majority of NLR genes were classified as TIR-type, similar to other Myrtaceae (Chakrabarty et al. 2023; Chen et al. 2023). The distribution of NLR genes across the *S. maire* genome was nonuniform, with a number of regions showing far higher density, a phenomenon also observed for *M. quinquenervia* (Chen et al. 2023). IW03 and IW05 were particularly enriched for these

genes, with those on IW03 belonging almost exclusively to the TIR-subfamily. This clustering of resistance genes has been observed previously for other plant species and is hypothesized to aid in evolution of recognition of a wide range of pathogens. Specifically, different loci may host genes providing resistance to different pathogens (Barbosa-da-Silva et al. 2005; Butler et al. 2016; Christie et al. 2016; van Wersch and Li 2019), while variation and duplication of genes at loci could enable plants to detect novel or variable pathogen strains (van Wersch and Li 2019). Furthermore, haplotypic genetic diversity within NLR genes in single genomes is essential to their function (Chen et al. 2023). Studies exploring not only inter-species, but also population and individual level variation in these genes are therefore necessary to fully uncover their role in conferring resistance to environmental challenge.

Given myrtle rust's significant effect on biodiversity, uncovering potential resistance mechanisms is an active field of study. Previous research has identified NLR (*Ppr*) genes putatively associated with beneficial effects on plant survival (Tobias et al. 2018; Santos et al. 2020; Yong et al. 2021). Our study has identified dense clusters of NLR genes, including potential orthologs for *Ppr* genes, on chromosome IW03 and IW06. We observe far fewer *Ppr* orthologs than reported for *Eucalyptus*, highlighting the evolutionary plasticity of these genes. In a study on *Syzygium luehmannii*, two specific TIR-NBS-LRR type transcripts were identified as strongly associated with resistance to the pathogen (Tobias et al. 2018). Intriguingly, these genes mirror homologs identified in *E. grandis* *Ppr1* locus on chromosome 3 associated with myrtle rust resistance. The conservation of the myrtle rust-resistance phenotype attributed to genes at this loci across diverse Myrtaceae lineages suggests that homologs in the *S. maire* genome would also be involved in determining susceptibility to this pathogen. To afford the best chance of resistance mechanism discovery, multiple level analyses should be promoted. For instance, metabolomics analysis lead to the identification of numerous secondary metabolites responsible for various myrtle rust resistance phenotypes in *M. quinquenervia* (Moffitt et al. 2022).

Understanding vulnerability to inbreeding and the loss of genetic variation as proxy for adaptive potential is a critical component of informing forest restoration, especially in regions with small populations or where local sourcing of seed may be preferred. How to define seed sourcing boundaries, and which germplasm to focus conservation efforts on are also critical questions (Heenan et al. 2023). Analysis of genetic structure and fixation within and between populations can provide evidence on which restoration decisions can be made.

Genetic distinction between the three small populations assessed here appear reasonably high, suggesting there are three genetically distinct clusters and some level of genetic

isolation between sites. Pairwise F_{ST} at the population level suggested that Ecoviaduct and Otangaki were the most distinct from each other, despite being geographically proximal. The difference between Otangaki and Te Waha o Te Kuri and Ecoviaduct and Te Waha o Te Kuri was only 0.03, however, even though Otangaki and Ecoviaduct share trees of the same cluster, are geographically closer and on the same side of a potential barrier to gene flow (Te Āpiti/Manawatū Gorge). It therefore does not appear that the gorge has had a noticeable effect on segregating these populations, but that other factors are at play.

High kinship values and high levels of inbreeding in the Ecoviaduct population may be partly responsible for the apparent genetic distinctiveness and inflated F_{ST} of the sub-population. Indeed, four individuals grouping together with Otangaki trees are less related to each other and all other trees at Ecoviaduct. Highly inbred individuals also showed some of the highest individual level pairwise F_{ST} values, further suggesting some of the genetic structure apparent in the analysis may be a result of inbreeding. The overall fixation index, however, was low and heterozygosity reasonably high and as such there are many outbred individuals within the dataset, providing genetic substrate for conservation.

Pairwise F_{ST} calculated among individuals suggested that Ecoviaduct and Te Waha o Te Kuri were less distinct compared to Otangaki (Fig. 8b). The Otangaki site also showed relatively low kinship levels with both stands, although kinship with Ecoviaduct was higher. These findings, together with the clustering, suggest that Otangaki may have been a larger population than Ecoviaduct, with some genetic exchange taking place across Te Āpiti/Manawatū Gorge, while isolation at Ecoviaduct has led to higher kinship and inbreeding. As Ecoviaduct is such a small population however, these findings may simply reflect the chance survival of a few locally highly related individuals during land clearance.

Due to the close spatial proximity of *S. maire* trees within swamp-forest stands (usually separated by no more than a few metres), localised structure is not surprising. Wotton and Kelly (2012) found a negative correlation between frugivore and seed size versus dispersal distance. This included the major fruit disperser of *S. maire*, a large wood pigeon (the kererū), which has a maximum seed dispersal distance of ~1.5 km, with the distance of dispersal at less than 100 m from the feeding site. Gamba and Muchhala (2020) have shown that animal pollinated plants develop genetic structure over short distances but that method of seed dispersal does not seem to have an effect. The dispersal of *Syzygium maire* relies on both animal pollination and seed deposition and it is likely that the short distances travelled by dispersers in this case increase local structure. Many seeds also fall and germinate directly under the parent tree. Local structure as a result of this dispersal state may be especially the case

in naturally and artificially patchy habitats, as there may be no suitable habitat within dispersal distance, reducing the chance of migration between subpopulations.

In a restoration context, it may therefore be advisable to maintain some level of distinction between these populations, especially Ecoviaduct/Otangaki to the west of Te Āpiti/Manawatū Gorge and Te Waha o Te Kuri to the east. Seed sourcing may best be focussed on Otangaki and Te Waha o Te Kuri trees, attempting to collect from as many adults as possible in order to capture the reasonably high diversity which appears to exist at these sites. However, it is not generally known how diverse the species is, whether these distinctions may break down at a larger spatial scale, nor what effects fragmentation and isolation of these subpopulations may be having on seed genetics. To improve recommendations for the conservation and restoration of *S. maire* more genetic data is required – both for mature trees across the geographic distribution of the species and for seed sourced from fragmented stands.

Conclusion

In conclusion, we have assembled a reference quality genome for *Syzygium maire*, a Myrtaceae species endemic to Aotearoa. Synteny across the genus and the Myrtaceae suggests high levels of genome stability among members of the family. Furthermore, we have demonstrated the usefulness of reference genomes to conservation through annotating genes potentially involved in determining susceptibility to *Austropuccinia psidii* and by revealing genetic structure among and within three subpopulations of *S. maire* nearby the tree from which the reference genome was derived. These findings allow us to make practical restoration recommendations for the species in a local context.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11295-024-01659-5>.

Acknowledgements We thank Pāpā Manu Kawana, Hohepa Isaac-Sharland and Terry Hapi from Rangitāne o Manawatū for the naming of the swamp maire genome, William Brokenshaw and Aaron Tautau (Te Ao Turoa Environmental Centre) for helping with sample collections, Renata Blissett (Rangitāne o Manawatū) and Keith Funnell (Plant & Food Research Palmerston North) for providing liaison between the research team and Rangitāne o Manawatū, Peter Bellingham (Manaaki-Whenua Landcare Research) and Mark Bullians (Plant & Food Research Auckland) for support with the Ngā Rakau Taketake project, Deepa Bowatte (Plant & Food Research Palmerston North) for support with DNA extraction, Prof Bruce Weir for discussions about kinship analysis, Dr Peri Tobias for discussions about NLR annotation and Karen van der Walt for insight into the ecology of swamp maire. Computational analyses were performed with the support of the New Zealand eScience Infrastructure (NeSI, <https://www.nesi.org.nz/>).

Author contributions CB performed sampling and data analysis with assistance from EMK, CW, CD, DC and PH. DC, JRD and PR aided in conceiving and developing the study. PH and WB provided ecological and cultural knowledge of the species and development of protocols relating to Māori knowledge and ownership of the data. CB wrote the manuscript with input from all authors.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions Sequencing was supported by a New Zealand Bioheritage Ngā Rakau Taketake project (<https://bioheritage.nz/about-us/nga-rakau-taketake/>, grant number 2021–44-017) allocated to DC and EMK. Laboratory and operational costs for the project were funded through a Wellington City Council Our Natural Capital Research Grant (grant number E3784 – 3339), and a New Zealand Bioheritage Ngā Rakau Taketake (grant number E3784 – 3339) allocated to JRD, a Genomics Aotearoa grant (grant number E3870- 2940) allocated to CB and JRD, and a Tararua Tramping Club and Wellington Botanical Society grant allocated to CB. DC, CD, CW and PR are supported by the Ministry for Business Innovation and Employment platform Genomics Aotearoa (<https://www.genomics-aotearoa.org.nz/>). CB was supported by a Te Herenga Waka Victoria University of Wellington Doctoral Scholarship.

Data archiving statement Raw data, assembled genome files and annotation information are available from the Aotearoa Genomic Data Repository (Te Aika et al. 2023) at <https://data.agdr.org.nz/> (<https://doi.org/10.57748/1q9v-pr20>) with permission access to be consented from Rangitāne o Manawatū. Biocultural Labels and Notices were associated with the data by Rangitāne o Manawatū and Plant & Food Research, respectively under Local Contexts (localcontexts.org) project 'Maire Tawake Genome'.

Declarations

Conflicts of interest The authors do not have any conflicts of or competing interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* 23:1–19
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Cambridge, United Kingdom, Babraham Bioinformatics, Babraham Institute
- Assessment ME (2005) Ecosystems and human well-being. Island Press, Washington, DC

- Barbosa-da-Silva A, Wanderley-Nogueira AC, Silva RR, Berlarmino LC, Soares-Cavalcanti NM, Benko-Iseppon AM (2005) In silico survey of resistance (R) genes in *Eucalyptus* transcriptome. *Genet Mol Biol* 28:562–574
- Beech E, Rivers M, Oldfield S, Smith P (2017) GlobalTreeSearch: The first complete global database of tree species and country distributions. *J Sustain* 36:454–489
- Bentley SB, Tomscha SA, Deslippe JR (2022) Indicators of wetland health improve following small-scale ecological restoration on private land. *Sci Total Environ* 837:155760
- Bettoni JC, van der Walt K, Souza JA, McLachlan A, Nadarajan J (2022) Sexual and asexual propagation of *Syzygium maire*, a critically endangered Myrtaceae species of New Zealand. *New Zealand J Bot* 62(1):35–52
- Betts MG, Wolf C, Ripple WJ, Phalan B, Millers KA, Duarte A, Butchart SH, Levi T (2017) Global forest loss disproportionately erodes biodiversity in intact landscapes. *Nature* 547:441–444
- Butler J, Freeman JS, Vaillancourt R, Potts BM, Glen M, Lee DJ, Pegg GS (2016) Evidence for different QTL underlying the immune and hypersensitive responses of *Eucalyptus globulus* to the rust pathogen *Puccinia psidii*. *Tree Genet Genomes* 12:1–13
- Butler J, Vaillancourt R, Potts BM, Lee DJ, King GJ, Baten A, Shepherd M, Freeman JS (2017) Comparative genomics of *Eucalyptus* and *Corymbia* reveals low rates of genome structural rearrangement. *BMC Genomics* 18:1–13
- Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196
- Chakrabarty S, Hsieh J-F, Chakraborty P, Foley WJ, Külheim C (2023) Evolutionary relationship of the NBS-LRR gene family in *Melaleuca* and *Eucalyptus* (Myrtaceae). *Tree Genet Genomes* 19:25
- Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890
- Chen SH, Martino AM, Luo Z, Schwessinger B, Jones A, Tolessa T, Bragg JG, Tobias PA, Edwards RJ (2023) A high-quality pseudo-phased genome for *Melaleuca quinquenervia* shows allelic diversity of NLR-type resistance genes. *GigaScience* 12:102
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* 18(12):170–175
- Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30:31–37
- Christie N, Tobias PA, Naidoo S, Külheim C (2016) The *Eucalyptus grandis* NBS-LRR gene family: physical clustering and expression hotspots. *Front Plant Sci* 6:1238
- Clarkson BR, Ausseil AE, Gerbeaux P (2013) Wetland ecosystem services. In: Dymond JR (ed) *Ecosystem services in New Zealand: conditions and trends*. Manaaki Whenua Press, Lincoln, pp 192–202
- Clarkson W (2014) Waiwaka: Taranaki's forgotten swamp tree
- Collier-Robinson L, Rayne A, Rupene M, Thoms C, Steeves T (2019) Embedding indigenous principles in genomic research of culturally significant species. *N Z J Ecol* 43:1–9
- Columbus J, Sirguy P, Tenzer R (2011) A free fully assessed 15 metre digital elevation model for New Zealand. *Survey Quarterly* 66:16–19
- Craven LA, Biffin E (2010) An infrageneric classification of *Syzygium* (Myrtaceae) Blumea-Biodiversity. *Evol Biogeogr Plants* 55(94):99
- Curtis PG, Slay CM, Harris NL, Tyukavina A, Hansen MC (2018) Classifying drivers of global forest loss. *Science* 361:1108–1111
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158
- Dawson M (1987) Contributions to a chromosome atlas of the New Zealand flora—29 Myrtaceae. *NZ J Bot* 25:367–369
- Dawson J, Lucas R, Connor J, Brownsey PJ (2011) New Zealand's native trees. Craig Potton Nelson, New Zealand
- De Lange PJ, Rolfe JR, Barkla JW, Courtney SP, SM B, KA F, B I, S I, H-W R (2018) Conservation status of New Zealand indigenous vascular plants, 2017. Department of Conservation, Wellington
- Dray S, Dufour A-B (2007) The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 22:1–20
- Driguez P, Bougouffa S, Carty K, Putra A, Jabbari K, Reddy M, Soppe R, Cheung MS, Fukasawa Y, Ermini L (2021) LeafGo: Leaf to Genome, a quick workflow to produce high-quality *de novo* plant genomes using long-read sequencing technology. *Genome Biol* 22:1–18
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797
- Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* 9:1–14
- Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048
- Ewers RM, Kliskey AD, Walker S, Rutledge D, Harding JS, Didham RK (2006) Past and future trajectories of forest loss in New Zealand. *Biol Cons* 133:312–325
- Feng C, Feng C, Lin X, Liu S, Li Y, Kang M (2021) A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (*Psidium guajava*). *Plant Biotechnol J* 19:717–730
- Gamba D, Muchhala N (2020) Global patterns of population genetic differentiation in seed plants. *Mol Ecol* 29:3413–3428
- García C, Guichoux E, Hampe A (2018) A comparative analysis between SNPs and SSRs to investigate genetic variation in a juniper species (*Juniperus phoenicea ssp. turbinata*). *Tree Genet Genomes* 14:87
- Gibbs JP (2000) Wetland loss and biodiversity conservation. *Conserv Biol* 14:314–317
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186
- Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes* 5:184–186
- 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 Genet Genomes* 8:463–508
- Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinf* 10:645–656
- Heenan, P. B., W. G. Lee, M. S. McGlone, J. K. McCarthy, C. M. Mitchell, M. J. Larcombe, and G. J. Houlston. 2023. Ecosourcing for resilience in a changing environment. *New Zealand Journal of Botany*:1–26. <https://doi.org/10.1080/0028825X.2023.2210289>
- Hilario E (2018) Plant nuclear genomic DNA preps. protocols.io. <https://doi.org/10.17504/protocols.io.rncd5aw>
- Hodgins KA, Moore JL (2016) Adapting to a warming world: Ecological restoration, climate change, and genomics. *Bot Soc Am* 103(4):590–592
- Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* 7:1–11

- Hsieh J-F, Chuah A, Patel HR, Sandhu KS, Foley WJ, Külheim C (2018) Transcriptome profiling of *Melaleuca quinquenervia* challenged by myrtle rust reveals differences in defense responses among resistant individuals. *Phytopathology* 108:495–509
- Hudson ML, Russell K (2009) The Treaty of Waitangi and research ethics in Aotearoa. *J Bioeth Inq* 6:61–68
- Jain C, Diltthey A, Koren S, Aluru S, Phillippy AM (2017) A fast approximate algorithm for mapping long reads to large reference databases. In: Cenk Sahinalp S (ed) International Conference on Research in Computational Molecular Biology. Springer, pp 66–81
- Jamieson IG, Grueber CE, Waters JM, Gleeson DM (2008) Managing genetic diversity in threatened populations: a New Zealand perspective. *New Zealand J Ecol* 130–137
- Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37:540–546
- Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:1–9
- Laetsch DR, Blaxter ML (2017) BlobTools: Interrogation of genome assemblies. *F1000Research* 6:1287
- Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* 9:123
- Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 49:W293–W296
- Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPPD (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079
- Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. Accessed Mar 2023
- Low YW, Rajaraman S, Tomlin CM, Ahmad JA, Ardi WH, Armstrong K, Athen P, Berhaman A, Bone RE, Cheek M (2022) Genomic insights into rapid speciation within the world's largest tree genus *Syzygium*. *Nat Commun* 13:1–15
- Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770
- Matra DD, Ritonga AW, Natawijaya A, Poerwanto R, Siregar UJ, Widodo WD, Inoue E (2019) Datasets for genome assembly of six underutilized Indonesian fruits. *Data Brief* 22:960–963
- McCarthy JK, Richardson SJ, Cooper JA, Bellingham PJ, Wiser SK (2019) Species distribution models of the native New Zealand Myrtaceae. Ministry for Primary Industries, Wellington
- McHale L, Tan X, Koehl P, Michelson RW (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol* 7:1–11
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303
- Moffitt MC, Wong-Bajracharya J, Shuey LS, Park RF, Pegg GS, Plett JM (2022) Both constitutive and infection-responsive secondary metabolites linked to resistance against *Austropuccinia psidii* (Myrtle Rust) in *Melaleuca quinquenervia*. *Microorganisms* 10:383
- Morgan ER, Perry NB, Chagné D (2019) Science at the intersection of cultures—Māori, Pākehā and mānuka. *N Z J Crop Hort Sci* 47:225–232
- Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D (2014) The genome of *Eucalyptus grandis*. *Nature* 510:356–362
- Nei M (1987) Molecular evolutionary genetics. Columbia University Press
- Ochoa A, Storey JD (2021) Estimating FST and kinship for arbitrary population structures. *PLoS Genet* 17:e1009241
- Ohri D, Bhargava A, Chatterjee A (2004) Nuclear DNA Amounts in 112 Species of Tropical Hardwoods—New Estimates. *Plant Biol* 6:555–561
- Ou S, Jiang N (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* 176:1410–1422
- Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* 46:e126–e126
- Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* 20:1–18
- Ouadi S, Sierro N, Goepfert S, Bovet L, Glauser G, Vallat A, Peitsch MC, Kessler F, Ivanov NV (2022) The clove (*Syzygium aromaticum*) genome provides insights into the eugenol biosynthesis pathway. *Commun Biol* 5:1–13
- Pegg GS, Giblin F, McTaggart A, Guymer G, Taylor H, Ireland K, Shivas RG, Perry S (2014) *Puccinia psidii* in Queensland, Australia: disease symptoms, distribution and impact. *Plant Pathol* 63:1005–1021
- Potter KM, Jetton RM, Bower A, Jacobs DF, Man G, Hipkins VD, Westwood M (2017) Banking on the future: progress, challenges and opportunities for the genetic conservation of forest trees. Springer
- Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842
- Sala OE, Chapin FS, Armesto JJ, Berlow E, Bloomfield J, Dirzo R, Huber-Sanwald E, Huenneke LF, Jackson RB, Kinzig A (2000) Global biodiversity scenarios for the year 2100. *Science* 287:1770–1774
- Santos SA, Vidigal PM, Guimarães LM, Mafia RG, Templeton MD, Alfenas AC (2020) Transcriptome analysis of *Eucalyptus grandis* genotypes reveals constitutive overexpression of genes related to rust (*Austropuccinia psidii*) resistance. *Plant Mol Biol* 104:339–357
- Schlaepfer DR, Braschler B, Rusterholz HP, Baur B (2018) Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere* 9:e02488
- Seidl R, Spies TA, Peterson DL, Stephens SL, Hicke JA (2016) Searching for resilience: addressing the impacts of changing disturbance regimes on forest ecosystem services. *J Appl Ecol* 53:120–129
- Seppy M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol (clifton, NJ)* 1962:227–245
- Shakirov EV, Chen JJ-L, Shippen DE (2022) Plant telomere biology: the green solution to the end-replication problem. *Plant Cell* 34:2492–2504
- Sim SB, Corpuz RL, Simmonds TJ, Geib SM (2022) HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence

- of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* 23:1–7
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34:W435–W439
- Taylor JS, Van de Peer Y, Meyer A (2001) Genome duplication, divergent resolution and speciation. *Trends Genet* 17:299–301
- Te Aika B, Liggins L, Rye C, Perkins EO, Huh J, Brauning R, Godfery T, Black MA (2023) Aotearoa genomic data repository: an āhuru mōwai for taonga species sequencing data. *Mol Ecol Resour*. <https://doi.org/10.1111/1755-0998.13866>
- Tobias PA, Guest DI (2014) Tree immunity: growing old without antibodies. *Trends Plant Sci* 19:367–370
- Tobias PA, Guest DI, Külheim C, Park RF (2018) De novo transcriptome study identifies candidate genes involved in resistance to *Austropuccinia psidii* (myrtle rust) in *Syzygium luehmannii* (Riberry). *Phytopathology* 108:627–640
- Tomscha S, Deslippe J, de Róiste M, Hartley S, Jackson B (2019) Uncovering the ecosystem service legacies of wetland loss using high-resolution models. *Ecosphere* 10:e02888
- Tomscha SA, Bentley S, Platzer E, Jackson B, de Roiste M, Hartley S, Norton K, Deslippe JR (2021) Multiple methods confirm wetland restoration improves ecosystem services. *Ecosyst People* 17:25–40
- Tomscha S, Jackson B, Benavidez R, de Róiste M, Hartley S, Deslippe J (2023) A multiscale perspective on how much wetland restoration is needed to achieve targets for ecosystem services. *Ecosyst Serv* 61:101527
- Toome-Heller M, Ho W, Ganley R, Elliott C, Quinn B, Pearson H, Alexander B (2020) Chasing myrtle rust in New Zealand: host range and distribution over the first year after invasion. *Australas Plant Pathol* 49:221–30
- Townsend AJ, de Lange PJ, Duffy CA, Miskelly CM, Molloy J, Norton DA (2008) New Zealand threat classification system manual. *Dep Conserv Wellingt* 16:2008–2011
- Van der Walt K, Kemp P, Sofkova-Bobcheva S, Burritt DJ, Nadarajan J (2020) Seed development, germination, and storage behaviour of *Syzygium maire* (Myrtaceae), a threatened endemic New Zealand tree. *New Zealand J Bot* 59(2):198–216
- van der Walt K, Nadarajan J, Mathew L, Bettoni JC, Souza JA (2023) Advances in cryopreservation of *Syzygium maire* (swamp maire, maire tawake) zygotic embryos, a critically endangered tree species endemic to New Zealand. *Front Conserv Sci* 4:1269881
- van Wersch S, Li X (2019) Stronger when together: clustering of plant NLR disease resistance genes. *Trends Plant Sci* 24:688–699
- 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
- Wei K, Aldaimalani R, Mai D, Zinshteyn D, Satyaki P, Blumenstiel JP, Kelleher ES, Brooks E (2022) Rethinking the “gypsy” retrotransposon: A roadmap for community-driven reconsideration of problematic gene names. <https://doi.org/10.31219/osf.io/fma57>
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370
- Weir BS, Goudet J (2017) A unified characterization of population structure and relatedness. *Genetics* 206:2085–2103
- Wotton DM, Kelly D (2012) Do larger frugivores move seeds further? Body size, seed dispersal distance, and a case study of a large, sedentary pigeon. *J Biogeogr* 39:1973–1983
- Wu C, Deng C, Hilario E, Albert NW, Lafferty D, Grierson ER, Plunkett BJ, Elborough C, Saei A, Günther CS (2022) A chromosome-scale assembly of the bilberry genome identifies a complex locus controlling berry anthocyanin composition. *Mol Ecol Resour* 22:345–360
- Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265–W268
- Yong WTL, Ades PK, Runa FA, Bossinger G, Sandhu KS, Potts BM, Tibbitts JF (2021) Genome-wide association study of myrtle rust (*Austropuccinia psidii*) resistance in *Eucalyptus obliqua* (subgenus *Eucalyptus*). *Tree Genet Genomes* 17:31

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.