Phylogenomics and Evolution of Polyploid *Azorella* (Apiaceae) in New Zealand

A thesis presented in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

in

Plant Biology

at Massey University, Manawatū, New Zealand.

Weixuan Ning

2023

This PhD thesis was undertaken in the Tate Lab, Massey University, by Weixuan Ning, from Nov 2018 to August 2023. The PhD thesis was supervised by Associate Professor Jennifer Tate (Massey University), Dr Heidi Meudt (Museum of New Zealand Te Papa Tongarewa), Professor Peter Lockhart (Massey University), and Professor William Lee (Landcare Research).

Thesis abstract

Polyploid plants have more than the usual two sets of chromosomes in every cell. Analysing the macroevolutionary patterns of polyploid plants can provide further insight into the mechanisms of polyploidization or whole genome duplication (WGD) in driving species diversification. The polyploid-rich lineage, *Azorella*, in New Zealand (NZ) has two sections, *Schizeilema* and *Stilbocarpa*, with a total of 17 described polyploid taxa (species, subspecies, or varieties) in three known ploidy levels (4*x*, 6*x* and 10*x*). The divergent leaf morphologies and distinct distribution range of polyploid taxa in NZ *Azorella* makes this lineage an ideal system to investigate the macroevolutionary outcomes of WGD in a polyploid-rich lineage.

This thesis aimed to 1) resolve the origins and species relationships of NZ *Azorella* using phylogenetic inference, and 2) compare the polyploidy-associated genomic, morphological, and ecological traits to understand the post-WGD diversification of *Azorella* polyploids.

In this thesis (Chapter 1), I first reviewed the current phylogenomic approaches for resolving species relationships in groups that have complex evolutionary histories, including polyploidization and reticulation. To resolve the NZ *Azorella* phylogenetic relationships (Chapter 2), I applied Hyb-Seq of the Angiosperms353 bait set via Illumina sequencing to amplify 353 target-enriched single copy nuclear genes. Additionally, nrDNA and whole chloroplast DNA were recovered via genome-skimming reads to represent high copy genes/regions that are traditionally used in phylogenetics. Hyb-Seq of Angiosperms353 loci was combined with a PacBio sequencing run to improve homeologous gene extraction (Chapter 3). Finally, NZ *Azorella* post-polyploidization diversification patterns (Chapter 3) were assessed using the variation in genome sizes (via flow cytometry), stomatal guard cell length (using scanning electron microscopy), and ecological niches (using the R package ENMTools).

Overall, from biogeographical analyses, I found two independent dispersal events of species in New Zealand *Azorella* sections *Stilbocarpa* and *Schizeilema*, respectively. Using the concordance factors among gene trees and single nucleotide polymorphisms from Hyb-Seq data, as well as the topological incongruence between single copy and high copy gene trees, the results indicated hybrid origins of several hexaploid (6*x*) species, reticulate relationships among tetraploids, and an allopolyploid origin of the 10*x* species *A. colensoi*. Furthermore, different post-polyploidization diversification patterns were compared among *Azorella* taxa in different ploidy levels, which showed that phylogenetic relationships (i.e., genome content), reticulate evolutionary histories, genomic modification processes (i.e., expansion or contraction), niche shifts, and the age of the

polyploid species are all important factors to predict the macroevolutionary patterns of polyploid species.

Thesis acknowledgement

My four-years' PhD journey (Nov 2018 to Dec 2022) turned out to be an amazing and unique adventure, for which I would like to express my sincerest gratitude to my supervisors Associate Professor Jennifer Tate and Dr Heidi Meudt. Thank you for providing me insightful and patient guidance for my thesis development, being strict about all the fine details, and being super supportive to make many novel ideas come true. In addition to all scientific knowledge, I learned a lot from you on how to carry out field trips, how to write grant proposals, and how to be a team player.

Special thanks to my co-supervisor, Bill Lee, for including me in this project and offering me kind help on plant leaf collection and measurement of genome sizes. Thanks also to co-supervisor Peter Lockhart for sharing ideas about polyploidy. I am also grateful to our collaborators Antoine Nicolas, Gregory Plunkett, and Peter Heenan for sharing valuable data and providing detailed background information for this study. Big thanks to my colleagues, Anne Thomas and Luke Liddell, for providing many insightful suggestions on data analysis. Special thanks also go to Xiaoxiao Lin, Massey Genome Service, for spending an enormous amount of time with me to set up the wet lab protocols.

I would like to thank the research funding from the Royal Society of New Zealand Marsden Fund, the School of Natural Sciences at Massey University, Hellaby Grasslands Trust Grant, Australasian Systematic Botany Society Hansjörg Eichler Scientific Research Fund, Royal Society of New Zealand Hutton Fund Award, and J P Skipworth Scholarship.

Many thanks to all the botanists and volunteers who helped me during my field trips or by sending me plant materials, including Sam, Justin, John, Alex, Cara-Lisa, David, Chris, Demet, Kay, Sophie, India, Jen, Heidi, Bill, Antoine, and Peter. For technical support, I am also thankful to Prashant, Caroline, Sidra, Thilini, Sofie, Trish, Lars, Todd, and Andrew.

Last but not least, I wish to thank to my friends Célia and Margarita for all your support and encouragement. Special thanks to my friend Glen for making Palmy feel like home to me. I could never have accomplished this thesis without support from my parents, especially my mother Xiaoying, who has always been there for me, offering me her unconditional support.

Table of Contents

List of figures

Chapter 1

Chapter 4

List of tables

Chapter 2

Chapter 3

Chapter 1. An introduction to phylogenomic methods for studying polyploid genera

Abstract

Phylogenetic inference of polyploid species can provide insight into their origins and taxonomic relationships. This is the first step towards understanding the diversification of polyploid species. However, in addition to the original whole genome duplication (WGD) event, evolutionary processes such as reticulation (hybridization) and repeated polyploidization can occur, which can further increase the genomic complexity of polyploids and complicate inference of their phylogenetic inference. In this paper, we review the limitations of inferring species relationships of polyploids using traditional phylogenetic sequencing approaches, as well as the mischaracterization of the species trees from single or multiple gene trees. We provide a roadmap to studying phylogenomic approaches for polyploid lineages by comparing and evaluating the application of five current phylogenetic or phylogenomic approaches (e.g., PCR, microfluidic PCR, RAD-Seq, target enrichment and genome-skimming sequencing) and one phylotranscriptomics approach (transcriptome sequencing), using different generations of sequencing platforms. For polyploid species tree reconstruction, we assess the following criteria: 1) the amount of prior information or tools required (e.g., primers, restriction enzymes or biotinylated RNA baits) to capture the region of interest; 2) the probability of recovering homeologs for polyploid species; and 3) the time-efficiency of downstream data analysis. Moreover, we discuss bioinformatic pipelines that can reconstruct networks of polyploid species relationships, which better reflect their evolutionary histories. In summary, although current phylogenomic approaches have much improved our understanding of reticulate species relationships in polyploid-rich genera, the difficulties of recovering reliable orthologous genes and sorting all homeologous copies for allopolyploids remain as a challenge. In the future, long-reads assembled sequence data will further assist the capture of all gene copies (i.e., homeologs), which can be particularly useful to reconstruct the multiple independent origins of polyploids. Furthermore, combining the species network with chromosome number and genome size data will offer additional insight into the WGD or post events -WGD chromosome modification among polyploid lineages.

Keywords: Genome-skimming; Hyb-Seq; PCR; microfluidic PCR; Phylogenomics; Polyploidy; Plants; Single copy genes; Sequencing; New Zealand.

1.1 Polyploidy and Species Diversification

Polyploidy results from whole genome duplication (WGD) and describes organisms that have more than two paired sets of chromosomes. Although the long-term evolutionary effects of WGD remain debatable (Soltis et al., 2014; Mayrose et al., 2015), that is, whether polyploidy is an evolutionary dead-end (e.g., polyploids have higher extinction rates) or ultimately a mechanism to promote speciation, genetic evidence suggests that all flowering plants have had at least once WGD event in their evolutionary history (Jiao et al., 2011a). Especially for plant lineages with many young polyploids (neopolyploids), WGD as a source of variation is particularly important for promoting species diversity (Rothfels, 2021; Clark and Donoghue, 2018; Wood et al., 2009). Therefore, to investigate plant evolution and diversification due to WGD, first and foremost the origins and relationships of polyploid species must be reconstructed (Rothfels, 2021; Oxelman et al., 2017).

Polyploids are broadly defined by their progenitor genome contributions. An allopolyploid forms from the combination of genomes from different species, while autopolyploids form by the duplication of a single species (Comai, 2005; Glover et al., 2016). For example, Fig. 1A shows allopolyploid genomes that contain multiple homeologous (partially homologous) copies of parental chromosomes, while autopolyploids contain multiple homologous sets. These two groups of defined polyploids occupy the extreme ends of a spectrum of formations, but it is important to realize that there are forms in between that can be difficult to categorize into these strict types. Sometimes, meiotic chromosome pairing behaviour can be used to help elucidate the type of polyploid, whether bivalent (allopolyploids) or multivalent (autopolyploids) pairing occurs (Glover et al., 2016). In any case, duplicated genomes in both auto- and allopolyploids can be triggers to initialize the genomic evolutionary, and species diversification of polyploid lineages (Michael et al., 2016; Soltis et al., 2015).

From a genomic point of view, polyploids tend to be more diverse than diploids, and this diversity is related to the differences in their formation and the diversity of their parents (i.e., autopolyploids and allopolyploids) (Fig. 1A). In general, allopolyploids can have more genetic incompatibility compared to autopolyploids due to the divergence between the parental genomes (i.e., the differences between duplicated homeologs vs duplicated homologs), which brings more challenges for the establishment of allopolyploids than autopolyploids, i.e., balancing coexisting divergent subgenomes (Edger et al., 2018). On the other hand, the divergent subgenomes in allopolyploids can provide more opportunities for novel traits to evolve, which often makes

allopolyploids easier to be identified than autopolyploids using morphological traits and genetic sequencing data (Qiu et al., 2020; Soltis et al., 2007).

In addition, because the cytoplasmic genomes in plants are uniparentally inherited (maternally in flowering plants) (Birky, 1995). Especially in allopolyploids, this situation can lead to further divergence of the polyploid lineage due to organellar-nuclear genome incompatibility (Sharbrough et al., 2017; Postel and Touzet, 2020). For example, Fig. 1B shows the formation of two types of tetraploids that have the same nuclear genome but different organellar genomes via reciprocal allopolyploidization [e.g., *Tragopogon miscellus* (Shan et al., 2020)].

Even for allopolyploid species with the same subgenome donors and organellar genomes, different post-polyploidization processes can also increase the diversity of a polyploid lineage (Li et al., 2021; Dodsworth et al., 2016). In particular, post-WGD molecular level changes, including chromosome reorganization, genomic/gene deletions or insertions, transcriptomic regulation, and epigenomic modification, can be further affected by various ecological conditions to promote the divergence of polyploids (Li et al., 2021; Vicient and Casacuberta, 2017; Chen, 2007). One example of this is the different post-WGD divergence among three polyploids in Fig. 1C that have a single allopolyploidization origin [e.g., *Nicotiana* section Repandae (Dodsworth et al., 2017; Clarkson et al., 2017; Wang et al., 2022; McCarthy et al., 2016)].

The formation of higher-level polyploids (i.e., more than two subgenomes donors) via repeated WGD (e.g., Novikova et al., 2018), as well as additional reticulation events including hybridization or introgression (Twyford and Ennos, 2012; Soltis and Soltis, 2009) may lead to further species complexity and diversity. This might include gene flow between species with the same ploidy level (Fig. 1C) or between different ploidy levels (Fig. 1D). Moreover, the formation of higher-level polyploids can occur via autopolyploidization of an existing allopolyploid or via additional hybridization events with a progenitor lineage (Fig. 1D), as exemplified by *Fragaria* (Wei et al., 2017), *Cerastium* (Brysting et al., 2011; Brysting et al., 2007), and *Rosa* (Debray et al., 2021).

Given the variable factors that can contribute to the diversity of polyploids, a system with closely related polyploids of variable chromosome numbers (hereafter polyploid lineages or polyploid-rich genera) can be particularly useful to investigate macroevolutionary patterns (e.g., Wang et al., 2021a; Meudt et al., 2015; Moraes et al., 2022; Karbstein et al., 2022a; Brittingham et al., 2018). Due to the initial WGD and additional reticulation events, such genera can have neopolyploid species with comparable and predictable genome sizes (C-value) and ploidy levels when compared to their parental species. Whereas the additive genome size or even the ploidy level in

neopolyploids from parental species can be rapidly altered following each round of post-polyploidization genomic changes that are mediated by different ecological conditions (Wang et al., 2021c; Leitch and Bennett, 2004). Eventually, neopolyploids or mesopolyploids (later generation polyploids) within a genus can have various genome sizes, diverse morphological characteristics, phenologies, life histories, and geographic distributions (e.g., Qiu et al., 2019; Lu et al., 2022; Debray et al., 2021; Han et al., 2020). Therefore, understanding the phylogenetic relationships of species in a polyploid-rich genus that may have gone through multiple rounds of WGD and with reticulate evolutionary histories is both critical and challenging.

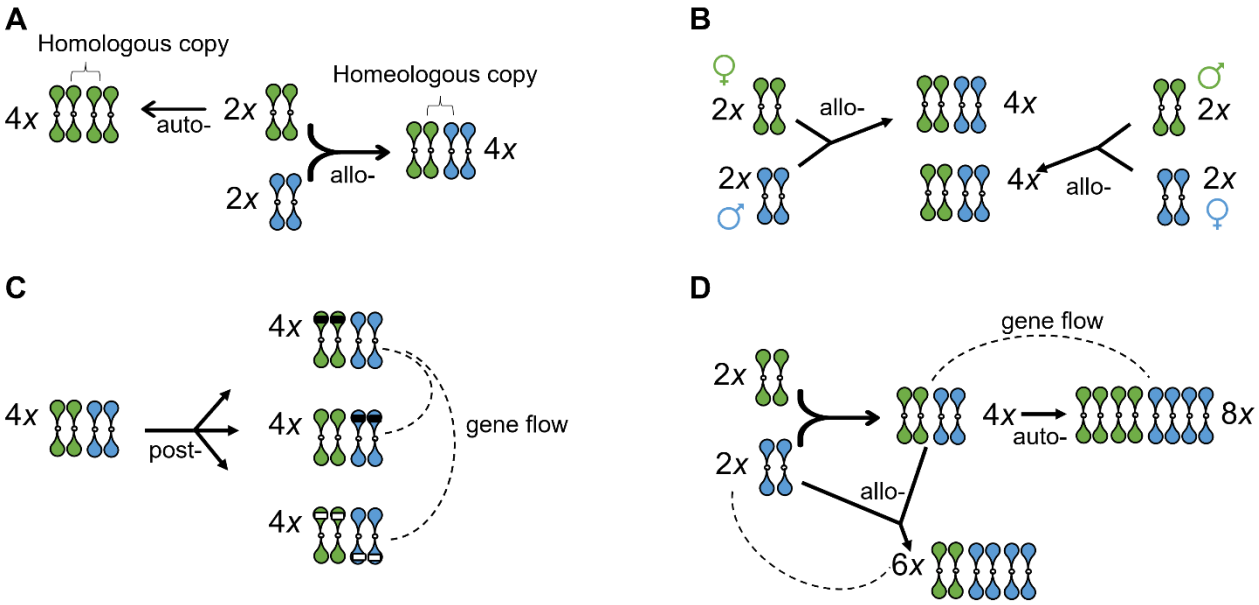


Figure 1 Origins and diversification of polyploid species. The genomes of two diploid ancestors are represented by green (2*x*) and blue (2*x*). **A)** An autotetraploid originates from whole genome duplication (WGD) of a single diploid ancestor vs. an allotetraploid that originates from WGD of two diverged diploid ancestors that brought homeologous chromosomes together in one cell. **B)** Reciprocal allopolyploidization that formed two allotetraploids with different organellar genomes. **C)** Different post-polyploidization processes, such as DNA deletion (white strips on the chromosome) or insertion (black strips on the chromosome) can lead to the diversification of a single allopolyploid lineage. Additional gene flow between tetraploids of different origins (represented by dashed line) generates further diversity among neopolyploid lineages. **D)** Formation of higher polyploids via repeated allo- or autopolyploidization events and with possible gene flow indicated by a dashed line.

1.2 Resolving Polyploid Relationships Using Orthologous Sequences

1.2.1 Orthologous Genes in Polyploids

Comparative genomic analysis of polyploid lineages is an ideal approach to identify the origins of homeologous genomes and to resolve species relationships (Yang et al., 2011). However, for non-model polyploid lineages that contain multiple ploidy levels, the steps required to generate and annotate sub-genomes are, in most cases, time-consuming and not cost-effective. By contrast, molecular phylogenetic inference of polyploid lineages provides a reliable and affordable approach to estimate species relationships and the origins of polyploids (Schmickl et al., 2016; Morales-Briones et al., 2018; Hart et al., 2016; Lozano-Fernandez, 2022).

In general, phylogenetic inference of species relationships can be estimated from the genealogy of a single genetic marker (i.e., a gene tree) or from genealogies derived from genome-wide markers (i.e., phylogenomic inference). In either case, these gene markers are expected to be orthologous sequences that were derived from the same ancestral locus or loci and diverged following speciation, as opposed to paralogous sequences that arise from gene duplication (reviewed by Small et al., 2004). The informative polymorphisms among orthologous genes can be used to infer phylogenetic relationships among species. However, for polyploid lineages with multiple sets of chromosomes (Fig. 1D), identifying and selecting informative genetic regions remains challenging (Rothfels, 2021; Small et al., 2004). This is because identifying orthologous genes that have all homeologs copies preserved after post-WDG can be difficult, as well as capturing and recovering all homeologous sequences, as discussed later.

Moreover, not all gene regions or orthologous markers can be successfully used to resolve phylogenetic relationships, as different regions of a locus or different loci can have their own unique evolutionary rates (i.e., nucleotide substitution rates) and characteristics (i.e., inheritance mode) (Small et al., 2004; Soltis et al., 2014). For example, within a single locus, the coding regions are more conserved across different taxa than non-coding regions, whereas non-coding regions contain more polymorphisms for phylogenetic inference due to faster evolutionary rates (Borsch et al., 2009; Pleines et al., 2009). Therefore, the selection of different genetic markers or even different regions within a locus (e.g., coding vs. non-coding regions) depends on the specific study questions of the plant lineages being investigated.

1.2.2 Phylogeny of High and Single Copy Genes in Plants

Here, taking a polyploid-rich group as an example, we focus on investigating the diversification of polyploid lineages that may have an ancient WGD origin, as well as a recent species radiation that involved both polyploidization and reticulation. Below, we discuss the utility and limitations of different genetic markers (i.e., high copy organellar and nuclear genes vs. single nuclear genes) in phylogenetic reconstruction, especially in identifying the parental species for taxa with hybrid or allopolyploid origins (Fig. 1A; Fig. 1D).

In flowering plants, organellar genes from the chloroplast (plastid DNA, cpDNA) or mitochondria (mtDNA) are often only uniparentally inherited from the maternal lineage (Birky, 1995). They can be particularly useful to identify the maternal lineage of allopolyploids (Fig. 1B) and to resolve complex species relationships especially in conjunction with nuclear genes (Šlenker et al., 2021; de Lima Ferreira et al., 2022; Debray et al., 2021). These genetic markers have high copy numbers present in the cell and are often conserved across taxa, two attributes that make them easy to capture and sequence (Small et al., 2004; McKain et al., 2018). Compared to the mitochondrial genome, which has more structural variation (i.e., high intramolecular recombination) and less sequence variation (i.e., low evolutionary rates from low nucleotide substitution rates), the chloroplast genome lacks recombination and has a faster evolutionary rate, and therefore chloroplast markers have been generally preferred for species-level phylogenetic inference (Small et al., 2004; Ravi et al., 2008). Nevertheless, the phylogeny utilizing organellar markers may have limited abilities to correctly infer polyploid relationships due to WGD and reticulate evolutionary histories. In particular, local introgression events between closely-related polyploid species will lead to a plastid phylogeny that shows local geographical structure instead of species relationships (Tsitrone et al., 2003).

In comparison, biparentally inherited nuclear genes (Small et al., 2004) often have faster evolutionary rates compared to cpDNA or mtDNA (e.g., Huang et al., 2012; Gaut, 1998). Nuclear genes can be more informative about reticulate polyploid species relationships, given the recent formation of neopolyploids that often contain largely intact subgenomes (Li et al., 2021). The orthologous nuclear genes can be further divided into two categories based on their copy number variation, i.e., high copy vs. single copy nuclear genes. Both high copy and single copy nuclear genes are useful to resolve phylogenetic relationships of polyploid lineages (Small et al., 2004). High copy nuclear genes (e.g., the internal transcribed spacers of the nuclear ribosomal DNA; ITS; nrDNA), similar to organellar genes, are also often conserved across taxa and are easy to capture and sequence (Baldwin et al., 1995). However, high copy nuclear genes may be of limited use for

phylogenetic inference because of insufficient polymorphic sites due to recent radiation among species, or the homogenization process (concerted evolution) of homeologous loci that can result in a single copy remaining in the genome (reviewed by Soltis et al., 2014; Álvarez and Wendel, 2003). By contrast, single copy nuclear genes are less likely to be subjected to concerted evolution and tend to have faster evolutionary rates with more informative sites (Small et al., 2004). Therefore, phylogenies reconstructed from single copy nuclear genes can be used to characterize the species relationships more precisely, especially when high copy genes alone, such as plastid markers and nrDNA, cannot provide resolved phylogenies for polyploid-rich genera whose species have gone through rapid species radiation, WGD and reticulation (Sang, 2002; Soltis et al., 2014).

1.3 An Overview of Current Phylogenetic Approaches

Phylogenetic approaches can capitalize on next-generation high-throughput sequencers to capture genome-wide markers, including both whole or partial organellar genomes and numerous biparentally inherited nuclear loci, to infer the phylogeny of polyploid species (McKain et al., 2018; Kapli et al., 2020). There are currently six popular phylogenetic approaches (e.g., including phylogenomics and phylotranscriptomics) that can be divided into four categories based on their prior required investment (e.g., cost and whether reference data are required) and the efficiency of downstream data analysis (Fig. 2) (reviewed by McKain et al., 2018).

Specifically, 1) Designing primer pairs for capturing an individual locus is time-consuming, but downstream analysis is efficient, e.g., PCR and microfluidic PCR; 2) Restriction enzyme based approaches that can reduce genomic complexity (usually little prior work needed to design restriction enzyme) for downstream analysis, e.g., restriction site-associated DNA sequencing (RAD-Seq); 3) Target enrichment sequencing (Hyb-Seq) that can capture specific genomic regions simultaneously using predesigned biotinylated RNA baits, including complex taxon-specific (requires prior selection of single copy genes using transcriptome or genome-skimming data) or simplified universal (no prior information required) bait set (Johnson et al., 2018), with reduced downstream computational analysis; 4) No prior information is required but downstream computational methods are demanding, e.g., genome-skimming sequencing with depth or transcriptome sequencing (i.e., phylotranscriptomics).

Selecting the most efficient sequencing platform for each phylogenetic approach plays an important role in recovery of informative orthologous gene sequences and their duplicated homeologous gene copies, which is essential for phylogenetic inference of polyploid species. Below,

we discuss the application of three generations of sequencing platforms combined with six approaches by comparing their efficiency on sequencing genome-wide markers in diploid vs. allopolyploids (4*x* and 6*x*), and the recovery of homeologs in polyploids (Fig. 2). These platforms include first-generation Sanger sequencing (e.g., ABI Applied Biosystems™ sequencer), second-generation high-throughput sequencing of short paired-end reads (e.g., Illumina® sequencer), and third-generation single molecule real-time sequencing of long reads (e.g., Pacific® Biosciences or Oxford Nanopore® sequencer).

1.3.1 PCR and Microfluidic PCR

PCR amplification requires a pair of primers that can bind to the targeted region. For targeted regions (e.g., a single copy nuclear gene) without any prior information in the studied group (e.g., a reference genome based on genome-skimming or transcriptomic sequences), designing and optimizing primers from a related genus that has published reference sequences can be time-consuming. By contrast, commonly used regions, such as plastid DNA or the nuclear ribosomal ITS region, which are conserved between taxa and present in high copy number in the genome, are easy to amplify via universal primers using PCR and can be informative about the relationships of diploid species without reticulate evolutionary histories (Hillis and Dixon, 1991; Shaw et al., 2014). For polyploid lineages, PCR-amplified biparentally inherited nuclear genes, such as the ITS homeologs that have not been affected by concerted evolution (i.e., concerted evolution has not occurred to the extant as to render homeologous copies uniform) or single copy nuclear gene homeologs, can be particularly useful to infer the reticulate species relationships within polyploid-rich genera (Rothfels et al., 2017; Osuna-Mascaró et al., 2022; Xu et al., 2017).

Separating PCR-amplified homeologous gene copies can be challenging, especially when combining PCR with traditional Sanger sequencing, which has a sequencing length limit for each targeted gene (< 800 bp) and may have heterozygous sites (as evidenced by multiple electropherogram peaks) from different homeologs (Kircher and Kelso, 2010). Sanger sequencing often requires additional cloning or homeolog-specific primers to separate the homeologous gene copies (Brysting et al., 2011). By contrast, combining PCR with second- or third-generation sequencers, the sequenced PCR amplicons can overcome the limitations using Sanger sequencing. For example, Rothfels et al. (2017) sequenced PCR amplicons (1 Kbp) of four single copy nuclear genes using PacBio, and by adding ambiguous sites among primers, they captured multiple phased (separated) individual homeolog sequences of the amplified genes in the resulting long reads.

Microfluidic PCR enhances the ability of parallel amplification of multiple loci by combining PCR with microarray technology, which allows one to amplify up to 4,608 PCRs per array (192.24 array Biomark™ HD) (Zhang and Ozdemir, 2009; Oshiki et al., 2018). The combination of microfluidic PCR with second-generation Illumina sequencers is a powerful tool that can capture and sequence hundreds of single copy nuclear genes, with additional homeologous sequences recovered from the sequenced reads (Uribe-Convers et al., 2016; Frost et al., 2021; Debray et al., 2021). Utilizing unique barcodes and pair-end sequence information, the individual locus sequences can be demultiplexed and assembled from the sequenced reads for each input taxa, via bioinformatic tools such as Pipeline for Untangling Reticulate Complexes (PURC) (Rothfels et al., 2017) or Fluidigm2PURC (Blischak et al., 2018). After filtering the chimeric sequences from PCR amplifications, the haplotype sequence across all input taxa at each locus can be clustered based on their sequence similarities to continue the downstream gene alignments (e.g., S4 in Fig. 2). However, Illumina sequencers can maximally return up to 300 bp for a single end read, therefore, the targeted gene length is often required to be shorter than 1 Kbp to recover the whole gene sequence without any gaps present.

Although PCR-based methods can efficiently capture the targeted loci, additional effort to test if the orthologous genes are conserved between all taxa can be time-consuming. Moreover, selecting the amplified loci that have all homeologous copies presented, designing the individual primer pairs for each locus and optimizing PCR conditions for each gene amplification will take additional time. Especially for non-model lineages, generating additional reference genomes or transcriptome sequences will be required before targeted loci selection.

1.3.2 RAD-Seq

RAD-Seq is often used for population genetic studies and its use has become more popular in phylogenomic studies (reviewed by Leaché and Oaks, 2017). This approach uses restriction enzymes that can recognize specific genome sites (e.g., 4-bp cutter, 6-bp cutter or 8-bp cutter) to shear DNA into simplified yet comparative DNA fragments, which contain informative sites for quantitative genetic and population genetic studies of individuals from different populations (reviewed by Davey et al., 2011). Similarly, using enzyme-digested DNA fragments between closely related taxa can also be used to compare their genetic divergence or reconstruct their phylogenetic relationships (Fig. 2) (Karbstein et al., 2022a; Wang et al., 2021b).

Andrews et al. (2016) reviewed RAD-Seq related approaches, such as double-digest RAD (ddRAD) and genotyping-by-sequencing (GBS), which often rely on single polymorphic sites

(SNPs) extracted from Illumina sequenced reads (maximum single-end 300 bp length) for downstream analysis. After acquiring the RAD-Seq reads data, the next step is to identify the orthologous DNA fragments based on their sequence similarities (S4 in Fig. 2). Bioinformatic tools such as ipyrad (Eaton and Overcast, 2020) can demultiplex the sequence reads into each individual using sequencing barcode information, then the similar reads can be sorted into each DNA fragment cluster (loci) with or without a reference genome. After assembling each DNA cluster block with overlapped reads into consensus sequences and cross-comparing all input taxa, eventually, a VCF file that contains all SNPs for each individual can be generated for downstream analysis.

Allopolyploids that contain diverged subgenome donors are also expected to have the more heterozygous SNP sites. Assigning or phasing the genome-wide SNPs of polyploids to each subgenome donor using RAD-Seq DNA fragments is not feasible without the reference genome of each subgenome donor, unless the RAD-Seq data include both the digested DNA fragments of polyploid species as well as their potential diploid subgenome donors (Wang et al., 2021b). On the other hand, inferring polyploid species phylogeny using RAD-Seq data often has lower requirement for SNPs phasing, but is often limited by the necessity of having diploid ancestor species (Karbstein et al., 2022a; Wang et al., 2021b) as discussed later.

Commercially available restriction enzymes can cut targeted DNA in specific places and to adjustable DNA fragment lengths, which can reduce the requirement of prior design. In addition, the RAD-Seq method often requires a low degree of divergence (or a low substitution rate) among the investigated taxa, but can also work for diverged taxa (Guo et al., 2023). However, the efficiency of RAD-Seq approach for study diverged taxa may depend on the proportion of missing SNPs since less common DNA cut sites between taxa would be expected (Eaton et al., 2017). On the other hand, the random distribution of the enzymatic sites means that no specific gene sequence (e.g., single copy genes) can be recovered (McKain et al., 2018). More importantly, RAD-Seq is also more sensitive to samples with highly degraded DNA (e.g., herbarium specimens) (Graham, 2008). Further, RAD-Seq results may be difficult to reproduce over time and among laboratories, which would usually prevent data being reused or repurposed with new results added from additional samples.

1.3.3 Target Enrichment Sequencing

Target enrichment sequencing or Hyb-Seq provides a straightforward way to capture desired genomic regions via predesigned biotinylated RNA baits (Weitemier et al., 2014; Andermann et al., 2020). These baits are often at least 100 bp in length and can capture specific genomic regions

based on the sequence similarities (Fig. 2). The bait set captured DNA fragments from the targeted genes can be enriched via PCR and sequenced by high throughput sequencing platforms (Weitemier et al., 2014). For phylogenetic inference, the exons of targeted genes are more conserved than introns or flanking regions across taxa, therefore, they are ideal regions for bait design (Schmickl et al., 2016; Weitemier et al., 2014).

Using a pipeline such as HybPiper (Johnson et al., 2016), the single copy genes can be extracted by mapping the sequenced captured-reads to a reference file which contains the reference sequences of the desired loci (e.g., McLay et al., 2021), and the mapping step can be done using the DNA nucleotide sequences or with the transcribed amino acid sequences. Each reads cluster can then be *de novo* assembled into contigs and assigned to correct orders by comparing to the exon sequences in the reference file, for final extraction of exons or whole gene sequences (including exon, introns and flanking regions). Furthermore, the limited number of targeted loci increases the opportunity to pool many individuals in one batch for mostly the second-generation sequencing. The read depth can also provide additional opportunity for each homeologous gene copy extraction. For example, below are a few bioinformatic tools that have been developed to detect and extract allelic variation from Hyb-Seq data.

ParalogWizard (Ufimov et al., 2022) or putative paralogs detection (PPD) pipeline (Zhou et al., 2022) can increase the performance of 'paralogs' (or different homeologous gene copies for allopolyploids) detection by comparing the divergence level among the assembled gene sequences. Whereas HybPhaser (Nauheimer et al., 2021) can phase the Hyb-Seq data of hybrid or polyploid species by splitting and mapping the sequenced reads via BSplit (BBMap; B. Bushnell, https://sourceforge.net/projects/bbmap/) to the predefined 'subgenome donors'. The method combines the Hyb-Seq data inferred species phylogeny (e.g., an ASTRAL tree; discussed later), to first identify the divergent clades that contain different potential subgenome donors. Then the representative species with low SNPs percentages (e.g., a 2*x* that should not contain large amount of SNPs) can be selected from each associated clade as references, to continue the Hyb-Seq reads phasing for polyploids. By contrast, PATÉ (Tiley et al., 2021) utilizes the ploidy level and Hyb-Seq sequencing depth (i.e., overlapped reads) information of polyploid species to phase the haplotype blocks for each targeted gene (e.g., Crowl et al., 2022) via GATK (DePristo et al., 2011) and HpoPG (Xie et al., 2016). Similarly, Hyb-Seq allele phasing pipeline (e.g., https://github.com/mossmatters/phyloscripts/tree/master/alleles_workflow) that combines GATK and WhatsHap (Patterson et al., 2015) can also extract the haplotypes of each targeted gene using the depth of the sequenced reads (e.g., Šlenker et al., 2021). SORTER pipeline (Jonas et al., 2023),

on the other hand, also increased the ability for filtering the paralogous gene sequences and phasing the Hyb-Seq reads via SAMtools (Li et al., 2009).

However, for polyploid lineages, the pipelines are still limited to extracting the complete homeologous sequences of the targeted single copy nuclear genes. Challenges remain due to lack of ploidy level information or the phasing programme (such as GATK) cannot properly handle polyploids with multiple subgenome donors. Phasing can be additionally challenging when polyploids have subgenomes with low divergence rate or when the universal bait set (e.g., Johnson et al., 2018) are applied, which will often yield similar exon sequences and lower recovery rate of introns and flanking regions (i.e., discontinues exons) compared to the lineage-specific bait sets (Yardeni et al., 2022; Hendriks et al., 2021). Eventually, the targeted loci may result in chimeric gene sequences that contain exons from different homeologous copies.

On the other hand, many published taxon-specific bait sets (at family level) have reduced the time required for prior baits design using reference sequences and increased the opportunities to extract the allelic variation from the Hyb-Seq reads due to higher rates of recovering genes with introns and flanking regions (Šlenker et al., 2021; Crowl et al., 2022). Furthermore, Johnson et al. (2018) identified 353 conserved single copy nuclear genes among angiosperms and published the Angiosperms353 baits set for phylogenomic inference of any flowering plant group. Breinholt et al. (2021) further published GoFlag451 baits set that can capture up to 248 single copy nuclear genes across flagellate plant lineages, including bryophytes, ferns, and all gymnosperms. These universal bait sets improved the flexibility of phylogenetic inference of the investigated group regardless of their divergence rate, such as the Tree of Life project (Baker et al., 2022). The off-target reads can also be used for high copy gene extraction, which provides additional phylogenetic signal for the inference of species relationships (de Lima Ferreira et al., 2022; Karimi et al., 2020). Moreover, the pipeline such as HybSeq-SNP-Extraction (Slimp et al., 2021) can extract SNPs among Hyb-Seq data and the output can be analysed using similar methods as RAD-Seq.

1.3.4 Genome-Skimming by Shotgun Sequencing

Genome-skimming by shotgun sequencing can generate sequences that represent the whole genome of the sequenced individual with shallow coverage of sequencing reads, and often uses second-generation (short reads) sequencers (Fig. 2). This method requires no prior primer designing and is flexible with respect to the divergence rate of the studied groups. It also provides an opportunity to recover the homeologs of extracted loci for polyploid lineages.

Depending on the sequencing coverage, shallow genome-skimming can be an efficient approach to recover high copy number organellar and nuclear genes, such as nrDNA and whole plastid genomes. High copy number genes can be *de novo* assembled from genome-skimming reads due to their abundance in the genome even with low sequencing coverage (e.g., 1x of whole genome sequencing coverage) (Straub et al., 2012). Taking the assembly of a plastome as an example, the seed-and-extend based *de novo* assembly (e.g., assembler compared in Freudenthal et al., 2020) or the reference-mapping based haplotype calling methods (e.g., Takamatsu et al., 2018) via GTAK HaplotypeCaller (DePristo et al., 2011) or SAMtools mpileup (Li et al., 2009) both can produce a complete plastome. For repetitively occurring nrDNA, this may also be treated as a 'circular' genome like the plastome to be *de novo* assembled with different initial seeds, e.g., using the assembler GetOrganelle (Jin et al., 2020). However, phasing the nrDNA allele or homeolog variation from the depth of sequence reads would require additional ploidy level information or the number of subgenome donors of the input taxa for bioinformatic tools like GTAK HaplotypeCaller.

On the other hand, genome-skimming with depth can be also useful to extract SNPs or single copy nuclear genes. For example, SNPs among genome-skimming reads between low-diverged input taxa can be extracted using the reference-mapping based method via joint GVCF calling in GATK (DePristo et al., 2011), and the SNP data can be analysed in sliding-window based method or used to infer the species phylogeny based on independent SNPs, similar to a RAD-Seq output. Similar to the assembly of single copy nuclear genes using Hyb-Seq data, user-friendly bioinformatic tools (methods compared in Michel et al., 2022) , e.g., HybPiper (Johnson et al., 2016; Jackson et al., 2021) and HybPhyloMaker (Fér and Schmickl, 2018), can also *de novo* assemble the deep genome-skimming reads into exons by mapping to a reference file that contains the exons of a related species, and can correct the order of exons within a locus by comparing to the selected reference gene. Moreover, the reference file can be custom-designed using published genome or transcriptome data via bioinformatic tools such as MarkerMiner (Chamala et al., 2015) or a universal angiosperm reference file that contains the reference gene sequences for Angiosperms353 loci (McLay et al., 2021).

Although genome-skimming with sufficient depth is an efficient way to recover genome-wide single copy orthologous genes and possiblly all homeologous sequences for nuclear genes that are biparentally inherited, determining the proper coverage for lineages with multiple ploidal levels is challenging, given that each homeologous copy will require sufficient sequencing coverage to be recovered. Liu et al. (2021) compared the number of successfully recovered single copy nuclear genes and the depth of genome-skimming reads (2x to 20x coverage) in Vitaceae. Their result

suggested at least 10x coverage was required for extracting over 800 single copy nuclear genes for phylogenomic inference. Moreover, the extensive genomic data (e.g., the high number of reads for sequencing coverage and depth) required for each sequenced sample also limits the number of input samples that can be added to one sequencing pool, which increases the total cost of sequencing. In addition, the downstream filtering and extracting of the informative loci from massive amounts of genome-skimming sequenced reads also increases computational analysis time.

1.3.5 Transcriptome Sequencing

Transcriptome sequencing or mRNA-Seq can also be used to capture genome-wide markers (including high copy and single copy genes) without the need for any prior primers or restriction enzyme designing steps (Cheon et al., 2020). In comparison to the phylogenomic approaches, the phylotranscriptomic approach compares orthologous variation extracted from mRNA instead of DNA sequences (Fig. 2). Moreover, annotated reference transcriptomes are increasingly being added to online databases (Leebens-Mack et al., 2019), and the pipelines for identifying orthologs using mRNA are continuously improving (Cheon et al., 2020). When specific experimental design targeting specific tissues or collection times is used, the resulting mRNA comparisons can further address the functional divergence of the loci under investigation (McKain et al., 2018).

After acquiring the mRNA-Seq data and cleaning the reads data (removing the adaptors, poly-N sequences and low quality bases), using bioinformatic tools, e.g., CD-HIT (Li and Godzik, 2006), the sequence reads can be distributed into each protein-coding gene cluster for each sequenced individual. Identifying the orthologous genes across all input taxa can be done in several ways (e.g., compared in Cheon et al., 2020). For example, HaMStR (Ebersberger et al., 2009) uses the 'core-ortholog' (i.e., pre-defined set of orthologs) to identify the ortholog clusters, the gene-tree-based homology searching method in Yang and Smith (2014), or the ortholog-group and gene-tree based approach in OrthoFinder (Emms and Kelly, 2015), etc.

Extracting the haplotype isoforms using short-reads Illumina sequencing can be particularly challenging without reference genomes, because the discontinuous mRNA sequences are from the exons only. Although the haplotypes of gene isoforms can also be phased using third-generation long-read sequencers to determine the origins of different subgenome donors (e.g., Leung et al., 2021; Cerca et al., 2022), mRNA data are often affected by post-polyploidization genomic and transcriptomic modifications (e.g., tissue-specific or allele-specific expression, gene loss, gene silencing, pseudo/sub-functionalization, and mRNA alternative splicing, etc.) (Soltis et al., 2014; Zhou et al., 2011), which can result in biased sequencing depth of homeologs or only one gene copy

retained. In addition, mRNA preservation from the collected samples requires more critical conditions, e.g., freshly collected samples need to be sorted under -80 degrees Celsius. The extraction of mRNA and preparation of transcriptome libraries are both more complex and costly than extracting DNA and preparing DNA genomic libraries (McKain et al., 2018), which makes phylotranscriptomics a less attractive option compared to phylogenomics.

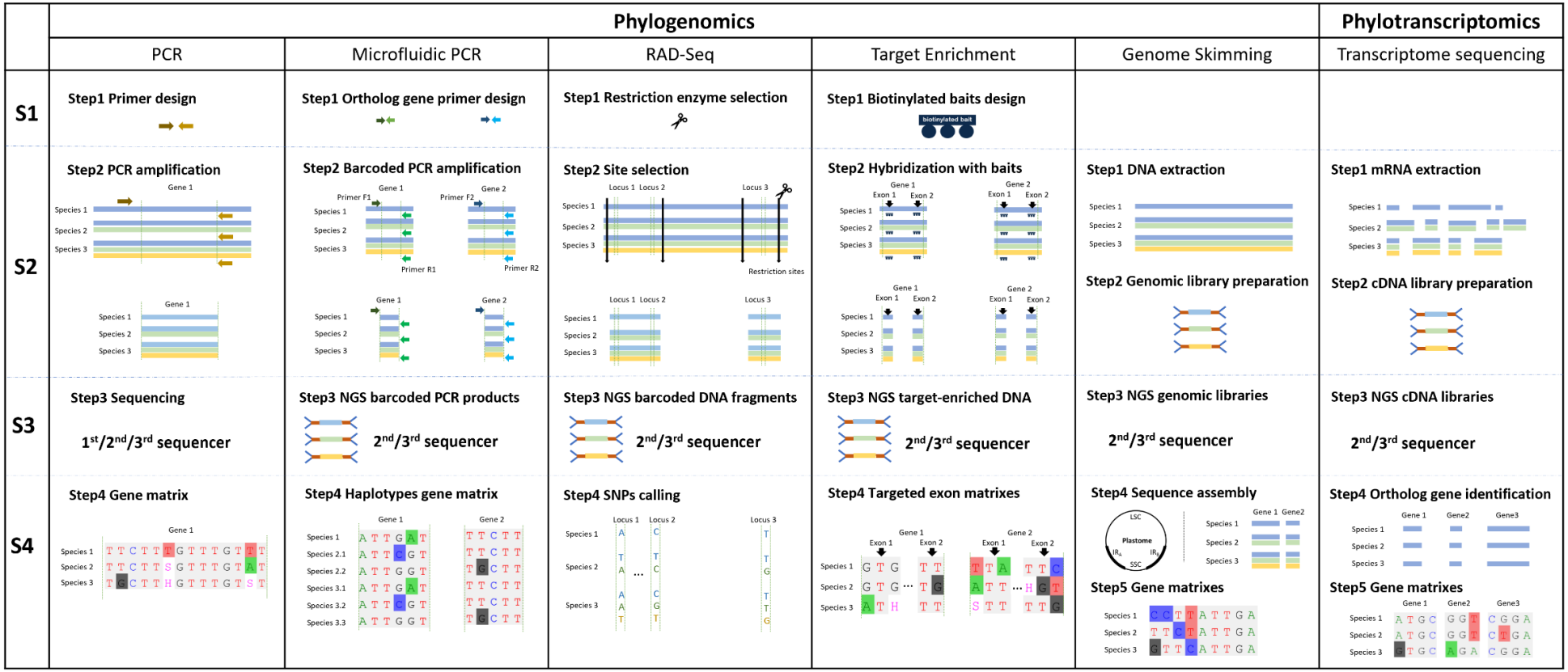


Figure 2 A comparison of the workflow for different phylogenomic approaches (from top panel left to right: PCR, microfluidic PCR, RAD-Seq, target enrichment, and genome-skimming sequencing) and phylotranscriptomics (transcriptome sequencing) for capturing targeted loci in four aspects (left column: S1 to S4): step 1 (S1) prior preparation, step 2 (S2) laboratory workflow, step 3 (S3) available sequencing platforms (e.g., first-generation Sanger sequencer, and the high-throughput of second- and third-generation sequencers), and step 4 (S4) sequencing outcomes and bioinformatic analysis. Taking the haploid genomes of diploid species 1, allotetraploid species 2, and allohexaploid species 3 as an example, **1)** PCR can capture and amplify the targeted locus (e.g., the sequence of Gene 1 indicated between two vertical dashed lines) with a pair of prior designed primers (a pair of arrows in S1). The PCR product can be sequenced using all three generation sequencers (S3). Using Sanger sequencing, the output (S4) will contain multiple heterozygous sites in the chromatogram representing different homeologs in a polyploid, whereas high-throughput sequencers will generate

separate homeologous sequences of the targeted gene, similar to microfluidic PCR. **2)** Microfluidic PCR also starts with the primer design (different colour of arrow pairs in S1) for each targeted locus (Gene 1 and 2 in between dashed lines in S2). The amplicons (S2) can be individually barcoded for each sample. Then the amplicons from different samples in one array can be multiplexed and sequenced using the second- or third-generation sequencers (S3). The sequenced result (S4) will contain variation derived from the parental genomes, e.g., the allotetraploid species 2 has two homeologous gene copies indicated by species 2.1 and species 2.2 and allohexaploid species 3 has three gene copies (3.1, 3.2, and 3.3). **3)** RAD-Seq utilizes restriction enzymes (indicated by scissors in S1) that can recognize specific restriction sites (black arrows in S2) and shred the genome into random, simplified, and comparative DNA fragments that contain informative DNA sites for sequencing (S3). The sequenced reads are often used to extract the SNP variation (S4). **4)** Target enrichment sequencing or Hyb-Seq uses predesigned RNA biotinylated baits (S1) to hybridize with individually barcoded genomic libraries in one pool (S2). Taking Angiosperms353 (Johnson et al., 2018) as an example (S2), the baits can combine with the conserved exons (e.g., Exon 1) of targeted genes (e.g., Gene 1). The post-capture enriched libraries can be sequenced on a high-throughput sequencer to recover the exons (S3; S4) that contain heterozygous sites that may not be able to be phased. **5)** Genome-skimming requires no prior information (S1), and the genomic libraries can be sequenced using a second or third-generation sequencer (S2; S3). The output of genome skimming depends on the sequencing depth. For example, shallow genome-skimming sequencing reads (e.g., the whole genome on average with only 1x coverage) can be used to extract high copy number genes or genomic regions (left in S4), such as the plastome or nrDNA. By contrast, deep genome-skimming sequencing reads (e.g., whole genome on average have more than 10x) can be used to extract high copy number genes, as well as haplotypes of single copy nuclear genes (right in S4). The final gene matrix for maternal-only inherited plastid DNA will only have one copy extracted (S5), whereas biparentally inherited genes can result in a haplotype gene matrix that is similar to microfluidic PCR (S5). **6)** Transcriptome sequencing also does not require any prior information (S1) and can be sequenced using the second- or third-generation sequencers (S2; S3). However, due to the sequencing reads being only from the exons and post-transcriptome modification processes (details see the main text),the identification of orthologs using mRNA data often cannot properly identify the genes with homeologous copies (S4). This figure was adapted from McKain et al., (2018).

1.4 Phylogenetic Inference of Species Relationships

1.4.1 Phylogenetic Inference of an Individual Locus (Gene trees)

After assembling the targeted loci (e.g., SNPs, cpDNA, nrDNA or single copy nuclear genes) from the sequenced reads data, the next step is to reconstruct the phylogeny for each individual locus. There are four main methods for phylogenetic inference of an individual gene tree, namely neighbour-joining (NJ), maximum parsimony (MP), maximum likelihood (ML), and Bayesian Inference (BI) (reviewed by Holder and Lewis, 2003; Kapli et al., 2020; Yang and Rannala, 2012).

An allopolyploid species is expected to have only one copy at uniparentally inherited loci (i.e., plastid) and multiple homeologous gene copies or heterozygous SNPs at each biparentally inherited nuclear locus (Fig. 2). In this way, homeologs would be expected to have greater sequence-level differences than homologs. For sequencing methods that can produce individual gene sequences, including PCR (with homeolog-specific primers or next-generation sequencers), microfluidic-PCR, Hyb-Seq, genome-skimming with depth and mRNA-Seq, aligning the gene sequences is the first step to identify the homology of a sequenced locus prior to phylogenetic tree construction by any of the four methods mentioned above (bioinformatic tools reviewed in Kapli et al., 2020; Guo et al., 2023). Taking the cytoplasmic marker as an example (Fig. 3C), a traditional bifurcating gene tree can be used to infer only the maternal lineage of a polyploid species, given that these markers often only have one copy present, with exceptions such as biparentally inherited plastid DNA (e.g., Barnard-Kubow et al., 2017). By contrast, a multi-labelled bifurcating gene tree (MUL-tree) can show multiple origins of divergent homeologous gene copies of a biparentally inherited locus (Fig. 3B) (Czabarka et al., 2013; Huber et al., 2008). A MUL-tree can be informative about reticulate species relationships, when the terminal branches come from the same individual that can be further merged together via Dendroscope (Huson and Scornavacca, 2012) to visualize network relationships of polyploids (Fig. 3D) (Morrison, 2014; Hibbins and Hahn, 2022).

By contrast, SNP data of known allopolyploids can be mapped back to the potential subgenome donors (e.g., the putative 2*x* ancestors) and the proportion of mapped reads may also indicate the origins of the species (e.g., Wang et al., 2021b). Long, shredded DNA fragments (e.g., assembled contigs larger than 200 bp) that contain SNPs can also be treated as an individual 'locus' for gene phylogeny inference (e.g., Wang et al., 2021b). Moreover, SNPs from the same DNA fragments can be phased and analysed together to infer species relationships (e.g., Karbstein et al., 2022a) via RADpainter and fineRADstructure (Malinsky et al., 2018b). On the other hand, the individual biallelic SNP (i.e., SNPs pruned in linkage disequilibrium) can often be analysed as each individual locus to reconstruct the 'gene trees' via quartet-inference of SVDquartets (Chifman and Kubatko,

2014) or BI based SNAPP (Bryant et al., 2012) under the coalescent model. Eventually, a consensus tree can be generated by considering the concordance levels among all individual SNP trees.

However, not all homeologous copies or DNA fragments of a targeted locus can be successfully recovered, because the post-polyploidization genomic modification process will also lead to duplicated genes with different fates (Fig. 1C). For example, the duplicated gene copies may be either retained or lost, or subfunctionalization or neofunctionalization of duplicated genes can occur (reviewed by Prince and Pickett, 2002; Comai, 2005). Even if all gene copies are retained at one locus and no recombination between homeologs has occurred, phasing the homeologs using short-reads can still be challenging, given the limited bioinformatic tools that cannot handle polyploids with multiple subgenome donors and the large amount of missing ploidy level information for non-model groups. Finally, the phylogeny of polyploid species may result in the incomplete gene tree inference with only one copy retrieved from one of the subgenome donors (Fig. 3A vs Fig. 3B). In addition, for all six phylogenomics or phylotranscriptomic methods, the misidentification of the orthologous gene copy or DNA fragment can happen due to the short divergence time between parental genomes or gene loss during post-WGD genomic modification (Unruh et al., 2018). Furthermore, Hyb-Seq and mRNA-Seq, which are most specific to recover only conserved exons (Weitemier et al., 2014; Cheon et al., 2020; Zhou et al., 2022), can be additionally problematic to extract the correct orthologs and their homeologous copies using the short-reads, because of the chance of getting chimeric concatenated-exons from different homeologs is possible, which may result the incorrect gene tree inference (Fig. 3E vs Fig. 3B).

Therefore, the species phylogenetic relationships can rarely be estimated from a single locus. Most studies now use a combination of genetic markers from different genomes (e.g., organellar vs. nuclear), consider the evolutionary background of each type of marker, and reconcile genome-wide signals to understand the origins and relationships of polyploids to improve their phylogeny reconstruction (Holder and Lewis, 2003; Soltis et al., 2014).

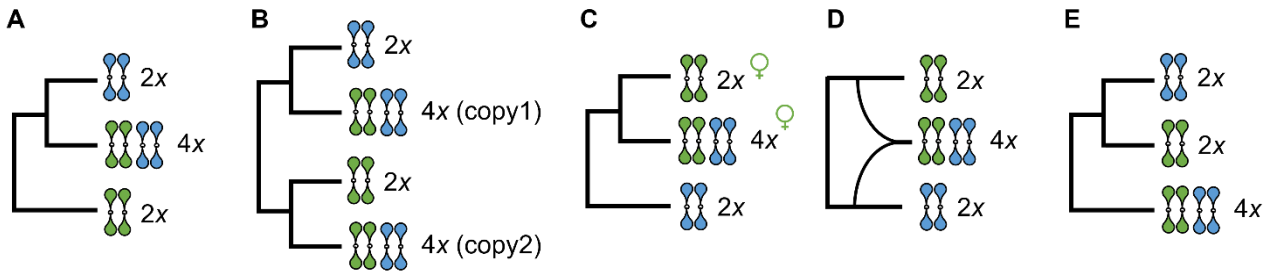


Figure 3 Phylogenetic inference of an allopolyploid species using different tree reconstruction methods. **A)** A traditional bifurcating phylogenetic tree based on a nuclear marker reconstructs a polyploid as sister to one parent. The other parent may be observed to be more phylogenetically

distant, depending on if gene loss has occurred, divergence between the two parents, and overall sampling in the tree. **B)** A multi-labelled nuclear gene tree shows two homeologous copies in an allotetraploid, each derived from one diploid parent. **C)** A bifurcating phylogenetic tree based on an organellar marker reflects the maternal progenitor of an allopolyploid. As with the nuclear-based bifurcating tree, the other parent may be more distantly related based on divergence between the two parents and overall sampling. **D)** A network based on a nuclear marker shows both parents that contributed to the allopolyploid genome. **E)** A bifurcating gene tree inferred from chimeric assembled gene sequence of a polyploid species or possible recombination between homeologs.

1.4.2 Inferring Species Tree from Gene Trees

In addition to post-WGD genomic modification processes, discordance between any one single gene tree (i.e., phylogenetic incongruence) may be caused by the independent evolutionary histories, i.e., origins or evolutionary rates of selected gene or genetic region as mentioned above, or additional biological factors (e.g., incomplete lineage sorting, reticulation, horizontal gene transfer and polyploidization, etc.) can also contribute to the gene trees being discordant with the underlying species phylogeny (reviewed by Maddison, 1997; Degnan and Rosenberg, 2009; Twyford and Ennos, 2012; Mallet et al., 2016). Moreover, the concordance level of genome-wide captured nuclear loci can be further used to resolve the reticulate relationships of species in polyploid-rich genera as discussed below (Solís-Lemus et al., 2017; Than et al., 2008).

Maddison and Knowles (2006) and Yan et al. (2022) showed that a reliable species phylogeny can be generated from a sufficient number of input nuclear loci, and there are two typical evolutionary models applied when inferring the species phylogeny from the list of loci. First, the species tree can be estimated by joining all the markers together (concatenation model), such that all markers are considered to have the same evolutionary history. The concatenation method may be more robust for species phylogeny estimation, but it can also result in overconfident node support values (Kubatko and Degnan, 2007). By contrast, each gene can be analysed individually (multispecies-coalescent model, MSC), and under the coalescent model, the list of gene trees can be used to infer the species tree (Degnan and Rosenberg, 2009). The MSC model assumes the gene trees are independently evolving, and it applies the coalescent-based theory to estimate the coalescence time (Heled and Drummond, 2010). This model also can be more consistent in identifying ILS, and therefore provide a more accurate estimation of species trees (Mirarab et al., 2014).

For polyploid lineages, summarizing gene trees for species phylogenetic inference should be considered in two aspects: i) if all gene trees of biparentally inherited loci have all homeologous copies from all subgenome donors (i.e., MUL-tree in Fig. 3B), or ii) if each gene tree only contains one set of orthologous gene copy from only one of subgenome donors (e.g., plastid ortholog gene

tree in Fig. 3C). Realistic data for studying polyploid-rich lineages mostly still rely on utilizing orthologs extracted from one subgenome donor, for example, recovery of orthologous genes using Hyb-Seq (Karbstein et al., 2022a; Thomas et al., 2021), mRNA-Seq (Morales-Briones et al., 2021; Zhang et al., 2021), or deep genome skimming (Liu et al., 2021). The gene trees can be summarised into a species tree using a concatenation model such as IQTree (Minh et al., 2020b) or a MSC model such as ASTRAL or StarBEAST2 (Ogilvie et al., 2017). Inferring the species phylogeny using MUL-trees across genome-wide biparentally inherited loci requires proper handling of each gene copy. Although a robust method such as Bayesian-based ASTRAL-Pro (Zhang et al., 2020) under the MSC model allows for multiple alleles of one individual to be present, which may improve the final species tree inference compared to the similar program ASTRAL (Mirarab et al., 2014) that only allows for single copy present in one individual, it nevertheless assumes bifurcating species relationships.

1.4.3 Networks in Polyploid Species

Under the assumptions that all recovered gene sequences or SNPs are all orthologous markers, the conflict between the topologies of gene or SNP trees can be used to infer the allopolyploidization or hybridization histories between polyploids, which can be calculated via gene concordance analysis (Smith et al., 2015; Bouckaert, 2010) or network inference (Solís-Lemus et al., 2017; Than et al., 2008). Using a traditional bifurcating approach under multispecies coalescent model, the conflicts among independent biallelic SNP trees generated by SNAPP can be visualized by DensiTree (Bouckaert, 2010). Moreover, the ABBA-BABA or D-statistic test (Patterson et al., 2012; Hibbins and Hahn, 2022) can calculate the overall genomic introgression signals using biallelic SNPs via bioinformatic tools such as Dsuite (Malinsky et al., 2021). In addition Bayesian-based methods, such as MCMC_BiMarkers (Zhu et al., 2018) implemented in PhyloNet (Than et al., 2008) or SnappNet (Rabier et al., 2021) implemented in BEAST2 (Bouckaert et al., 2014), both are extended SNAPP that can infer network relationships of polyploids using biallelic SNPs.

Similarly, the concordance levels within a set of orthologous gene trees or between individual gene trees and the inferred species tree can indicate the ILS and complex evolutionary history of included taxa (i.e., WGD or reticulation) (Minh et al., 2020a). Concordance analysis methods such as DensiTree can also visualize the topological incongruence between a set of gene trees (e.g., Zhou et al., 2020), and PhyParts (Smith et al., 2015) also calculates the concordance level between a set of gene trees for each internode of a species tree. On the other hand, the likelihood programmes that infer species relationships with a reticulate model and ILS such as InferNetwork_ML and

InferNetwork_MPL as implemented in PhyloNet (Than et al., 2008) or SNaQ as implemented in PhyloNetworks (Solís-Lemus et al., 2017) can produce a network of species relationships by calculating the concordance level from a list of gene trees, which can be useful to identify reticulation or allopolyploidization events. However, neither tool considers the subgenome origins of the gene copy and assumes strictly two progenitors per hybrid lineage (e.g., Karbstein et al., 2022a). They require testing for the number of 'hybrid or WGD' nodes based on pseudo-likelihood or maximum-likelihood, which can be computationally demanding when inferring polyploid-rich groups with large amount of input taxa, as well as multiple rounds of hybridization and allopolyploidization events (e.g., 20 individuals with 10 hybridization events may take over one month computational time in). In addition, PhyloNet only can allow polyploids with no more than two subgenomes, i.e., maximum two alleles per gene. In addition, the robust method SplitsTree (Huson and Bryant, 2006) can be used to show the conflicts between concatenated gene sequences using a 'network' approach. Whereas SpeciesNetwork (Zhang et al., 2018) in BEAST uses a MSC model to infer each gene tree with a Bayesian method and summarises the final species tree with reticulations.

By contrast, a Bayesian-based pipeline such as Homologizer (Freyman et al., 2023) or gene-tree based method AlleleSorting (https://github.com/MarekSlenker/AlleleSorting) (Šlenker et al., 2021) can first identify the origins of each gene copy and assign the gene copy to the correct subgenome donor (i.e., phased-MUL-tree). Then the phased-MUL-trees can be summarized into a species-MUL-tree via a programme such as ASTRAL where each tree tip corresponds to each subgenome of a polyploid species (e.g., discussed in Debray et al., 2021). Moreover, phased-MUL-tree can be used to infer the species network and ILS via AlloppNET (Jones et al., 2013) with the MSC model implemented in BEAST, which also infers the network of polyploid species that have a maximum of two subgenome donors (e.g., Rothfels et al., 2017; Eriksson et al., 2018). However, phasing the allele variation and assigning to the correct subgenomes can be challenging due to bioinformatic difficulties (Ufimov et al., 2022; Zhou et al., 2022) and post-WGD genomic modification uncertainties (Li et al., 2021).

1.5 Polyploidy in the New Zealand Flora

Compared to a diploid-polyploid system (Marchant et al., 2016), a polyploid-rich group provides a more complex biological system involving WGD and reticulation histories, to investigate the diversification of closely related polyploids, which may have undergone different post-WGD macroevolutionary changes, i.e., genome modification (represented by genome size variation) or

niche shifts (e.g., Moraes et al., 2022; Wang et al., 2021a; Brittingham et al., 2018; Karbstein et al., 2022a). Moreover, to explore the interactions between post-WGD diversification and environmental conditions, insular endemic polyploid-rich genera can be particularly useful models (Soltis et al., 2009; Rice et al., 2019), because each island group can be considered as an isolated, unique system often with highly variable environmental conditions (Whittaker et al., 2017; Meudt et al., 2021).

Islands can be divided into three major categories based on their origins: oceanic islands (of volcanic or coralline origin, and never connected to the mainland); continental fragment islands (long-isolated from a continent due to the formation of new mid-ocean rifts); and continental islands (separated from the mainland due to interglacial sea-level rise) (reviewed in Whittaker and Fernández-Palacios, 2007). In particular, the oceanic islands and continental fragment islands can have many endemic plants with ancestors derived from long-distance dispersal and a colonization evolutionary history (Kier et al., 2009) that may be facilitated by wind, water, or animals [e.g., Hawaiian flora in (Price and Wagner, 2018)].

New Zealand (NZ) is a continental fragment island system that separated from the super continent Gondwana around c. 80 million years ago (Ma) (Stevens, 1980). New Zealand has a high percentage of endemic vascular plants ~82% (2551 species), which is similar to some oceanic islands, such as the Hawaiian Islands (~ 88%, ~1233 taxa) (reviewed in Meudt et al., 2021). In addition, most of the native NZ flora have ancestors that arrived via long-distance dispersal from overseas (Winkworth et al., 2002; Pole, 1994). The current NZ flora is relatively young, a result of recent dispersal and diversification. Heenan and McGlone (2019) estimated that 89% of extant taxa evolved since the end of the Miocene Thermal Optimum (17.0 to 14.45 Ma), including 50% taxa that evolved in only the last 5 Ma. As these dates correspond to similar timeframes for the origin of many oceanic island systems, NZ is often treated as an oceanic archipelago (Trewick et al., 2007; Meudt et al., 2021).

Compared to other global polyploid-rich hotspots (Rice et al., 2019), New Zealand has a high proportion of its flora with a known chromosome number (77% of extant taxa) (Murray et al., 2005), and it is also identified as a polyploid-rich island (Hair, 1966; Meudt et al., 2021). Meudt et al. (2021) calculated that 88 out of 430 native genera have species representing at least two different ploidies and one genus may contain species with as many as six different ploidy levels, ranging from 2*x* to 20*x*. This high proportion of polyploids indicates the important role that polyploidization and reticulation may have on NZ plant species diversification (Murray et al., 2005).

In addition, high species richness and endemism of insular floras can be driven by environmental heterogeneity (Barajas-Barbosa et al., 2020). The New Zealand archipelago has three main islands: North Island, South Island and Stewart Island, as well as hundreds of other offshore and smaller islands (including subantarctic islands, such as Snares Islands, Auckland Islands, Campbell Islands, and Antipodes Islands, etc.) (reviewed in Wallis and Trewick, 2009). Each main island, in particular, is characterized by its own unique biogeography, geological history, topography, soils, and landscapes (Shepherd et al., 2022; Hewitt and Dymond, 2013).

The three main islands of NZ were at different times joined together, which provided opportunities for plant migration, e.g., the North and South Islands were still joined until the mid-Pleistocene (c. 1 Ma) (Trewick and Bland, 2012), and the connection between Stewart Island and the South Island still existed during the Pleistocene before the sea level rose around 1 Ma (Lockhart et al., 2001; Cullen, 1967). However, different geological histories that may affect climates also shaped different regions of the North and South Islands. Specifically, mountain uplift on the South Island started in the early Miocene (~25 Ma) (Trewick et al., 2007; Graham, 2008). Until the Pliocene (5 to 2 Ma), the rapid uplift of the Southern Alps resulted from the formation of the Alpine Fault system and additional tectonic activity (Batt et al., 2000; Kamp et al., 1989). In particular, the glacial cycles during the Pliocene-Pleistocene (c. 5 to 0 Ma) may have prompted further divergence of ecological habitats, especially in the South Island alpine region (Winkworth et al., 2005). For example, the advance or retreat of glaciers in the Southern Alps was related to the ecological or genetic diversification of polyploid lineages such as *Ranunculus* (Ranunculaceae) (Lockhart et al., 2001), *Pachycladon* (Brassicaceae) (Joly et al., 2014), *Plantago* (Plantaginaceae) (Meudt, 2011) and *Azorella* (Apiaceae) (see Chapter 2). By contrast, mountain uplift and several significant volcanic eruptions, which started c. 1 Ma on the North Island (Holloway and McCaskill, 1982), provided novel habitats for the migration of alpine flora or the establishment of new lineages (Trewick and Bland, 2012).

1.6 Conclusion

The development of phylogenomic approaches has significantly improved our understanding of species diversification in polyploid-rich genera (Rothfels, 2021; Rothfels et al., 2017; Johnson et al., 2018; Debray et al., 2021). The financial cost of capturing genome-wide markers has decreased largely because of the more advanced sequencing technologies, increasingly available universal primers, or reference genome sequences, as well as more commercially available taxon-specific Hyb-Seq baits. Recently developed bioinformatic tools can further increase the accuracy of

extracting the orthologous loci from multiple subsets of genomes of polyploid-rich groups. However, most bioinformatic tools are still limited by the requirement of having related diploid ancestors, small number of input taxa, low number of ploidy levels or strictly only two subgenome donors. In the future, improving the bioinformatic pipelines that can tackle multiple origins of polyploids, adapting the phylogenomic approaches to third-generation sequencers and using long-reads-assisted extraction of homeologs or SNP phasing, as well as combining chromosome number and genome size data (Heslop-Harrison et al., 2022), will all further the phylogenomic inference of polyploid lineages.

1.7 Aims and Study Questions

This thesis aimed to infer polyploid species phylogenetic relationships using phylogenomic approaches, and to further understand the post-WGD macroevolutionary patterns of closely related polyploids. New Zealand polyploid species in *Azorella* sections *Schizeilema* and *Stilbocarpa* were selected to achieve the goals, given the diversity of ploidy levels and morphological traits in the New Zealand *Azorella* species. Below, three specific study questions were addressed:

1. Using two phylogenomic approaches, including target-enrichment sequencing (Hyb-Seq) and genome-skimming sequencing, can the phylogenetic relationships among *Azorella* species be resolved?

2. Will increasing the reads length of target-enrichment sequencing reads help to improve the homeologs extraction of *Azorella* polyploids?

3. Are ploidy level-associated traits informative about the post-WGD macroevolutionary patterns in *Azorella*?

Structure of the thesis:

Chapter 1 summarised and reviewed the genomic methods for studying polyploid species relationships.

Chapter 2 reconstructed the *Azorella* species relationships using Hyb-Seq with Angiosperm353 bait set, which aimed to capture up to 353 single copy nuclear genes, and whole plastid DNA and nrDNA recovered from the genome skimming reads. In addition, networks of *Azorella* polyploids were inferred from Hyb-Seq reads using gene sequences, as well as SNP variations.

Chapter 3 aimed to combine Hyb-Seq with PacBio sequence data to increase the read length and to improve homeolog extraction for the targeted genes. Moreover, the ploidy-level associated traits, such as the variation in genome sizes (via flow cytometry), stomatal guard cell length (using scanning electron microscopy), and ecological niches (using the R package ENMTools) were assessed and compared between species in different ploidy levels.

Chapter 4 provided an overview of the thesis main outcomes and future perspectives.

Chapters 1, 2, 3 are intended to be separate publications and are formatted for the appropriate journal to which each will be submitted.

1.8 References Cited

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution,* 29: 417-434.

Andermann T, Torres Jiménez MF, Matos-Maraví P, Batista R, Blanco-Pastor JL, Gustafsson ALS, *et al.* 2020. A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics,* 10.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics,* 17: 81-92.

Baker WJ, Bailey P, Barber V, Barker A, Bellot S, Bishop D, *et al.* 2022. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology,* 71: 301-319.

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*: 247-277.

Barajas-Barbosa MP, Weigelt P, Borregaard MK, Keppel G, Kreft H. 2020. Environmental heterogeneity dynamics drive plant diversity on oceanic islands. *Journal of Biogeography,* 47: 2248-2260.

Barnard-Kubow KB, McCoy MA, Galloway LF. 2017. Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. *New Phytologist,* 213: 1466-1476.

Batt GE, Braun J, Kohn BP, McDougall I. 2000. Thermochronological analysis of the dynamics of the Southern Alps, New Zealand. *GSA Bulletin,* 112: 250-266.

Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences,* 92: 11331-11338.

Blischak PD, Latvis M, Morales-Briones DF, Johnson JC, Di Stilio VS, Wolfe AD, *et al.* 2018. Fluidigm2PURC: automated processing and haplotype inference for double-barcoded PCR amplicons. *Applications in Plant Sciences,* 6: e01156.

Borsch T, Quandt D, Koch M. 2009. Molecular evolution and phylogenetic utility of non-coding DNA: applications from species to deep level questions. *Plant Systematics and Evolution,* 282: 107-108.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, *et al.* 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Computational Biology,* 10: e1003537.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics,* 26: 1372-1373.

Breinholt JW, Carey SB, Tiley GP, Davis EC, Endara L, McDaniel SF, *et al.* 2021. A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences,* 9: e11406.

Brittingham HA, Koski MH, Ashman T-L. 2018. Higher ploidy is associated with reduced range breadth in the Potentilleae tribe. *American Journal of Botany,* 105: 700-710.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution,* 29: 1917-1932.

Brysting AK, Mathiesen C, Marcussen T. 2011. Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine *Cerastium alpinum* complex. *TAXON,* 60: 333-347.

Brysting AK, Oxelman B, Huber KT, Moulton V, Brochmann C. 2007. Untangling complex histories of genome mergings in high polyploids. *Systematic Biology,* 56: 467-476.

Cerca J, Petersen B, Lazaro-Guevara JM, Rivera-Colón A, Birkeland S, Vizueta J, *et al.* 2022. The genomic basis of the plant island syndrome in Darwin's giant daisies. *Nature Communications,* 13: 3729.

Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, *et al.* 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences,* 3: 1400115.
Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology,* 58: 377-406.
Cheon S, Zhang J, Park C. 2020. Is phylotranscriptomics as reliable as phylogenomics? *Molecular Biology and Evolution,* 37: 3672-3683.
Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics,* 30: 3317-3324.
Clark JW, Donoghue PCJ. 2018. Whole-genome duplication and plant macroevolution. *Trends in Plant Science,* 23: 933-945.
Clarkson JJ, Dodsworth S, Chase MW. 2017. Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (Solanaceae). *Plant Systematics and Evolution,* 303: 1001-1012.
Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics,* 6: 836-846.
Crowl AA, Fritsch PW, Tiley GP, Lynch NP, Ranney TG, Ashrafi H, *et al.* 2022. A first complete phylogenomic hypothesis for diploid blueberries (*Vaccinium* section *Cyanococcus*). *American Journal of Botany,* 109: 1596-1606.
Cullen DJ. 1967. Submarine evidence from New Zealand of a rapid rise in sea level about 11,000 years B.P. *Palaeogeography, Palaeoclimatology, Palaeoecology,* 3: 289-298.
Czabarka É, Erdős PL, Johnson V, Moulton V. 2013. Generating functions for multi-labeled trees. *Discrete Applied Mathematics,* 161: 107-117.
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* 12: 499-510.
de Lima Ferreira P, Batista R, Andermann T, Groppo M, Bacon CD, Antonelli A. 2022. Target sequence capture of Barnadesioideae (Compositae) demonstrates the utility of low coverage loci in phylogenomic analyses. *Molecular Phylogenetics and Evolution*: 107432.
Debray K, Le Paslier M-C, Bérard A, Thouroude T, Michel G, Marie-Magdelaine J, *et al.* 2021. Unveiling the patterns of reticulated evolutionary processes with phylogenomics: hybridization and polyploidy in the genus *Rosa*. *Systematic Biology,* 71: 547-569.
Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution,* 24: 332-340.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* 43: 491-498.
Dodsworth S, Chase MW, Leitch AR. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society,* 180: 1-5.
Dodsworth S, Jang T-S, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. 2017. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution,* 303: 1013-1020.
Eaton DAR, Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics,* 36: 2592-2594.
Eaton DAR, Spriggs EL, Park B, Donoghue MJ. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology,* 66: 399-412.
Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology,* 9: 157.

Edger PP, McKain MR, Bird KA, VanBuren R. 2018. Subgenome assignment in allopolyploids: challenges and future directions. *Current Opinion in Plant Biology*, 42: 76-80.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16: 157.
Eriksson JS, de Sousa F, Bertrand YJK, Antonelli A, Oxelman B, Pfeil BE. 2018. Allele phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M. strasseri* (Fabaceae). *BMC Evolutionary Biology*, 18: 9.
Fér T, Schmickl RE. 2018. HybPhyloMaker: target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics*, 14: 1176934317742613.
Freudenthal JA, Pfaff S, Terhoeven N, Korte A, Ankenbrand MJ, Förster F. 2020. A systematic comparison of chloroplast genome assembly tools. *Genome Biology*, 21: 254.
Freyman WA, Johnson MG, Rothfels CJ. 2023. Homologizer: phylogenetic phasing of gene copies into polyploid subgenomes. *Methods in Ecology and Evolution*, 14: 1230-1244.
Frost LA, O'Leary N, Lagomarsino LP, Tank DC, Olmstead RG. 2021. Phylogeny, classification, and character evolution of tribe Citharexyleae (Verbenaceae). *American Journal of Botany*, 108: 1982-2001.
Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, Macintyre RJ, Clegg MT, eds. *Evolutionary Biology*. Boston, MA: Springer US.
Glover NM, Redestig H, Dessimoz C. 2016. Homoeologs: what are they and how do we infer them? *Trends in Plant Science*, 21: 609-621.
Graham IJ. 2008. *A continent on the move: New Zealand geoscience into the 21st century*: Geological Society of New Zealand.
Guo C, Luo Y, Gao L-M, Yi T-S, Li H-T, Yang J-B, *et al.* 2023. Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology*, 65: 299-323.
Hair JB. 1966. Biosystematics of the New Zealand flora, 1945–1964. *New Zealand Journal of Botany*, 4: 559-595.
Han T-S, Zheng Q-J, Onstein RE, Rojas-Andrés BM, Hauenschild F, Muellner-Riehl AN, *et al.* 2020. Polyploidy promotes species diversification of *Allium* through ecological shifts. *New Phytologist*, 225: 571-583.
Hart ML, Forrest LL, Nicholls JA, Kidner CA. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon*, 65: 1081-1092.
Heenan PB, McGlone MS. 2019. Cenozoic formation and colonisation history of the New Zealand vascular flora based on molecular clock dating of the plastid rbcL gene. *New Zealand Journal of Botany*, 57: 204-226.
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27: 570-580.
Hendriks KP, Mandáková T, Hay NM, Ly E, Hooft van Huysduynen A, Tamrakar R, *et al.* 2021. The best of both worlds: combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences*, 9: e11438.
Heslop-Harrison JS, Schwarzacher T, Liu Q. 2022. Polyploidy: its consequences and enabling role in plant diversification and evolution. *Annals of Botany*: mcac132.
Hewitt A, Dymond J. 2013. Survey of New Zealand soil orders. *Ecosystem services in New Zealand: conditions and trends*. Lincoln, Canterbury, New Zealand: Manaaki Whenua Press.
Hibbins MS, Hahn MW. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220: iyab173.
Hillis DM, Dixon MT. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology* 66: 411-53.
Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4: 275-284.

Holloway JT, McCaskill LW. 1982. *The mountain lands of New Zealand*: Tussock Grasslands & Mountain Lands Institute, Lincoln College.
Huang C-C, Hung K-H, Wang W-K, Ho C-W, Huang C-L, Hsu T-W, *et al.* 2012. Evolutionary rates of commonly used nuclear and organelle markers of *Arabidopsis* relatives (Brassicaceae). *Gene,* 499: 194-201.
Huber KT, Lott M, Moulton V, Spillner A. 2008. The complexity of deriving multi-labeled trees from bipartitions. *Journal of Computational Biology,* 15: 639-651.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution,* 23: 254-267.
Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology,* 61: 1061-1067.
Jackson C, McLay T, Schmidt-Lebuhn AN. 2021. hybpiper-rbgv and yang-and-smith-rbgv: Containerization and additional options for assembly and paralog detection in target enrichment data. *bioRxiv*: 2021.11.08.467817.
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature,* 473: 97-100.
Jin J-J, Yu W-B, Yang J-B, Song Y, DePamphilis CW, Yi T-S, *et al.* 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology,* 21: 1-31.
Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, *et al.* 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences,* 4: apps.1600016.
Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, *et al.* 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology,* 68: 594-606.
Joly S, Heenan PB, Lockhart PJ. 2014. Species radiation by niche shifts in New Zealand's rockcresses (*Pachycladon*, Brassicaceae). *Systematic Biology,* 63: 192-202.
Jonas M-R, Burleigh JG, Erin MS. 2023. Target capture methods offer insight into the evolution of rapidly diverged taxa and resolve allopolyploid homeologs in the fern genus *Polypodium* s.s. *Systematic Botany,* 48: 96-109.
Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology,* 62: 467-478.
Kamp PJJ, Green PF, White SH. 1989. Fission track analysis reveals character of collisional tectonics in New Zealand. *Tectonics,* 8: 169-195.
Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics,* 21: 428-444.
Karbstein K, Tomasello S, Hodač L, Wagner N, Marinček P, Barke BH, *et al.* 2022. Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large Ranunculus auricomus species complex. *New Phytologist,* 235: 2081-2098.
Karimi N, Grover CE, Gallagher JP, Wendel JF, Ané C, Baum DA. 2020. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; Bombacoideae; Malvaceae). *Systematic Biology,* 69: 462-478.
Kier G, Kreft H, Lee TM, Jetz W, Ibisch PL, Nowicki C, *et al.* 2009. A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences,* 106: 9322-9327.
Kircher M, Kelso J. 2010. High-throughput DNA sequencing – concepts and limitations. *BioEssays,* 32: 524-536.
Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology,* 56: 17-24.

Leaché AD, Oaks JR. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution and Systematics,* 48: 69-84.
Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, *et al.* 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature,* 574: 679-685.
Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society,* 82: 651-663.
Leung SK, Jeffries AR, Castanho I, Jordan BT, Moore K, Davies JP, *et al.* 2021. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports,* 37: 110022.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25: 2078-2079.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22: 1658-1659.
Li Z, McKibben MTW, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology,* 72: 387-410.
Liu BB, Ma ZY, Ren C, Hodel RG, Sun M, Liu XQ, *et al.* 2021. Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae. *Journal of Systematics and Evolution,* 59: 1124-1138.
Lockhart PJ, McLenachan PA, Havell D, Glenny D, Huson D, Jensen U. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. *Annals of the Missouri Botanical Garden,* 88: 458-477.
Lozano-Fernandez J. 2022. A practical guide to design and assess a phylogenomic study. *Genome Biology and Evolution,* 14: evac129.
Lu W-X, Hu X-Y, Wang Z-Z, Rao G-Y. 2022. Hyb-Seq provides new insights into the phylogeny and evolution of the *Chrysanthemum zawadskii* species complex in China. *Cladistics,* n/a.
Maddison WP. 1997. Gene trees in species trees. *Systematic Biology,* 46: 523-536.
Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology,* 55: 21-30.
Malinsky M, Matschiner M, Svardal H. 2021. Dsuite-fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources,* 21: 584-595.
Malinsky M, Trucchi E, Lawson DJ, Falush D. 2018. RADpainter and fineRADstructure: population Inference from RADseq Data. *Molecular Biology and Evolution,* 35: 1284-1290.
Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays,* 38: 140-149.
Marchant BD, Soltis DE, Soltis PS. 2016. Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytologist,* 212: 708-718.
Mayrose I, Zhan SH, Rothfels CJ, Arrigo N, Barker MS, Rieseberg LH, *et al.* 2015. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytologist,* 206: 27-35.
McCarthy EW, Chase MW, Knapp S, Litt A, Leitch AR, Le Comber SC. 2016. Transgressive phenotypes and generalist pollination in the floral evolution of *Nicotiana* polyploids. *Nature Plants,* 2: 16119.
McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences,* 6: e1038.
McLay TGB, Birch JL, Gunn BF, Ning W, Tate JA, Nauheimer L, *et al.* 2021. New targets acquired: improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences,* 9.
Meudt HM. 2011. Amplified fragment length polymorphism data reveal a history of auto- and allopolyploidy in New Zealand endemic species of *Plantago* (Plantaginaceae): new

perspectives on a taxonomically challenging group. *International Journal of Plant Sciences,* 172: 220-237.
Meudt HM, Albach DC, Tanentzap AJ, Igea J, Newmarch SC, Brandt AJ, *et al.* 2021. Polyploidy on islands: its emergence and importance for diversification. *Frontiers in Plant Science,* 12.
Meudt HM, Rojas-Andrés BM, Prebble JM, Low E, Garnock-Jones PJ, Albach DC. 2015. Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Botanical Journal of the Linnean Society,* 178: 243-266.
Michael SB, Nils A, Anthony EB, Zheng L, Donald AL. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist,* 210: 391-398.
Michel T, Tseng Y-H, Wilson H, Chung K-F, Kidner C. 2022. A hybrid capture bait set for *Begonia*. *Edinburgh Journal of Botany,* 79: 1-33.
Minh BQ, Hahn MW, Lanfear R. 2020a. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution,* 37: 2727-2733.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, *et al.* 2020b. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution,* 37: 1530-1534.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics,* 30: i541-i548.
Moraes AP, Engel TBJ, Forni-Martins ER, de Barros F, Felix LP, Cabral JS. 2022. Are chromosome number and genome size associated with habit and environmental niche variables? Insights from the Neotropical orchids. *Annals of Botany,* 130: 11-25.
Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, *et al.* 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl. *Systematic Biology,* 70: 219-235.
Morales-Briones DF, Liston A, Tank DC. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytologist,* 218: 1668-1684.
Morrison D. 2014. Phylogenetic networks: a review of methods to display evolutionary history. *Annual Research & Review in Biology,* 4: 1518-1543.
Murray BG, De Lange PJ, Ferguson AR. 2005. Nuclear DNA variation, chromosome numbers and polyploidy in the endemic and indigenous grass flora of New Zealand. *Annals of Botany,* 96: 1293-1305.
Nauheimer L, Weigner N, Joyce E, Crayn D, Clarke C, Nargar K. 2021. HybPhaser: a workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences,* 9: e11441.
Novikova PY, Hohmann N, Van de Peer Y. 2018. Polyploid *Arabidopsis* species originated around recent glaciation maxima. *Current Opinion in Plant Biology,* 42: 8-15.
Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution,* 34: 2101-2114.
Oshiki M, Miura T, Kazama S, Segawa T, Ishii S, Hatamoto M, *et al.* 2018. Microfluidic PCR Amplification and MiSeq Amplicon Sequencing Techniques for High-Throughput Detection and Genotyping of Human Pathogenic RNA Viruses in Human Feces, Sewage, and Oysters. *Frontiers in Microbiology,* 9: 830.
Osuna-Mascaró C, de Casas RR, Berbel M, Gómez JM, Perfectti F. 2022. Lack of ITS sequence homogenization in congeneric plant species with different ploidy levels. *bioRxiv*: 2022.05.29.493735.
Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE. 2017. Phylogenetics of allopolyploids. *Annual Review of Ecology, Evolution, and Systematics,* 48: 543-557.

Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, *et al.* 2015. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology,* 22: 498-509.
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, *et al.* 2012. Ancient admixture in human history. *Genetics,* 192: 1065-1093.
Pleines T, Jakob SS, Blattner FR. 2009. Application of non-coding DNA regions in intraspecific analyses. *Plant Systematics and Evolution,* 282: 281-294.
Pole M. 1994. The New Zealand flora-entirely long-distance dispersal? *Journal of Biogeography,* 21: 625-635.
Postel Z, Touzet P. 2020. Cytonuclear genetic incompatibilities in plant speciation. *Plants,* 9: 487.
Price JP, Wagner WL. 2018. Origins of the Hawaiian flora: phylogenies and biogeography reveal patterns of long-distance dispersal. *Journal of Systematics and Evolution,* 56: 600-620.
Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics,* 3: 827-837.
Qiu F, Baack EJ, Whitney KD, Bock DG, Tetreault HM, Rieseberg LH, *et al.* 2019. Phylogenetic trends and environmental correlates of nuclear genome size variation in *Helianthus* sunflowers. *New Phytologist,* 221: 1609-1618.
Qiu T, Liu Z, Liu B. 2020. The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Molecular Biology Reports,* 47: 5549-5558.
Rabier C-E, Berry V, Stoltz M, Santos JD, Wang W, Glaszmann J-C, *et al.* 2021. On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLOS Computational Biology,* 17: e1008380.
Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution,* 271: 101-122.
Rice A, Šmarda P, Novosolov M, Drori M, Glick L, Sabath N, *et al.* 2019. The global biogeography of polyploid plants. *Nature Ecology and Evolution,* 3: 265-273.
Rothfels CJ. 2021. Polyploid phylogenetics. *New Phytologist,* 230: 66-72.
Rothfels CJ, Pryer KM, Li FW. 2017. Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist,* 213: 413-429.
Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology,* 37: 121-147.
Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, *et al.* 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources,* 16: 1124-1135.
Shan S, Boatwright JL, Liu X, Chanderbali AS, Fu C, Soltis PS, *et al.* 2020. Transcriptome dynamics of the inflorescence in reciprocally formed allopolyploid *Tragopogon miscellus* (Asteraceae). *Frontiers in Genetics,* 11.
Sharbrough J, Conover JL, Tate JA, Wendel JF, Sloan DB. 2017. Cytonuclear responses to genome doubling. *American Journal of Botany,* 104: 1277-1280.
Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany,* 101: 1987-2004.
Shepherd L, Simon C, Langton-Myers S, Morgan-Richards M. 2022. Insights into Aotearoa New Zealand's biogeographic history provided by the study of natural hybrid zones. *Journal of the Royal Society of New Zealand*: 1-20.
Šlenker M, Kantor A, Marhold K, Schmickl R, Mandáková T, Lysak MA, *et al.* 2021. Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq data: a

case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. *Frontiers in Plant Science,* 12.
Slimp M, Williams LD, Hale H, Johnson MG. 2021. On the potential of Angiosperms353 for population genomic studies. *Applications in Plant Sciences,* 9.
Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany,* 17: 145-170.
Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology,* 15: 1-15.
Solís-Lemus C, Bastide P, Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. *Molecular Biology and Evolution,* 34: 3292-3298.
Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, *et al.* 2009. Polyploidy and angiosperm diversification. *American Journal of Botany,* 96: 336-348.
Soltis DE, Soltis PS, Schemske DW, Hancock JF, Thompson JN, Husband BC, *et al.* 2007. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon,* 56: 13-30.
Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then…and now: Stebbins revisited. *American Journal of Botany,* 101: 1057-1078.
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development,* 35: 119-125.
Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology,* 60: 561-588.
Stevens GR. 1980. *New Zealand adrift: the theory of continental drift in a New Zealand setting*: AH & AW Reed.
Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany,* 99: 349-364.
Takamatsu T, Baslam M, Inomata T, Oikawa K, Itoh K, Ohnishi T, *et al.* 2018. Optimized method of extracting rice chloroplast DNA for high-Quality plastome resequencing and *de novo* assembly. *Frontiers in Plant Science,* 9.
Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics,* 9: 322.
Thomas AE, Igea J, Meudt HM, Albach DC, Lee WG, Tanentzap AJ. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *American Journal of Botany,* 108: 1289-1306.
Tiley GP, Crowl AA, Manos PS, Sessa EB, Solís-Lemus C, Yoder AD, *et al.* 2021. Phasing alleles improves network inference with allopolyploids. *bioRxiv*: 2021.05.04.442457.
Trewick SA, Bland KJ. 2012. Fire and slice: palaeogeography for biogeography at New Zealand's North Island/South Island juncture. *Journal of the Royal Society of New Zealand,* 42: 153-183.
Trewick SA, Paterson AM, Campbell HJ. 2007. Guest Editorial: Hello New Zealand. *Journal of Biogeography,* 34: 1-6.
Tsitrone A, Kirkpatrick M, Levin DA. 2003. A model for chloroplast capture. *Evolution,* 57: 1776-1782.
Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. *Heredity,* 108: 179-189.
Ufimov R, Gorospe JM, Fér T, Kandziora M, Salomon L, van Loo M, *et al.* 2022. Utilizing paralogues for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data. *Molecular Ecology Resources,* 22: 3018-3034.

Unruh SA, McKain MR, Lee Y-I, Yukawa T, McCormick MK, Shefferson RP, *et al.* 2018. Phylotranscriptomic analysis and genome evolution of the Cypripedioideae (Orchidaceae). *American Journal of Botany*, 105: 631-640.
Uribe-Convers S, Settles ML, Tank DC. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLOS One*, 11: e0148203.
Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, 120: 195-207.
Wallis GP, Trewick SA. 2009. New Zealand phylogeography: evolution on a small continent. *Molecular Ecology*, 18: 3548-3580.
Wang G, Zhou N, Chen Q, Yang Y, Yang Y, Duan Y. 2021a. Gradual genome size evolution and polyploidy in *Allium* from the Qinghai–Tibetan Plateau. *Annals of Botany*: 109–122.
Wang N, Kelly LJ, McAllister HA, Zohren J, Buggs RJ. 2021b. Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees. *Molecular Phylogenetics and Evolution*, 160: 107126.
Wang S, Gao J, Chao H, Li Z, Pu W, Wang Y, *et al.* 2022. Comparative chloroplast genomes of *Nicotiana* species (Solanaceae): insights into the genetic variation, phylogenetic relationship, and polyploid speciation. *Frontiers in Plant Science*, 13.
Wang X, Morton JA, Pellicer J, Leitch IJ, Leitch AR. 2021c. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal*, 107: 1003-1015.
Wei N, Tennessen JA, Liston A, Ashman T-L. 2017. Present-day sympatry belies the evolutionary origin of a high-order polyploid. *New Phytologist*, 216: 279-290.
Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, *et al.* 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, 2: 1400042.
Whittaker RJ, Fernández-Palacios JM. 2007. *Island biogeography: ecology, evolution, and conservation*: Oxford University Press.
Whittaker RJ, Fernández-Palacios JM, Matthews TJ, Borregaard MK, Triantis KA. 2017. Island biogeography: taking the long view of nature's laboratories. *Science*, 357: eaam8326.
Winkworth RC, Wagstaff SJ, Glenny D, Lockhart PJ. 2002. Plant dispersal N.E.W.S from New Zealand. *Trends in Ecology & Evolution*, 17: 514-520.
Winkworth RC, Wagstaff SJ, Glenny D, Lockhart PJ. 2005. Evolution of the New Zealand mountain flora: origins, diversification and dispersal. *Organisms, Diversity & Evolution*, 5: 237-247.
Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106: 13875-13879.
Xie M, Wu Q, Wang J, Jiang T. 2016. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics*, 32: 3735-3744.
Xu B, Zeng X-M, Gao X-F, Jin D-P, Zhang L-B. 2017. ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Scientific Reports*, 7: 1-15.
Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. 2022. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology*, 71: 367-381.
Yang X, Ye C-Y, Cheng Z-M, Tschaplinski TJ, Wullschleger SD, Yin W, *et al.* 2011. Genomic aspects of research involving polyploid plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 104: 387-397.
Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, 31: 3081-3092.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics,* 13: 303-314.
Yardeni G, Viruel J, Paris M, Hess J, Groot Crego C, de La Harpe M, *et al.* 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources,* 22: 927-945.
Zhang C, Huang C-H, Liu M, Hu Y, Panero JL, Luebert F, *et al.* 2021. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *Journal of Integrative Plant Biology,* 63: 1273-1293.
Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution,* 35: 504-517.
Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution,* 37: 3292-3307.
Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics,* 16: 409-420.
Zhang Y, Ozdemir P. 2009. Microfluidic DNA amplification—A review. *Analytica Chimica Acta,* 638: 115-125.
Zhou R, Moshgabadi N, Adams KL. 2011. Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. *Proceedings of the National Academy of Sciences,* 108: 16122-16127.
Zhou W, Soghigian J, Xiang Q-Y. 2022. A new pipeline for removing paralogs in target enrichment data. *Systematic Biology,* 71: 410-425.
Zhou W, Xiang Q-Y, Wen J. 2020. Phylogenomics, biogeography, and evolution of morphology and ecological niche of the eastern Asian–eastern North American *Nyssa* (Nyssaceae). *Journal of Systematics and Evolution,* 58: 571-603.
Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLOS Computational Biology,* 14: e1005932.

Chapter 2. Resolving reticulate evolutionary histories of polyploid *Azorella* (Apiaceae) species in New Zealand

Abstract

Genera with species of multiple ploidal levels are important models for investigating the role of polyploidization and reticulation in plant diversification. Here, we studied 17 polyploid taxa (species, subspecies, or varieties) in *Azorella* sections *Schizeilema* and *Stilbocarpa* that have divergent leaf morphologies, distinct distributional ranges, and varying ploidal levels (4*x*, 6*x*, and 10*x*). All are endemic to New Zealand, except the Australian endemic *A. fragosea*. Our goals were to assess the biogeographic origins of the New Zealand species, resolve species relationships, and determine the origins of the higher polyploids (i.e., 6*x* and 10*x*). Including the South American outgroup species, we reconstructed the phylogeny of 125 individuals representing 20 *Azorella* taxa collected from 72 sites using the Angiosperms353 bait set and Hyb-Seq captured single copy nuclear genes (SCNGs). We also reconstructed phylogenies of complete plastomes and nrDNA that were extracted from genome-skimming reads from 104 individuals representing 19 *Azorella* taxa. Our results showed topological incongruence between the SCNG, plastome, nrDNA phylogenies. In the SCNG phylogeny, species in the monophyletic section *Schizeilema* were sister to their South American relatives, whereas section *Stilbocarpa* formed a separate clade sister to the South American relatives and section *Schizeilema* clade. Biogeographical analyses indicated that New Zealand section *Schizeilema* likely originated from South American ancestors (c. 12.1 Ma) via long distance dispersal, whereas the ancestor to megaherbs in section *Stilbocarpa* may have dispersed from South America or arrived via Antarctica to the subantarctic islands c. 25.85 Ma. Through SCNG network analyses, genomic introgression test, and comparison of incongruent plastome and nrDNA phylogenies, we identified the following reticulate evolutionary histories in section *Schizeilema*: the hybrid origins of three hexaploids, *A. hookeri*, *A. nitens* and *A. cockaynei*; the only decaploid *A. colensoi* originated from allopolyploidization between tetraploid *A. allanii* and hexaploid *A. hookeri*; a monophyletic group that included the South Island endemic alpine species that are mostly tetraploids (*A. exigua*, *A. haastii* subsp. *haastii*, *A. haastii* subsp. *cyanopetala*, *A. hydrocotyloides* and *A. pallida*) may have originated from reticulate events between ancestors of the tetraploids *A. allanii* and *A. roughii* from New Zealand with Australian *A. fragosea*. Overall, we found the Angiosperms353 baits, in combination with genome-skimming data, are useful to reveal the origins and reticulate relationships between species in a polyploid-rich genus.

Keywords: Allopolyploidy; Angiosperms353; Biogeography; Hybridization; Hyb-Seq; Genome-skimming; Phylogenomics.

2.1 Introduction

Polyploidization or whole genome duplication (WGD) is a major source of evolutionary variation that can promote short-term radiation and long-term divergence among plant species (Soltis et al., 2009; Clark and Donoghue, 2018; Jiao et al., 2011b). Such events are important for the diversification of angiosperm species, especially for species with more recent WGD events (i.e., neopolyploids) (Van de Peer et al., 2017). In particular, whole genome duplication can occur repeatedly within a plant genus to form a polyploid-rich genus that contains multiple species with differing chromosome numbers and thereby ploidal levels, e.g., genera reviewed by Meudt et al. (2021). Species within a polyploid-rich genus can have diverse morphological traits, as well as the potential to adapt to broad geographical ranges or diverse ecological habitats (Winkworth et al., 2005; Joly et al., 2014; Qiu et al., 2019). Therefore, polyploid-rich genera are ideal models to understand the mechanisms of polyploidization-driven genomic evolution and species diversification.

Understanding the origins and genetic relationships of polyploid species with phylogenetic approaches is the first step to examining their diversification patterns. Species may originate from the duplication of a single genomic source (autopolyploidization) or the combination of different genomic sources (allopolyploidization) (reviewed by Otto and Whitton, 2000). In a polyploid-rich genus, additional reticulate events between closely related polyploids, via hybridization and/or introgression, can also mix multiple copies of the same gene from different species (homeologs) into one genome, further increasing the complexity of phylogenetic inference (reviewed in Rothfels, 2021). Reconstructing their origins with traditional phylogenetic approaches (i.e., a bifurcating phylogenetic tree using a single locus) is challenging (Small et al., 2004). In particular, inferring origins for higher polyploids (e.g., hexaploids, octoploids, etc.) that may have experienced multiple rounds of whole genome duplication poses additional difficulties. To be successful in unravelling the origins of polyploid species, including genetic markers that have more informative sites and different evolutionary histories, as well as utilizing additional network analyses that allow reticulate relationships are required (reviewed in Soltis et al., 2014; Hibbins and Hahn, 2022).

Traditionally, a handful of high copy markers, such as plastid DNA markers (cpDNA) or nuclear ribosomal DNA (nrDNA), specifically the internal and external transcribed spacer (ITS and ETS) regions, have been widely used for reconstructing phylogenetic relationships of plants (Olmstead and Palmer, 1994). These high copy markers are easily amplified using polymerase chain reaction (PCR) because they are abundant in cells, and the conservation of genes allows universal primers to be used across taxa (Hillis and Dixon, 1991; Baldwin et al., 1995; Shaw et al., 2007). Another

efficient approach to recover these genes, especially the whole plastome as well as homeologous copies of biparentally inherited nrDNA, is by extracting them from genome-skimming sequence reads (Straub et al., 2012). In particular, homeologs of ITS that have not been affected by concerted evolution can be informative regarding the reticulate relationships of polyploid species (Rauscher et al., 2004; Wan et al., 2014; Xu et al., 2017). However, for genera with species that have experienced multiple whole genome duplication and reticulation events, such high copy markers may be of limited value.

For example, plastomes are typically uniparentally inherited in flowering plants, therefore, only contain information from the maternal lineage (Birky, 1995). Additionally, cpDNA may exhibit geographical structure (due to local introgression) instead of reflecting phylogenetic or taxonomic relationships (Tsitrone et al., 2003). Moreover, nrDNA may not contain sufficient informative sites, resulting in polytomies in the phylogenetic tree, or the homogenization of nrDNA via concerted evolution may have occurred, such that the historical signatures of hybridization or allopolyploidization have been lost (Álvarez and Wendel, 2003). Given the challenges of using high copy genes, as well as the limited resolution provided by these markers, a phylogenomic approach that takes advantage of genome wide signals is required to investigate polyploid groups with complex evolutionary histories.

Genome-wide single copy nuclear genes (SCNGs), in contrast to high copy markers, may be more variable and have the potential to resolve phylogenetic relationships of closely related species (Karbstein et al., 2022b; Wang et al., 2021c; Thomas et al., 2021), as well as estimate population-level variation within species (Beck et al., 2021; Slimp et al., 2021). Combining a target-enrichment approach (Hyb-Seq) with lineage-specific or universal baits allows the capture of hundreds of SCNGs in one reaction (Weitemier et al., 2014; Johnson et al., 2018; Hendriks et al., 2021; Yardeni et al., 2022). As samples can be individually barcoded and sequenced in parallel using a next-generation sequencing platform, the efficiency in generating SCNG data for hundreds of samples increases phylogenetic power to resolve their relationships (Baker et al., 2022). Furthermore, the biparentally inherited SCNGs are not subjected to concerted evolution, therefore recovering homeologs from Hyb-Seq sequenced reads can further improve the genealogical reconstruction of each gene (Nauheimer et al., 2021; Tiley et al., 2021). However, Hyb-Seq reads phasing can still be limited by, e.g., the short sequencing read length; bioinformatic pipelines that cannot handle polyploids with multiple subgenome donors or low divergence between homeologous sequences; and downstream analyses that assume strictly two parental lineages (Johnson et al., 2016; McKain et al., 2018; Solís-Lemus et al., 2017; Than et al., 2008; Karbstein et al., 2022a).

In this study, we aim to reconstruct the origins and evolutionary history of New Zealand polyploid species in the genus *Azorella* (Apiaceae) (Table S1). Recent phylogenetic studies of *Azorella* and five other closely related genera from the southern hemisphere led to the transfer of species from six genera to *Azorella* (Plunkett and Nicolas, 2017). The genus now contains more than 60 species from South America, New Zealand, south-eastern Australia, and the subantarctic islands. In New Zealand, *Azorella* comprises two well-defined sections - sect. *Schizeilema* and sect. *Stilbocarpa* - according to the most recent phylogenetic reclassification based on nrDNA and two plastid regions (Plunkett and Nicolas, 2017). There are 17 perennial species, subspecies, and varieties (hereafter, taxa) described in these two sections, which vary in ploidal level [4*x*, 6*x*, and 10*x*, where *x* = 8; (Hair, 1980; Beuzenberg and Hair, 1983)], leaf morphological traits (Fig. 1; Fig. S1) (reviewed in Allan, 1961), and geographical distributions (Fig. S2).

The three species in section *Stilbocarpa, A. polaris* (6*x*), *A. robusta* (?*x* = unknown ploidal level) and *A. lyallii* (?*x*), are megaherbs endemic to Stewart Island and the subantarctic islands of New Zealand (Fig S1) (Beuzenberg and Hair, 1983; Mitchell et al., 1999; McGlone, 2002). By contrast, the 14 taxa in section *Schizeilema* are smaller rhizomatous rosette herbs and are all endemic to mainland New Zealand and Stewart Island. The exceptions are *A. schizeilema* (4*x*) (Fig. S2), which is endemic to the subantarctic Auckland Islands and Campbell Island, and the Australian *A. fragosea* (?*x*), which is endemic to Australia (New South Wales and Victoria). Previously counted chromosome numbers of New Zealand *Azorella* plants suggested that most species in section *Schizeilema* are tetraploids, except for *A. cockaynei*, *A. nitens*, *A. hookeri* and *A. pallida*, which are hexaploids, and *A. colensoi*, the only decaploid (Hair, 1980). Biogeographical patterns inferred from phylogenetic analysis of two plastid markers indicated that *Azorella* sections *Schizeilema* and *Stilbocarpa* originated from ancestors in Chile and Argentina, and were independently dispersed to New Zealand and the subantarctic islands (Nicolas and Plunkett, 2014; Nicolas and Plunkett, 2012). However, the relationships among New Zealand *Azorella* species were not fully resolved in previous phylogenetic analyses of high copy genes (Plunkett and Nicolas, 2017; Fernández et al., 2017), which also could not resolve a polytomy of branches leading to sections *Azorella*, *Huanaca*, *Schizeilema*, *Stilbocarpa* and *Ranunculus*. In particular, given the various ploidal levels among the New Zealand endemic *Azorella* species, there may have been multiple hybridization and polyploidization events within the sections, which require further sampling and additional phylogenomic analyses utilizing SCNGs to understand their phylogenetic relationships.

To resolve the phylogenetic relationships of 17 New Zealand *Azorella* taxa, we combined two sequencing approaches, namely Hyb-Seq and genome-skimming, to extract SCNGs and high copy

genes, respectively. Specifically, the Angiosperms353 baits kit was applied to capture 353 SCNGs (Johnson, et al. 2018), and the high copy genes of complete nrDNA and plastomes were recovered from genome-skim sequencing. We also included two South American species, *A. ranunculus* and *A. burkartii* from section *Ranunculus* as outgroups to sections *Stilbocarpa* and *Schizeilema*, based on previous ITS trees (Fernández et al., 2017; Plunkett and Nicolas, 2017). The main aims of this study were to use a phylogenetic approach to 1) estimate the divergence times and the biogeographical history of New Zealand *Azorella*; and 2) resolve the origins of the New Zealand polyploid species. In this way, we aim to test the utility of SCNGs and different analysis methods to reconstruct the reticulate history of the polyploid-rich genus *Azorella* in New Zealand, and discover whether network approaches are useful to explain the topological incongruence between trees from different data sets.

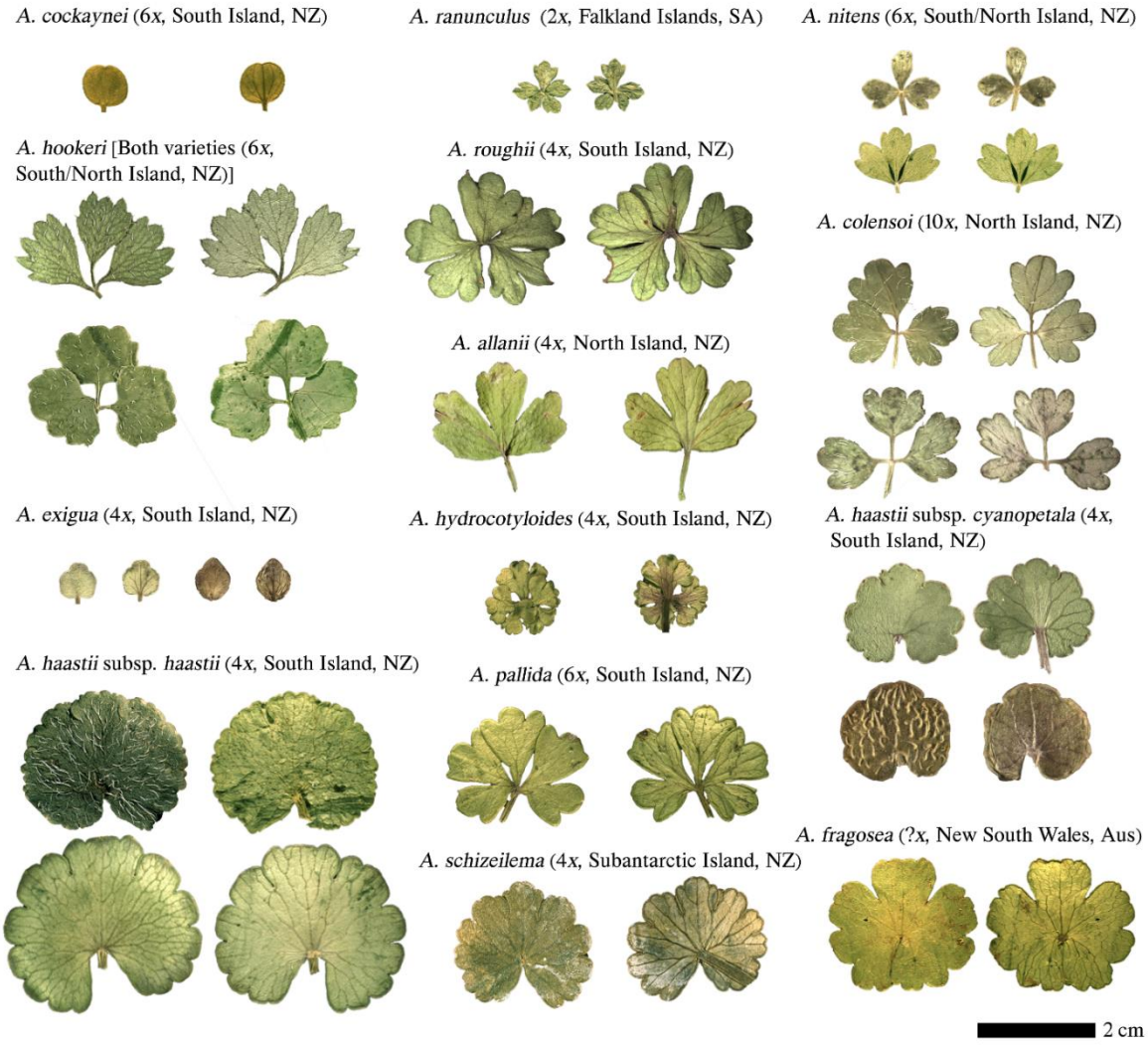


FIGURE 1 Comparison of leaf morphology, ploidal level and geographic distribution of 14 defined *Azorella* species in section *Schizeilema* that are endemic to New Zealand (NZ; including two

undescribed varieties of *A. hookeri*) and Australia (Au; *A. fragosea*; unknown ploidal level), and one South American (SA) relative *A. ranunculus* in section *Ranunculus*.

2.2 Materials and Methods

2.2.1 Hyb-Seq Taxon Sampling

Following the descriptions in Allan (1961) and the taxonomy in Plunkett and Nicolas (2017) (Table S1), we collected multiple accessions of each taxon in *Azorella* sections *Schizeilema* and *Stilbocarpa* from throughout New Zealand (Fig. 2) based on their herbarium specimen distributions (Fig. S2). For each field collection, voucher specimens were deposited at herbaria (WELT or MPN), and leaves from multiple individuals were sampled and stored in silica gel. We also sampled leaves of some New Zealand endemic species from herbarium specimens (CHR, AK, and MPN) to increase their sampling in our study (Table S2). Herbarium specimens were also sampled for three *Azorella* species that occur outside of New Zealand, i.e., *A. fragosea* (Australia; Au) from section *Schizeilema* (CANB), and *A. ranunculus* and *A. burkartii* (South America; SA) from section *Ranunculus* (NYBG). Two individual specimens of a putative new species from New Zealand that does not fit the taxonomy of Allan (1961) or Plunkett and Nicolas (2017) were also sampled (*Azorella* sp.: A.sp_CHR617214 and A.sp_CHR617283). For outgroups, we included the South American species *A. lycopodioides* (NYBG) in section *Glabratae*, which is more distantly related to sections *Schizeilema*, *Stilbocarpa* and *Ranunculus* in a published ITS tree (Plunkett and Nicolas, 2017), and one individual of *Hydrocotyle novae-zelandiae* var. *montana* Kirk (hereafter, *Hydrocotyle*) as a more distant outgroup to *Azorella*.

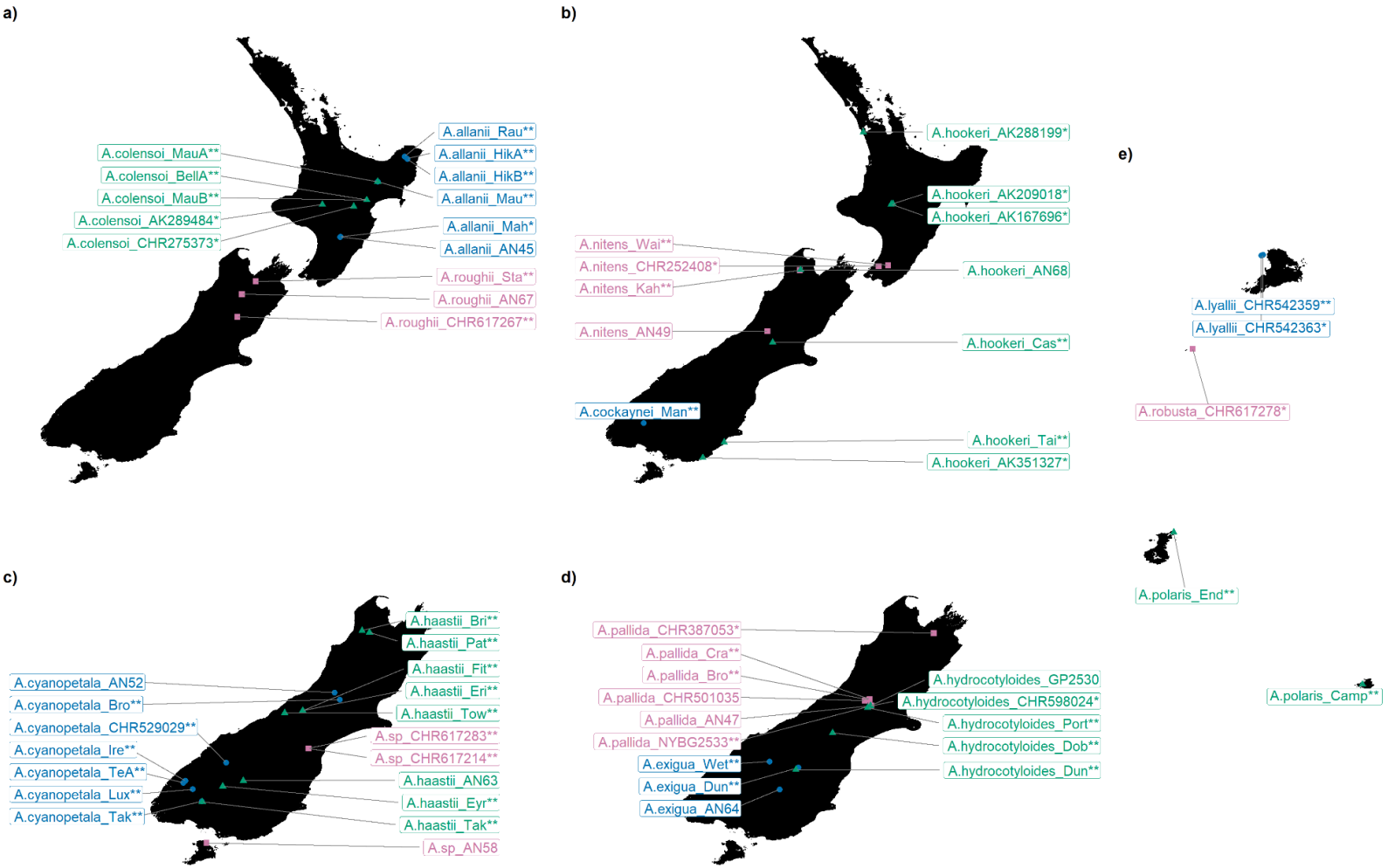


FIGURE 2 Sampling of New Zealand *Azorella* section *Schizeilema* (13 species from a. to d.), and section *Stilbocarpa* (three species in e). Each species is represented by a shape and colour in each subplot. a) *A. allanii* (4*x*), *A. roughii* (4*x*), and *A. colensoi* (10*x*). b) *A. hookeri* (6x), *A. nitens* (6x), and *A. cockaynei* (6x). c) Two South Island endemic tetraploid subspecies *A. haastii* subsp. *cyanopetala* (labelled as "A.cyanopetala") and *A. haastii*

subsp. *haastii* (labelled as "A.haastii"), and two undescribed taxa A.sp_CHR617283/CHR617214 and A.sp_AN58. d) South Island endemic *A. pallida* (6*x*), *A. exigua* (4*x*) and *A. hydrocotyloides* (4*x*). e) *A. lyallii* on Stewart Island, *A. robusta* on the Snare Islands, and *A. polaris* (6*x*) on both Auckland Islands and Campbell Islands. The individuals are labelled with species name and the collection site if field-collected or herbarium specimen accession number/collection number if sampled from a specimen. The individuals or accessions without an asterisk represent samples with only genome-skimming (nuclear ribosomal DNA and plastome DNA), or with only one asterisk (*) represent the only Hyb-Seq (Angiosperms353 single-copy nuclear genes) available. Individuals labelled with two asterisks (**) represent those with data from both genome-skimming and Hyb-Seq available. Note individual sample names start with "A.", do not have a space between the full stop and the unitalicized species name, and then have an underscore and the herbarium accession number or population code; see Table S2 for additional details.

2.2.2 DNA Extraction and Genomic Library Preparation

DNA was extracted from leaf tissue using a standard CTAB method (Doyle and Doyle, 1987) or the DNeasy® Plant Mini Kit (QIAGEN). Extracted DNA integrity levels were checked on 1% agarose gels and concentrations were measured by the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific). Genomic libraries were constructed using the NEBNext® UltraTM II Library Prep kit (New England Biolabs) following the manufacturer's protocol but with half of the recommended volumes for all steps. Initially, 10 to 250 ng input DNA was fragmented into less than 1 kbp (base pairs) pieces using the NEBNext® Ultra II FS Enzyme Mix. Fragmentation time (from 0 to 10 minutes) was adjusted according to the DNA quantity and degradation level of each sample. NEB adaptors were ligated to two-ends of fragmented DNA pieces. For samples with low input DNA (5 ng to 50 ng), the adaptor concentration was diluted 10-fold to avoid dimer-ligation and then cleaned with AMPure XP Beads (Beckman Coulter). By contrast, high input samples (higher than 50 ng) were size selected for inserts between 400 to 500 bp using AMPure XP Beads after adaptor ligation. Eventually, each sample was barcoded with a pair of NEBNext® Multiplex Oligos for Illumina. Genomic libraries were quantified using the Qubit™ dsDNA HS Assay Kit, and the average bp was calculated by the LabChip® GX Touch™ nucleic acid analyzer (PerkinElmer).

2.2.3 Target Enrichment and Genome-Skimming Sequencing

Genomic libraries with less than 10 ng or a profile shorter than 200 bp were excluded from pooling for the hybridization reactions. All DNA libraries of *A. schizeilema*, a subantarctic species in section *Schizeilema*, were filtered out at this stage due to high degradation of herbarium specimen DNA and a lack of fresh material. A total of 125 DNA libraries from 72 populations of 20 taxa (including one undescribed NZ taxon *A.* sp.; Table S2) were divided into four batches for Hyb-Seq with the Angiosperms353 universal baits (Johnson, et al. 2018). Samples with sufficient DNA library concentration were pooled into equimolar amounts in each batch to assure equivalent sequenced reads.

The first two trial Hyb-Seq batches 1 and 2 included a total of 39 individuals. Batch 1 had 12 individuals (including two replicates from the same individual A.roughii_Sta4A and A.roughii_Sta4B) and batch 2 had 17 individuals (including the replicates A.colensoi_AK289484A and A.colensoi_AK289484B; A.hookeri_AK351327A and A.hookeri_AK351327B). We followed the myBaits® Kit Manual V4 to capture the targeted DNA fragments (incubation at 65^{o}C for 20 hours), and the post-captured libraries were then sequenced on two Illumina Miseq™ runs (Massey

Genome Service, New Zealand) to produce 150 bp paired-end reads. The next two sequencing batches 3 and 4 had in total 94 samples using myBaits® Kit Manual V5 to produce the same length reads in one run on an Illumina Hiseq™ 2000 (Novogene, Singapore). The biological replicates of A.haastii_Pat10B and A.haastii_Tak1B were sequenced on one additional MiSeq run for comparison with the Hi-seq sequencing results from the same individuals in batches 3 and 4 (A.haastii_Pat10A and A.haastii_Tak1A).

To recover high copy genes, we reused 92 selected Hyb-Seq DNA libraries to produce 150 bp genome-skimming paired-end reads on one run of Illumina Hiseq™ 2000 (Novogene, Singapore). Additional genome-skimming data from 12 section *Schizeilema* individuals generated independently by G.P. and A.N. (unpubl. data) were also included (Table S2). Another potentially undescribed species from Stewart Island, New Zealand, included in the genome-skimming run as individual A.sp_AN58 does not fit the current taxonomic classification (Allan 1961) and may be distinct from the two individuals of another putative new species mentioned above, A.sp_CHR617214 and A.sp_CHR617283 (Canterbury, New Zealand). In total, we obtained genome-skimming data for 104 *Azorella* individuals collected from 56 sites and representing 19 taxa.

Sequence reads of Hyb-Seq and genome-skimming were trimmed using Trimmomatic v.0.39 (Bolger, et al. 2014) to remove the adaptor sequences and the low-quality bases with the following settings: ILLUMINACLIP:TruSeq2-PE.fa:2:20:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50.

2.2.4 Exon Recovery Rate and Polymorphisms Among Targeted Genes

Recovering the targeted exons from Hyb-Seq data involved three steps in the script 'reads_first.py' using Hybpiper v.1.3.1 (Johnson et al., 2016). First, the sequencing reads were mapped to the Angiosperms353 bait set reference sequences (Johnson et al., 2018) using BWA v. 0.7.17 (Li, 2013), and to improve the exon recovery rate, we used the pipeline 'BYO_transcriptome.py' by McLay et al. (2021) to include extra reference sequences from the order Apiales. Second, mapped reads were sorted to each targeted gene and *de novo* assembled into large contigs with SPAdes v.3.11.1 (Bankevich et al., 2012). Finally, the exons in assembled contigs were extracted and concatenated according to the reference sequences with Exonerate v.2.2.0 (Slater and Birney, 2005). We also calculated the recovery rates of exons by comparing the recovered exon length with the length of references for each gene.

De novo assembled contigs also contain partial introns or flanking regions that can be useful for taxonomic study of closely related species (Johnson et al., 2016; Thomas et al., 2021). We extracted the supercontigs of genes containing exons, introns, and flanking regions using the command 'intronerate.py' in Hybpiper. Paralogs (homeologs for polyploids) among the recovered genes were detected using 'paralogy.py' in Hybpiper. The low divergence levels between homeologous sequences may limit paralog assembly in SPAdes (Johnson et al., 2016; McKain et al., 2018), therefore, we mapped the trimmed reads to the extracted supercontigs of each gene in HybPhaser v.2.0 (Nauheimer et al., 2021), and calculated the proportion of polymorphic sites among each target gene (SNP percentage) for all sequenced samples to reveal the allele divergence among SCNGs.

2.2.5 Single Copy Nuclear Gene (SCNG) Trees and Species Tree Reconstruction

Genealogy reconstructions of each target-enriched gene started by aligning the supercontig sequences with the '--auto' option in MAFFT v.7.429 (Katoh and Standley, 2013). We filtered alignments with fewer than 90 individuals and that were lacking outgroup species after removing the gap sites present in 30% or more of the sequences (-gt 0.7) using trimAl v.1.4 (Capella-Gutiérrez et al., 2009). Gene trees were reconstructed with IQ-TREE2 (Minh et al., 2020b) to automatically detect the best model for each gene via ModelFinder (Kalyaanamoorthy et al., 2017), and 1000 bootstrap (bs) replicates were run (-B 1000). We collapsed the nodes among gene trees with bootstrap values less than 30% using the script 'i & b<=30' in 'nw_ed' (https://github.com/tjunier/newick_utils)(Junier and Zdobnov, 2010) to improve the accuracy of the final species tree (Zhang et al., 2017). Finally, the collapsed gene trees were used to generate the multispecies coalescent tree of 125 individuals in ASTRAL v.5.7.7 (Zhang et al., 2017; Mirarab et al., 2014), and each node was calculated with a local posterior probability.

From the 125-individual ASTRAL tree, which was rooted using *Hydrocotyle* as the outgroup, we filtered out two individuals, one that had a potential misidentification and another that was too distantly related to the ingroup species (see Results). Using the same thresholds as above for the filtered 336 genes, the phylogenies of the remaining 123 individuals were reconstructed using two approaches: the multi-species coalescent model-based approach in ASTRAL with node-collapsed individual gene trees, and the concatenation model in IQ-TREE2 (-p) using the supermatrix of the concatenated gene alignments (see Results).

2.2.6 Gene Tree Concordance Analysis

Gene concordance factors (gcf) measure the proportion of gene trees with support for each node in the species tree (Minh et al., 2020b), which can also be informative about reticulate relationships (Baum, 2007). Each of the 336 node-collapsed gene trees were first rooted with selected outgroup species using 'reroot_trees.py' (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts). The portion of concordant or discordant gene trees that supported each ASTRAL tree node was calculated in PhyParts (https://bitbucket.org/blackrim/phyparts) (Smith et al., 2015). The gcf pie charts were calculated for each node in the species tree via 'phypartspiecharts.py' (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts) and visualised by ggtree package (Yu et al., 2017) in R v.4.0.1 (R Core Development Team, 2013). Similarly, we calculated site concordance factors (scf) that show the number of nucleotide sites among concatenated gene alignments that support the topology of the species tree and the gcf for concatenation models in IQ-TREE2.

2.2.7 Genomic SNP Variation

Intraspecific variation was estimated for four taxa (i.e., *A. allanii*, *A. colensoi*, *A. haastii* subsp. *haastii*, and *A. haastii* subsp. *cyanopetala*) that had more than 10 individuals representing at least three different sites. One individual of each of these four taxa with a high number (>340) of assembled genes and high average exon recovery rate (>80%) was selected as the reference sample (also see Results). We mapped the Hyb-Seq reads of all individuals in each taxon to the supercontigs of the same reference and extracted their SNPs (i.e., joint-called SNPs) using the 'HybSeq-SNP-Extraction' (https://github.com/lindsawi/HybSeq-SNP-Extraction) (Slimp et al., 2021) pipeline. The SNPs were filtered using the following quality thresholds: 'QD < 5.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0' in GATK v.4.1.8.1 (Poplin et al., 2017). We removed the SNP sites with a missing genotype rate higher than 0% (--geno 0) and samples with an overall missing genotype rate higher than 40% (--mind 0.4) in PLINK v.1.90b6.16 (Purcell et al., 2007). Given the potential gaps between stitched exons in reference supercontigs (Johnson et al., 2016), the remaining SNPs were only thinned with a 10 bp window size in PLINK(--indep 10 5 3) to break down the linkage disequilibrium (LD) and to keep more polymorphic sites. The eigenvectors for the first 20 principal components (PCs) were calculated in PLINK, but only the first two highest PCs were plotted in the principal component analysis (PCA) for each species.

2.2.8 Phylogenetic Analyses of nrDNA and Plastome Sequences

High copy markers from the plastome and nrDNA were *de novo* assembled from genome-skimming reads of 104 individuals with the script 'get_organelle_from_reads.py' in GetOrganelle v.1.7.5 (Jin et al., 2020) using default settings. For the 12 samples for which a complete plastome could not be assembled (Table S3), we mapped their largely assembled plastid scaffolds to the reference plastome of *A. fragosea_CANB797887* [annotated using GeSeq v.2.03; (Tillich et al., 2017)] to assist with their consensus sequence extraction in Geneious v.9.1.8. For nrDNA, we only selected the cistron ETS-18S-ITS1-5.8S-ITS2-26S for phylogenetic reconstruction. Notably, ETS was not able to be extracted for two individuals (A.allanii_HikA1 and A.allanii_Mau10). The intergenic spacer (IGS) region was excluded because of the excessive variable sites that were difficult to align (Jin et al., 2020). After aligning the plastome and nrDNA sequences with MAFFT, respectively, the phylogenetic trees of high copy genes were reconstructed using IQ-TREE2 with 1000 bootstrap replicates. We further confirmed the number of polymorphisms and level of homogenization among nrDNA cistrons by mapping the sequenced genome-skimming reads to the extracted nrDNA sequences in PATÉ v.1 (Tiley et al., 2021), which can calculate the proportion of SNP sites and use the paired-end read information to phase the homeologous sequences.

2.2.9 Network Estimation with Gene Trees

To reduce computational running time and minimize conflicting intraspecific signals, one individual was selected to represent each *Azorella* taxon based on its exon recovery rate and whether the selected individual had genome-skimming data available. Additionally, two individuals representing the two different lineages of *A. haastii* subsp. *haastii* and one individual representing each plastome type of *A. pallida* and *A. haastii* subsp. *cyanopetala* were included (see Results). We first reconstructed the ASTRAL tree for 22 individuals and calculated the gcf for each node with filtered and node-collapsed gene trees using the same thresholds as above, except only gene trees that had all 22 individuals were selected to reduce the effect of missing data for gcf calculation. The alignments of filtered genes were concatenated in Geneious and used to visualize the conflicting signals among the concatenated gene alignment via a network approach with the MedianNetwork method in SplitsTree4 (Huson and Bryant, 2006).

Two additional maximum pseudo-likelihood approaches were applied to estimate the reticulation events using the concordance factors among gene trees. Specifically, a list of unrooted quartet concordance factors (qcf) was calculated in SNaQ (Species Networks applying Quartets) as implemented in PhyloNetworks v.0.14.2 (Solís-Lemus et al., 2017). SNaQ used the ASTRAL tree

as a starting tree to test for possible hybridization events (-h) that we set from 0 to 10 in qcf list, and each event had 10 parallel runs. The best model that describes the number of hybridization events was selected based on the lowest negative log pseudo-likelihood value plotted in R. Additionally, the rooted triplets of gene trees were extracted by the 'InferNetwork_MPL' model in PhyloNet v.3.6.9 (Than et al., 2008). We used the default setting to simulate hybridization events from 0 to 9, and each event had 5 parallel runs. The lowest log likelihood ratio of each generated model was applied to select the best model with the corrected Akaike information criterion (AICc) (Sugiura, 1978) and Bayesian information criterion (BIC) (Schwarz, 1978), which used the number of estimated hybridization events and simulated branch number as the number of parameters for model selection. Finally, we also performed a Tree Incongruence Checking (TICR) test with 'test.one.species.tree' function from PHYLOLM package (Tung Ho and Ané, 2014) in R to confirm if the discordance among gene trees could be explained by only incomplete lineage sorting (ILS) without reticulation events under a coalescent model.

The *de novo* assembled nrDNA from genome-skimming sequenced reads had complete and continuous sequencing structure for homeolog extraction. We used the cistron of extracted nrDNA as a reference to phase homeologs (i.e., assemble the homeologs correctly) for the selected 22 individuals via PATÉ and to confirm the networks in nrDNA data (see Results). The nrDNA multi-labelled (MUL) tree was constructed in IQ-TREE2 and the network was visualized in Dendroscope V3.7.6 (Huson and Scornavacca 2012) using a cluster-based method.

2.2.10 Detection of Genomic Introgression

Patterson's D-statistic test, also known as the ABBA-BABA test (Patterson, et al. 2012; Malinsky, et al. 2018), was used to identify genomic introgression signals within Hyb-Seq SNP data. To extract SNPs for all 22 selected individuals, we assigned the supercontigs of the outgroup species *A. lycopodioides* as the reference sequences for joint-calling SNPs with the pipeline 'HybSeq-SNP-Extraction'. Considering two Miseq samples among the 22 individuals (A.ranunculus_NYBG2447 and A.robusta_CHR617278; see Results) may have lower sequencing depth and therefore more missing data, the SNPs were filtered by bcftools (Li, 2013) to keep only biallelic sites allowing 20% missing data ('F_MISSING > 0.2' -m2 -M2). Similarly, we used VCFtools (Danecek et al., 2011) to select independent SNPs with only a window size of 10 bp (--thin 10) to maintain more polymorphic sites (see Results). Dsuite (Malinsky et al., 2021) was applied to calculate the D-statistics and the related estimate of f4-ratio (admixture fraction f) from all possible species trios among filtered SNPs. We then used the f4-ratios to calculate the f-branch

metric (Fbranch) (Malinsky et al., 2018a), which can assign gene flow to specific nodes or tips of the previously constructed ASTRAL 22-taxon tree.

2.2.11 Bayesian Inference of Species Relationships and Network with SNPs

A phylogenetic tree was reconstructed with the multi-species coalescent model-based approach SVDQuartets (Chifman and Kubatko, 2014) using the joint-called SNPs of 22 individuals as input. The SNPs were filtered using bcftools to keep only biallelic sites without missing data (‘'F_MISSING>0' -m2 -M2’) and thinned in VCFtools (--thin 10). The remaining SNPs were converted into NEXUS format in vcfphitools v2.0 (Ortiz, 2019) and the species tree estimated with 1000 bootstrap replicates via SVDQuartets, which is implemented in PAUP* v.4.0a (Swofford and Sullivan, 2003).

The phylogeny and divergence times of 22 individuals were estimated using filtered SNPs and the Bayesian based multi-species coalescent approach of SNAPP [SNP and AFLP Package for Phylogenetic analysis; (Bryant, et al. 2012)] implemented in BEAST2 (Bouckaert et al., 2014). To generate the input format for the SNAPP analysis, which considers each SNP as an independent marker, all homozygous sites among filtered biallelic SNP sites were converted into ‘0’ or ‘2’ at random and all the heterozygous sites were assigned as ‘1’ by ‘snapp_prep.rb’ (https://github.com/mmatschiner/snapp_prep). The Markov Chain Monte Carlo (MCMC) chain length was specified to 5,000,000 and the output trees were saved every 250 iterations.

Because *Azorella* has no fossil data for molecular dating calibration of our datasets, we extracted the previous estimated divergence times between *A. lycopodioides* with sections *Schizeilema*, *Stilbocarpa*, *Ranunculus*, *Huanaca* and *Azorella* [28.723 Ma (million years ago) with 95% HPD (highest posterior density) between 19.99 and 37.766 Ma] from the time-calibrated Apiales tree (Nicolas and Plunkett 2014) as the crown age for our SNAPP trees. Eventually, we calculated the final effective sample size (ESS) for Bayesian posterior distribution in Tracer (Rambaut, et al. 2018), and generated the final consensus SNAPP tree after specifying the burn-in percentage to 10% and setting the mean heights as node heights in TreeAnnotator (Drummond and Rambaut, 2007). The concordance level between SNAPP trees and consensus tree was visualized in DensiTree (Bouckaert, 2010).

Reticulation signals among filtered biallelic SNPs were analysed by the SNAPP-based Bayesian approach ‘MCMC_BiMarkers’ (Zhu, et al. 2018) as implemented in PhyloNet. We specified the MCMC chain length to 1,500,000 and the burn-in period as 200,000, and saved the output result every 500 iterations. The maximum number of hybridization events was assigned to eight

(according to the PhyloNet result; see Results). We selected the network with the highest maximum a posteriori (MAP) to represent the reticulate relationships among the SNP network of 22 individuals.

2.2.12 Divergence Times and Biogeographic History of New Zealand Azorella

The biogeographic history of New Zealand *Azorella* was analysed using the BioGeoBEARS package (Matzke, 2013) in R. BioGeoBEARS requires a bifurcating phylogenetic tree and the distribution ranges of each taxon for ancestral range modelling. After comparing the concordance levels of SNAPP trees and the SNAPP network result for 22 individuals, we excluded the individual of decaploid *A. colensoi* that affected the topology in the bifurcating tree (see Results). For the remaining 21 individuals, we used the reselected SNPs and generated their consensus SNAPP tree in BEAST2 with the same thresholds as above. According to the known distribution of selected *Azorella* species in Fig. S2, we defined six geographically separated regions to determine the possible long distance dispersal events from South America to New Zealand, as well as to estimate the species colonization histories within New Zealand, including South America (A), Subantarctic Islands, New Zealand (B), Australia (C), North Island, New Zealand (D), South Island, New Zealand (E) and Stewart Island, New Zealand (F). The maximum number of areas in which a species could occur was set to two. After comparing the result of six models (DEC, DEC+J, DIVALIKE, DIVALIKE+J, BAYAREALIKE, and BAYAREALIKE+J) in BioGeoBEARS, we selected the model with the highest AICc weight (AICc_wt) value to infer the biogeographical history of the sampled *Azorella* species.

2.3 Results

2.3.1 Hyb-Seq Sequencing Results

A comparison of the Hyb-Seq results from all 125 samples of *Azorella* and outgroups based on the type of leaf material (field-collected, silica-dried samples vs. herbarium specimens) and sequencing platform (MiSeq vs. HiSeq) is shown in Table 1. Among the 353 target-enriched genes, the number of genes from each sample with assembled exons varied from 119 to 349. The average exon recovery rate for these genes ranged from 28.99% to 88.02% (Table S3). For all 125 samples, gene 6514 had no exons assembled for any of the sequenced individuals.

On average, the 94 individuals sequenced by HiSeq generated around five times the number of reads for each sample compared to 31 individuals sequenced by MiSeq, which also led to ten times more mapped reads and a two-fold increase in mapped percentage (Table 1). The HiSeq sequenced

samples had on average 54 more genes with exons assembled and each gene was on average 179 bp longer than the MiSeq samples. Further, the captured read differences between two sequencing platforms caused a nearly 1.4-fold increase in exon recovery rate and 1.8-fold extracted supercontigs length in the HiSeq data. Two replicate samples (A.haastii_Pat10A/B and A.haastii_Tak1A/B) were sequenced by both HiSeq and MiSeq and showed a similar trend (Table S3). The source of the samples (field-collected vs. herbarium specimen) did not affect sequencing outcomes using the HiSeq platform, which underscores the flexibility and utility of the Hyb-Seq method, such that it is possible to get a sufficient number of targeted genes when sequencing herbarium specimen material.

TABLE 1 Comparison of Hyb-Seq results for 125 samples of *Azorella* and outgroups based on the type of leaf material extracted and sequenced (field-collected and silica-dried [F] or herbarium specimens [S]) and sequencing platform (HiSeq or MiSeq). For all the samples in each group, the mean mapped percentage represents the ratio between the mean of mapped reads and the mean of sequenced reads. The average number of genes with exons assembled and their mean exon recovery rates were calculated by comparison to the reference sequences. The length (bp) of exons and supercontigs were averaged for each group.

Group	No. Samples	Mean No. Sequenced Reads	Mean No. Mapped Reads	Mean Mapped Percentage (%)	Mean No. Genes with Exon	Mean Exon Recovery (%)	Average Exon Length (bp)	Average Supercontigs Length (bp)
MiSeq_S	16	1,255,242	294,061	24.28%	263	50.94%	412	1016
MiSeq_F	15	1,767,826	204,928	14.17%	317	66.48%	537	1538
HiSeq_S	11	7,566,882	2,680,628	36.94%	344	81.55%	647	2314
HiSeq_F	83	8,145,887	2,978,729	36.80%	344	82.97%	661	2465
Total Miseq	31	1,511,534	249,495	19.23%	290	58.71%	475	1,277
Total Hiseq	94	7,856,385	2,829,679	36.87%	344	82.26%	654	2,390

2.3.2 Phylogeny, Concordance and Allele Divergence of Single Copy Nuclear Genes (SCNGs)

After removing genes with a large amount of missing data, the gene trees of each of the 337 SCNGs were selected to reconstruct the ASTRAL tree for 125 individuals (Fig. S3). Each trimmed gene alignment had on average 1.7 kbp aligned sites and 435 informative sites. An initial sample filtering was performed based on the number of assembled genes, the average exon recovery rate of assembled genes, and the topology of the 125-individual ASTRAL tree.

Three herbarium specimen samples sequenced on MiSeq had fewer than 200 targeted genes with assembled exons, and each gene was 40% shorter than the reference gene length, including A.pallida_CHR387053 (119 genes with 28.99%), A.fragosea_CANB798456 (128 genes with 32.58%), and A.nitens_CHR252408 (179 genes with 33.66%). Of these, in the 125-individual ASTRAL tree (Fig. S3), only A.pallida_CHR387053 exhibited a different placement from all other sampled *A. pallida* individuals, and it was also sampled from a different locality (Nelson, South Island, New Zealand; Fig. 2). The result indicates that A.pallida_CHR387053 might be misidentified or its placement is an artifact of the low exon recovery rate for this sample. The tree also confirmed a more suitable outgroup species of *A. lycopodioides* for the New Zealand *Azorella* species. Therefore, the two individuals, *Hydrocotyle* and A.pallida_CHR387053, were removed for all downstream analyses.

The remaining 123 individuals represent 19 *Azorella* taxa; each taxon on average had 324 targeted genes assembled and each gene had on average around 2 kbp of supercontig length recovered (Table 2). In total, we selected 336 filtered genes for ASTRAL tree reconstruction and gene concordance factors (gcf) analysis (Fig. 3). For ease of discussion, we identified five main groups in the 123-individual ASTRAL tree based on their geographical distribution and phylogenetic relationships: 1) and 2) two mainland New Zealand groups 1 & 2 (NZ1 & NZ2), respectively, both with six taxa from section *Schizeilema* each; 3) *A. fragosea* from Australia (Au); 4) two species in section *Ranunculus* from South America (SA); and 5) the three species of section *Stilbocarpa* from the New Zealand subantarctic islands (Sub) (Fig. 3).

In Fig. 3, within the monophyletic section *Schizeilema,* Au emerged as sister to a monophyletic NZ2, while NZ1 was reconstructed as paraphyletic. The SA clade was sister to the clade of section *Schizeilema* (NZ1, Au, NZ2), with the three species of Sub reconstructed as monophyletic and sister to this larger group (SA, NZ1, Au, NZ2). Although the relationships of the five groups were all supported by high local posterior probabilities, the gene concordance analysis showed the species relationships within and between the two NZ groups and Au contained substantial gene conflict. A similar topology and gcf level were observed in the tree derived from the concatenated supermatrix

in IQTREE-2 (Fig. S4), however, the species relationships within the NZ2 clade were not congruent in the ASTRAL and IQTREE-2 trees (i.e., compare Fig. 3 and Fig. S4).

Within NZ1, both approaches (Fig. 3; Fig. S4) supported a clade comprising two tetraploid species, *A. allanii* and *A. roughii*, and a paraphyletic group of higher polyploids, including *A. hookeri* (6*x*), *A. nitens* (6*x*), *A. cockaynei* (6*x*) and *A. colensoi* (10*x*). Within NZ2, there are five monophyletic taxa shown in the ASTRAL tree: *A. exigua* (4*x*), *A. haastii*. subsp. *cyanopetala* (4*x*, labelled as A.cyanopetala), *A.* sp. (?*x*; A.sp_CHR617214 and A.sp_CHR617283), *A. hydrocotyloides* (4*x*), and *A. pallida* (6*x*), which were also supported by the concatenated supermatrix analysis in IQTREE-2. The exception was for A.pallida_CHR501035, which had 259 assembled genes with 54.68% exon recovery rate and was paraphyletic to *A. fragosea* (Fig. S4). Furthermore, *A. haastii* subsp. *haastii* (4*x*, labelled as A.haastii) in NZ2 comprised three different lineages in both analyses that were more closely related to *A. pallida*, *A.* sp. and *A. hydrocotyloides* than to each other.

Samples with a low percentage of target-captured reads will decrease the exon recovery rates and homeolog detection for each species. Therefore, we filtered out 13 samples with fewer than 300 assembled genes and 60% average exon recovered rate before calculating the allele divergence (proportion of SNPs among supercontigs) and the number of genes with homeologs for each species (Table 2; Fig. 3). Three levels of mean allele divergence were observed for the filtered 113 individuals representing 18 taxa: 1) c. 1% allele divergence for the outgroup and SA species; 2) c. 2.5% for all the tetraploid species in NZ2 (including hexaploid *A. pallida*) and the two tetraploids in NZ1; 3) c. 3-4% for the higher polyploids (6*x* and 10*x*) in NZ1 and Sub. In comparison to the large number of genes with more than 2% allele divergence in each species, HybPiper only assembled on average 4 to 29 genes with homeologs in these species (Table 2).

TABLE 2 Summary of the species sampling and targeted gene recovery rates by HybPiper and HybPhaser for 123 individuals of *Azorella*. The number of included individuals and sampled sites are listed in the first two columns, respectively. Species are organized according to the groups based on the ASTRAL tree topology in Fig. 3. Ploidal levels were estimated from chromosome numbers in the literature (Table S1). The number of assembled genes, the recovered gene percentage, and the exon or supercontigs length were averaged across all the individuals for each species. The mean allele divergence, number of genes with allele divergence of more than 2% or homeologs (i.e., paralogs warning via HybPiper; Table S3) were calculated for 110 filtered individuals with more than 300 assembled genes and over 60% exon recovery rates.

Species	**No. Sample**	**No. Sites**	**Group**	**Ploidal level**	**Mean No. Assembled Genes**	**Mean Exon Recovery Rate (%)**	**Mean Exon Length (bp)**	**Mean Supercontig Length (bp)**	**No. Filtered Samples**	**Mean Allele Divergence (%)**	**Mean No. Genes with Allele Divergence > 2%**	**Mean No. Genes with homeologs**
Azorella allanii	15	5	NZ1	4*x*	336	78.59%	629	2198	14	2.41%	194	14
A. roughii	7	2	NZ1	4*x*	328	76.16%	613	2106	6	2.38%	186	10
A. cockaynei	3	1	NZ1	6*x*	344	82.81%	660	2538	3	4.44%	308	21
A. hookeri	7	6	NZ1	6*x*	316	65.10%	519	1536	5	3.12%	236	9
A. nitens	5	3	NZ1	6*x*	310	71.54%	572	1973	4	4.32%	303	24
A. colensoi	12	5	NZ1	10*x*	335	75.87%	605	2057	10	4.20%	308	12
A. fragosea	3	3	Au	?*x*	270	62.13%	498	1401	2	1.97%	149	6
A. haastii subsp. *cyanopetala*	16	6	NZ2	4*x*	344	85.06%	677	2654	16	2.50%	198	11
A. haastii subsp. *haastii*	19	7	NZ2	4*x*	339	79.85%	638	2295	19	2.23%	176	7
A. exigua	6	2	NZ2	4*x*	342	81.48%	650	2352	6	2.38%	186	9
A. hydrocotyloides	7	4	NZ2	4*x*	334	79.58%	636	2282	6	2.36%	190	12
A. pallida	8	4	NZ2	6*x*	333	79.98%	639	2286	7	2.54%	201	8
A. sp	2	1	NZ2	?*x*	344	81.75%	648	2227	2	2.10%	173	9
A. lyallii	2	2	Sub	?*x*	303	62.89%	496	1494	1	3.98%	261	29
A. polaris	6	2	Sub	?*x*	329	69.78%	557	1755	6	3.18%	211	15
A. robusta	1	1	Sub	6*x*	307	58.76%	470	1228	NA	NA	NA	NA
A. burkartii	2	2	SA	?*x*	345	80.43%	636	2214	2	1.07%	59	4
A. ranunculus	1	1	SA	2*x*	249	45.80%	377	819	NA	NA	NA	NA
A. lycopodioides	1	1	Outgroup	2*x*	349	86.73%	685	2877	1	0.87%	48	9

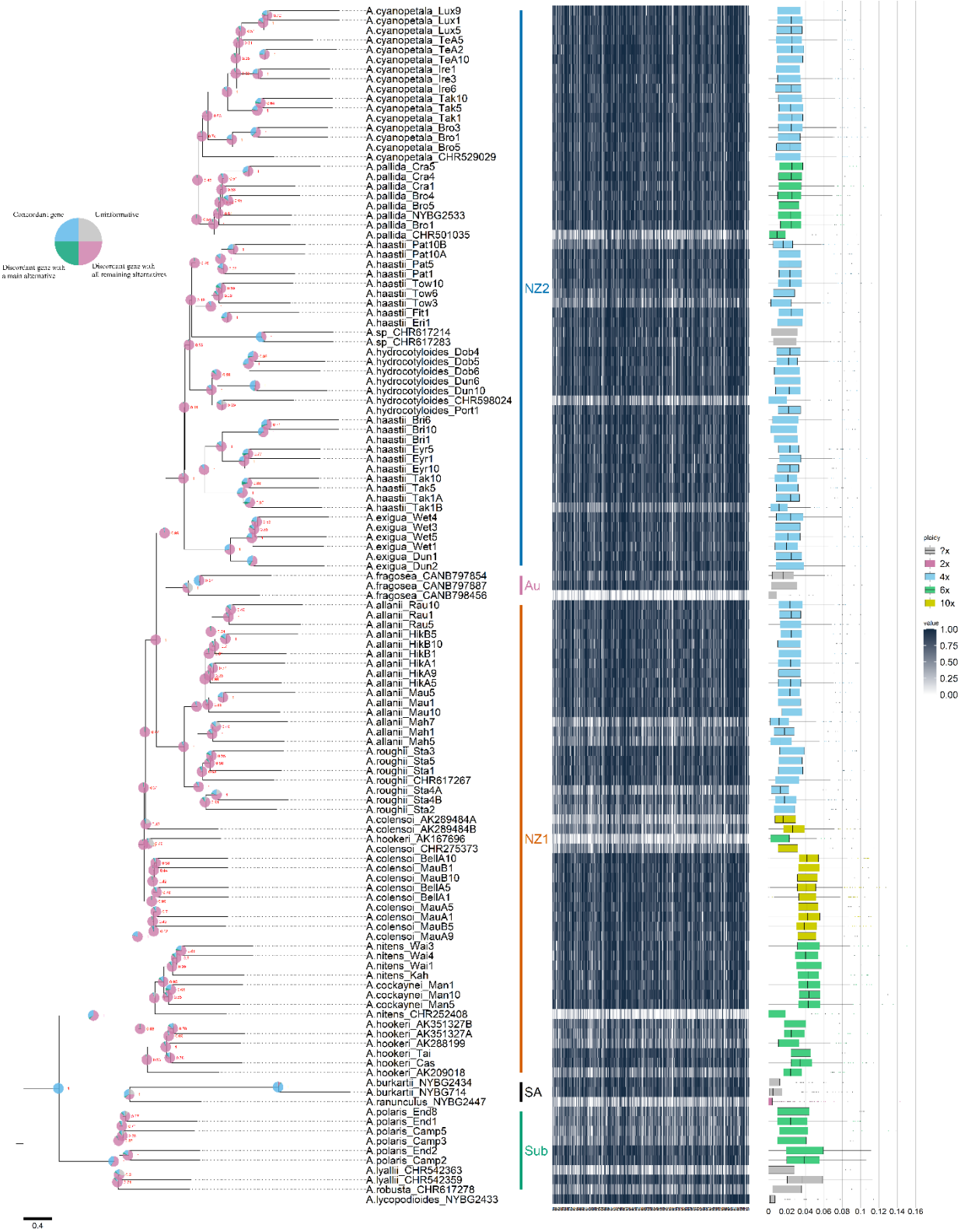


FIGURE 3 A multispecies coalescent phylogenomic ASTRAL tree of 123 *Azorella* individuals using 336 single copy nuclear genes captured using Angiosperms353 baits. All the terminal branch length was set to 1 in ASTRAL tree. The phylogeny shows five groups: New Zealand 1 (NZ1), New Zealand 2 (NZ2), Australia (Au), South America (SA) and Subantarctic islands (Sub). The individual sample names represent the species name and the sampling sources, i.e., the name of the field site or herbarium specimen accession number, plus the individual number of each field

collection (Table S2). Five individuals with biological replicates were annotated as A and B at the end of each sample name (e.g., the replicate pair A.haastii_Pat10A and A.haastii_Pat10B). Each node is supported by a local posterior probability (maximum = 1) and a gene concordant pie chart (blue = concordant, green = discordant with a main alternative, pink = discordant with all remaining alternatives, and grey = uninformative). For each individual, the exon recovery rates of the targeted 353 genes are shown in the heatmap with colour gradient from 0 (white) to 100% (dark blue). Each column of the heatmap represents one targeted gene and the y-axis is correlated with the tree tips. Boxplots show the allele divergence (0 to 16%) for the corresponding tip and the color represents ploidal level [?*x* = unknown (grey), 2*x* (pink), 4*x* (blue), 6*x* (green) and 10*x* (yellow)] of each species, based on published chromosome numbers (Table S1).

2.3.3 Phylogeny and Variation among High Copy nrDNA and Plastome Markers

High copy markers of whole plastome and nrDNA were *de novo* assembled for 104 individuals of 18 *Azorella* taxa (22 individuals) via genome-skimming data. The nrDNA cistron had a recovered length of (on average) 6,960 bp of which 674 sites were parsimony-informative , while the whole plastome was 153,306 bp with 3,578 parsimony-informative sites. The homogenization levels among biparentally inherited nrDNA cistrons were confirmed by calculating the allele divergence from genome-skimming sequenced reads using PATÉ. Especially for the 97 sequenced individuals from *Azorella* section *Schizeilema*, on average, they only had 0.12% allele divergence, in other words, there were on average only nine DNA sites out of 7 kbp nrDNA cistron sequences that were heterozygous. By contrast, the intergenic spacer (IGS) region of nrDNA exhibited a high level of intraspecific divergence for populations within the following three species: *A. haastii* subsp. *cyanopetala* (A.cyanopetala_Lux and A.cyanopetala_Tak), *A. haastii* subsp. *haastii* (A.haastii_Pat and A.haastii_Tow), and the undescribed *A. sp* (A.sp_CHR617214 and A.sp_CHR617283), which contained 8 to 27 assembled IGS copies (Table S4).

The plastome tree inferred using IQ-TREE2 (Fig. 4a) showed different relationships relative to the SCNG ASTRAL tree (Fig. 3). The megaherbs of section *Stilbocarpa* (Sub) formed a clade sister to the SA taxa. This clade was in turn sister to a clade comprising New Zealand taxa with *A. fragosea* (Au) embedded within it. The species in NZ2 were split into two groups that we referred to as NZ2_cp1 and NZ2_cp2 (Fig. 4a). Notably, several taxa were not resolved as monophyletic in these clades, including *A. haastii* subsp. *cyanopetala*, *A. haastii* subsp. *haastii*, and *A. pallida*, which were split across the two NZ2 clades. Within the NZ2_cp2 or NZ2_cp1 clades, no species were reconstructed as monophyletic except *A. exigua*. The plastomes of the NZ1 clade showed similar patterns, with most species being reconstructed as non-monophyletic, except *A. cockaynei* and *A. nitens*, which were reciprocally monophyletic and sister to one another. Otherwise, *A.*

hookeri and *A. colensoi* formed a clade in NZ1, within the paraphyletic group of *A. allanii* and *A. roughii*.

In several cases, the structure in the plastome tree reflected geographical relationships, in particular for *A. pallida* and related taxa, rather than taxonomy (Fig. 4a; S7). For example, within NZ1_cp1, plastome variation was mixed between populations in *A. pallida* (NYBG2533, AN47 and three individuals in Cra) and *A. hydrocotyloides* (GP2530 and Port1), which were sampled from the same region (Canterbury, South Island, New Zealand; Fig. 2). Within NZ1_cp2, the populations of *A. haastii* subsp. *cyanopetala* (A.cyanopetala_Bro) and *A. pallida* (A.pallida_Bro) both collected at the same site (Broken River Ski Area, Canterbury; Fig. 2) had nearly identical plastome sequences (Fig. 4a).

The nrDNA IQ-TREE2 tree (Fig. 4b), on the other hand, displayed a different topology compared to the plastome tree (Fig. 4a) as well as the SCNGs ASTRAL tree (Fig. 3). The six species previously reconstructed in a single clade NZ1 (SCNGs and plastome) were reconstructed in two different clades in the nrDNA tree: one clade included two tetraploids, *A. allanii* and *A. roughii* (NZ1_nr1), and the other included the four higher polyploids, *A. cockaynei* (6*x*), *A. nitens* (6*x*), *A. hookeri* (6*x*) and *A. colensoi* (10*x*) (NZ1_nr2). The SA taxa were reconstructed as sister to NZ1_nr2, while the Sub taxa were sister to a clade comprising Au and all other New Zealand species (NZ1_nr1 and NZ2). In the nrDNA tree (Fig. 4b), congruent with the SCNG tree (Fig. 3), NZ2 was reconstructed as a clade, again with most species not monophyletic, except *A. exigua*.

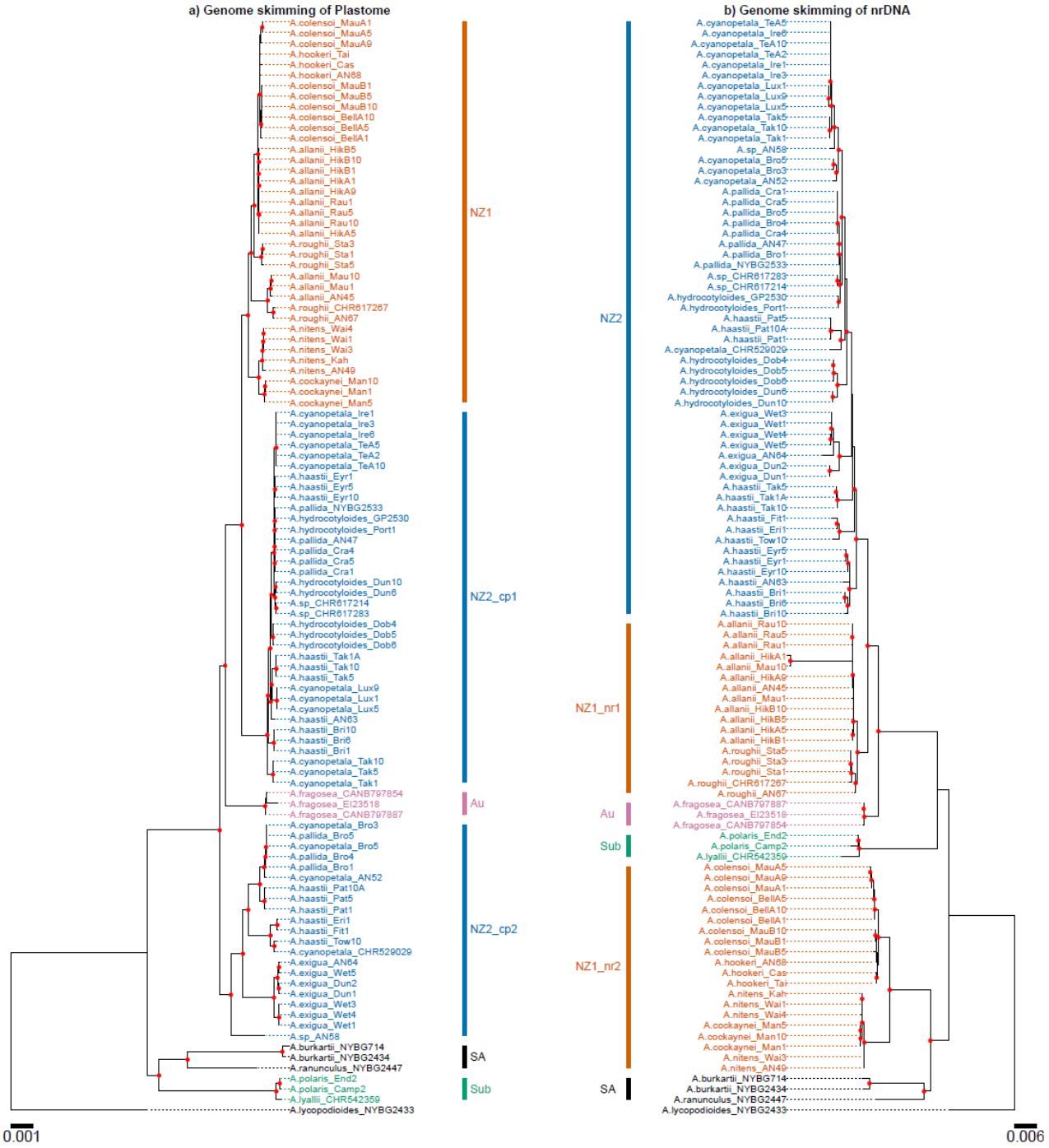


FIGURE 4 Comparison of phylogenetic relationships derived from a) plastome and b) nrDNA data for 104 samples of *Azorella* sections *Schizeilema*, *Stilbocarpa* and *Ranunculus*. Nodes with bootstrap support values higher than 90% are indicated with red dots. The individual names represent the species name and accession ID in Table S2.

2.3.4 Genomic SNP Variation for SCNGs

Intraspecific variation in four taxa from *Azorella* section *Schizeilema* with more than 10 individuals sampled in the SCNG tree, including the monophyletic taxa *A. allanii* and *A. haastii*

subsp. *cyanopetala* and the paraphyletic taxa *A. colensoi* and *A. haastii* subsp. *haastii* (Fig. 3), were analysed by comparing the genomic SNP variants (Table 2). For each taxon, the joint-called SNPs were extracted using the supercontigs of the following individuals as references, i.e., individuals A.allanii_HikB1 (345 genes with 81.70% average exon recovery rate), A. haastii_Eyr5 (345 genes with 82.49%), A.cyanopetala_TeA2 (344 genes with 88.02%), and A.colensoi_MauB5 (347 genes with 81.59%). The number of extracted SNPs for each species was 32,000 (*A. allanii*), 44,201 (*A. colensoi*), 57,021(*A. haastii* subsp. *cyanopetala*) and 64,335 (*A. haastii* subsp. *haastii*). Two individuals of *A. colensoi* sampled from herbarium collections (A.colensoi_CHR275373 and A.colensoi_AK289584A) were filtered out by PLINK because of the large number of missing SNPs. For the remaining samples in each species, only 12% to 25% extracted SNPs were selected for PCA calculation (Fig. 5). The first two major PCs (PC1 & PC2) divided all four species into three groups (Fig. 5), which were related to the geographic distribution of the sampled individuals (Fig. 5a; Fig. 5b; Fig. 5c), except for *A. haastii* subsp. *haastii* (Fig. 5d). The individuals of *A. haastii* subsp. *haastii* clustered in the same three groups as in the SCNG ASTRAL tree (Fig. 3).

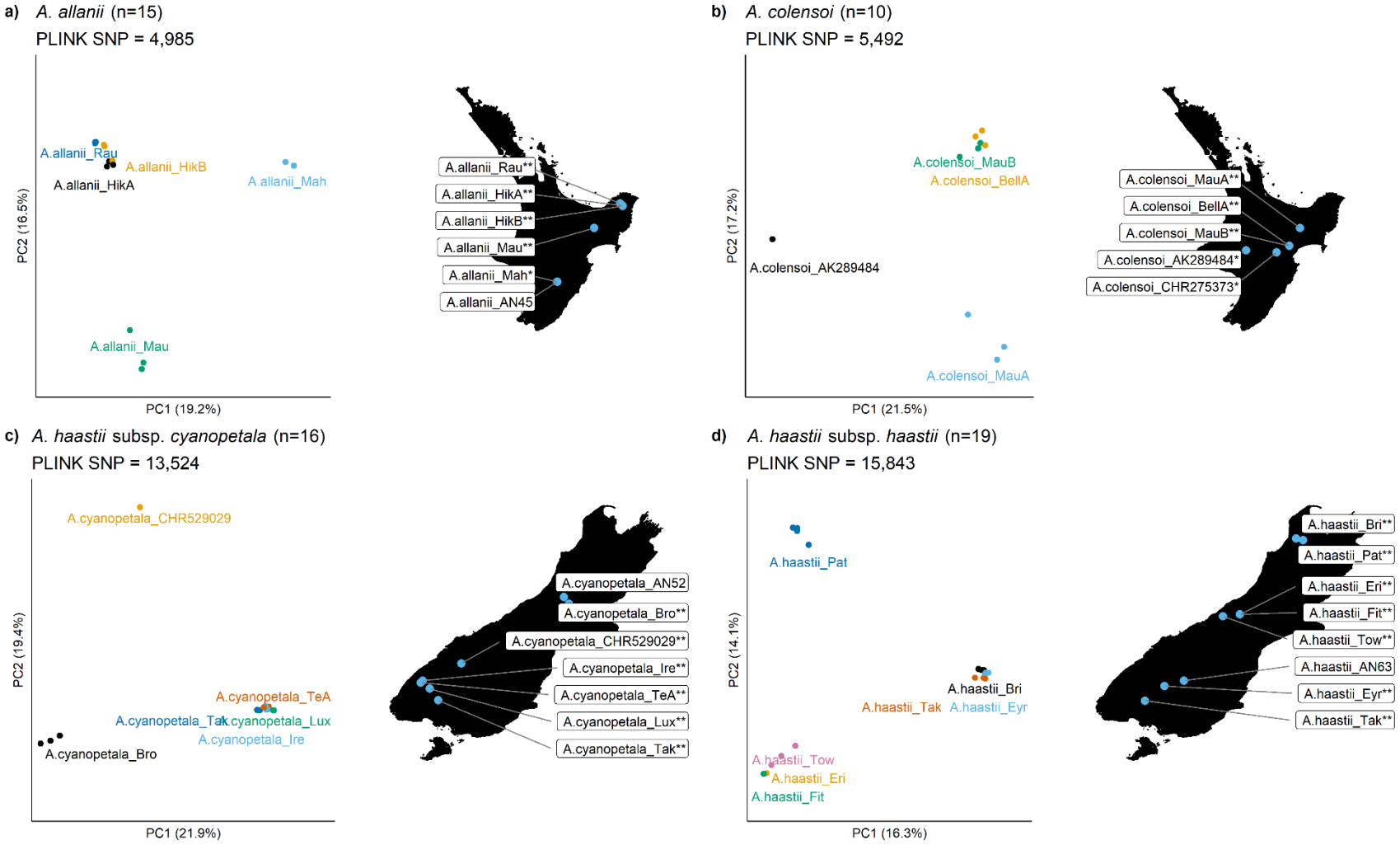


FIGURE 5 PCA plot of intraspecific genetic variation and correlation to geographical distributions of a) *A. allanii*, b) *A. colensoi*, c) *A. haastii* subsp. *cyanopetala* (as "A.cyanopetala"), or subspecies variation within species d) *A. haastii* subsp. *haastii* (as "A.haastii"). These four species had at least ten individuals sequenced with the Angiopserms353 baits. Within each subplot, each dot represents one individual and all the individuals collected from

the same population are labelled in the same colour with their accession numbers annotated in the same colour. The distributions of the samples are shown in the map with the same sample names shown in Fig. 2.

2.3.5 Analysis of Reticulation in Azorella Species

Based on the number of assembled genes and the exon recovery rate, one individual per taxon was selected to reveal reticulation among *Azorella* species. We included additional individuals for species that had more than one lineage in the SCNG tree (Fig. 3; Fig. 5d) or with more than one chloroplast type (Fig. 4a), i.e., three individuals of *A. haastii* subsp. *haastii* from each of the three polyphyletic groups (Fig. 3; Fig. 5d), and two individuals each of *A. pallida* and *A. haastii* subsp. *cyanopetala* to represent their two plastome types (Fig. 4a). We first reconstructed the bifurcating phylogeny of the selected individuals and calculated the gcf among gene trees, i.e., an ASTRAL tree of 22 individuals representing 18 *Azorella* taxa was generated from 225 selected SCNG trees, in which each gene tree contained all 22 individuals and had around 2 kbp supercontig length and averaged 207 informative sites (Fig. 6a).

The topology of the 22-individual ASTRAL tree (Fig. 6a) showed similar species relationships as the 123-individual ASTRAL tree (Fig. 3), including *A. haastii* subsp. *haastii* which consistently exhibited the same structure of three separate lineages. By excluding the conflicting intraspecific gene signals and reducing the missing data among targeted genes, the ASTRAL tree of 22 individuals (Fig. 6a) showed improvements of gene concordance levels in NZ1, whereas the high discordance levels of NZ2 and Au remained high, except for the two individuals of *A. pallida*. Notably, three nodes in Fig. 6a had a high portion of genes that supported a main alternative topology, as can be seen by their pie charts that are nearly one-quarter green, i.e., A.colensoi_BellA10, A.hookeri_Tai, and the node leading to A.lyallii_CHR542359 and A.polaris_Camp2. This indicates alternative species relationships at these nodes should be considered.

Network relationships estimated by SplitsTree used the concatenated 225 gene alignments of the selected 22 individuals. We removed the outgroup species *A. lycopodioides* due to the long branch length in the SplitsTree result (Fig. 6b). The remaining samples displayed four clear previously defined groups: Sub, SA, NZ1 and NZ2, while the Au species *A. fragosea* was nested within NZ1. Overall, the network showed conflicting phylogenetic signals in NZ1, especially between A.hookeri_Tai and A.colensoi_BellA10 (as indicated by the box-like relationships among species), as well as in NZ2, which had short and reticulate branches leading to each species. Conflicting signal was also detected among the subantarctic taxa of section *Stilbocarpa* (Sub).

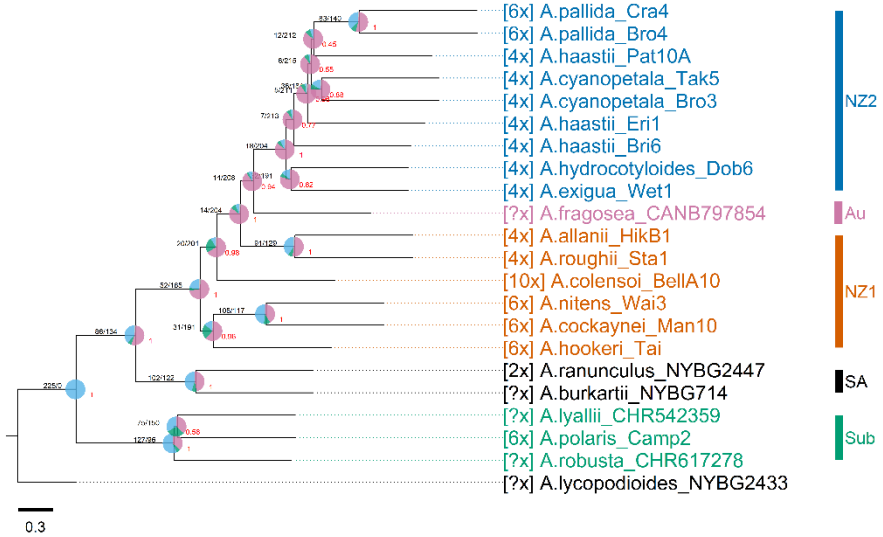
The SNaQ network used the topology of the 22-individual ASTRAL tree (Fig. 6a) as a starting tree to search for reticulation events that could explain the quartet concordance factors among the 225 gene trees. The network scores calculated by SNaQ suggested that there were most likely six

hybridization events (h =6; network score = 3876.26; Fig. S6) in the representative 22 individuals (Fig. 6c). By contrast, PhyloNet showed the estimated reticulation models of the same 225 gene trees had the lowest AIC and BIC values when the number of hybridization events equalled eight (h = 8; Table S7). Furthermore, although incomplete lineage sorting (ILS) could also contribute to the discordance of gene trees, the TICR result rejected the ILS-only model and recommended that reticulation events should be considered (X^2 = 57.61214, P value = 1.902093×10^{-12}).

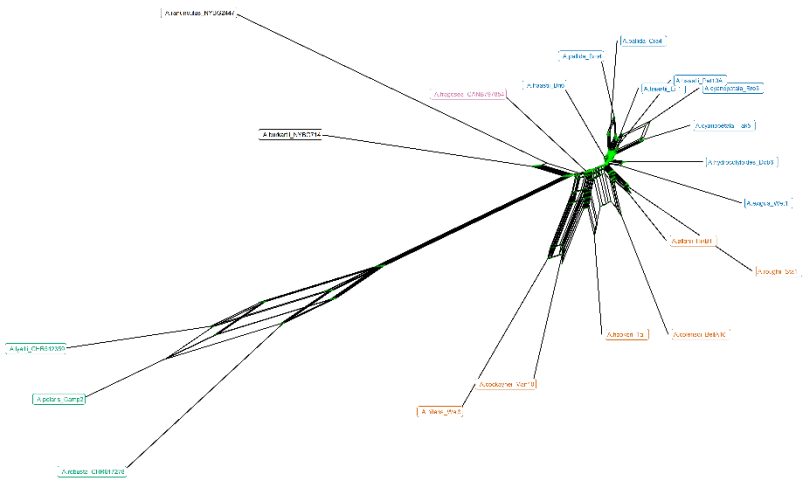
The SNaQ and PhyloNet network results showed complex reticulate evolutionary histories among the five groups (Fig. 6c; Fig. 6d). Specifically, the SNaQ and PhyloNet results both suggested a hybrid origin for A.ranunculus_NYBG2447 (2x; SA) that may involve A.bukartii_NYBG714 (SA) and another ancestor. The Sub group exhibited hybridization signals from the outgroup A.lycopodioides_NYBG2433 (Fig. 6c; Fig. 6d) and the A.allanii_HikB1 lineage (Fig 6d). The topologies of both networks divided NZ1 into a group of two tetraploids (A.allanii_HikB1 and A.roughii_Sta1) and a group of three hexaploids (A.hookeri_Tai, A.cockaynei_Man10 and A.nitens_Wai3), whereas the only decaploid (A.colensoi_BellA10) may have a hybrid origin from A.hookeri_Tai and one of the species in the group of two NZ1 tetraploids (A.allanii_HikB1 and A.roughii_Sta1) (Fig. 6c; Fig. 6d). By allowing reticulate relationships, the networks showed two tetraploids (A.allanii_HikB1 and A.roughii_Sta1) in NZ1 more likely originated from Au (*A. fragosea*) or its ancestors, and they were derived independently of the hexaploids in NZ1. The PhyloNet result in Fig. 6d supported additional ancient hybridization origins of the three hexaploids in NZ1, in which A. hookeri_Tai may be derived independently from A_nitens_Wai3 and A.cockaynei_Man10 (Fig. 6d). As for taxa in NZ2, the result in both networks indicated a single hybridization origin between Australian A.fragosea_CANB797854 and the clade comprising A.allanii_HikB1 and A.roughii_Sta1 (Fig. 6c; Fig. 6d). SNaQ further indicated gene flow between taxa in NZ2 (Fig. 6c).

Moreover, the homeologs of the nrDNA cistron, which were based on the phasing implemented in PATÉ, for the selected 22 individuals (excluding A.robusta_CHR6172728 that lacked genome-skimming data) showed low average allele divergence 0.09% (Table S4). The network estimated by Dendroscope with the multi-labelled tree (Fig. S8a) showed no reticulation signals were found in nrDNA (Fig. S8b).

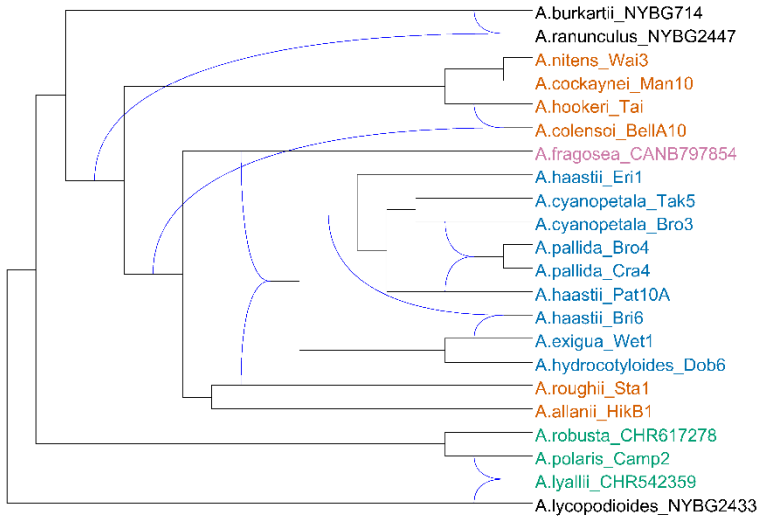
a) ASTRAL Tree



b) SplitsTree



c) SNaQ Network



d) PhyloNet Network

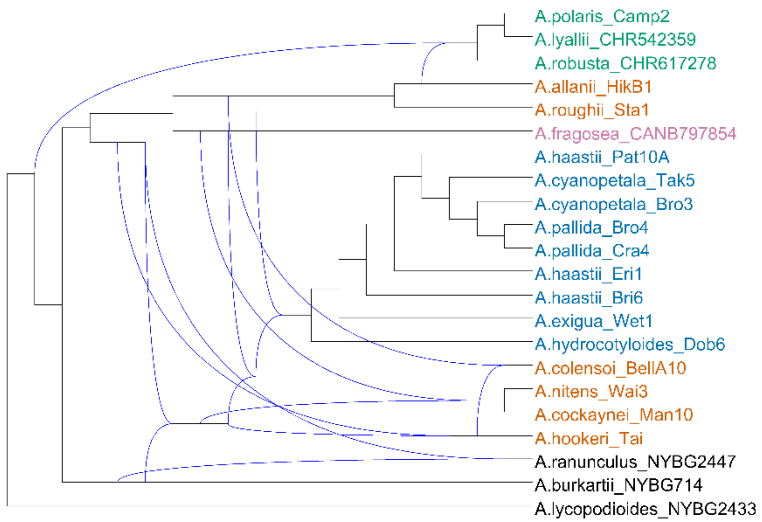


FIGURE 6 Network analysis of 22 *Azorella* individuals representing 14 New Zealand species, one Australian species, and two South American species, plus the outgroup *A. lycopodioides*. The colour of the taxa represents the five identified groups for three *Azorella* sections as in Figure 1 (orange = NZ1, blue = NZ2, green = Sub, pink = Au, black = SA; see text for details about each group). a) The 22-individual ASTRAL tree of selected *Azorella* individuals using 225 filtered SCNGs. The local posterior probabilities, gene concordance pie charts (i.e., blue = concordant, green = discordant with a main alternative, pink = discordant with all remaining alternatives, and grey = uninformative), and the number of supported gene trees (i.e., concordant gene vs all the remaining portions) are labelled at each node. b) SplitsTree network for 21 of the same 22 individuals but excluding the outgroup species *A. lycopodioides*. The box nodes are highlighted by green dots. c) SNaQ and d) PhyloNet estimated networks for the 22-individuals dataset. The blue lines represent hypothesized hybridization events between species or lineages.

2.3.6 ABBA-BABA Test with Genomic SNP Data

Genomic introgression signals among all 22 selected individuals were also calculated by Patterson's D-statistic, and the related admixture fraction f4-ratios and f-branch (fb) values using genomic SNP data. We selected one reference sample that had low allele divergence but a high number of assembled genes, i.e., A.lycopodioides_NYBG2433, with 0.87% allele divergence and 349 assembled genes. In total, there were 35,323 biallelic SNPs selected with 2% missing genotype rate from 163,498 extracted SNPs. The highest fb values corresponding to ASTRAL tree lineages (Fig. 7) can be found between A.colensoi_BellA10 and three hexaploids, especially with A.hookeri_Tai. In addition, A.colensoi_BellA10 also showed signs of introgression from the tetraploid species of A.allanii_HikB1 and A.roughii_Sta1. There were three lineages that showed introgression signals across multiple taxa (Fig. 7), e.g., the individuals of A.burkartii_NYBG714 and A.hydrocotyloides_Dob6, the node between A.polaris_Camp2 and A.lyallii_CHR542359. Moreover, the NZ1 tetraploid group (A.allanii_HikB1 and A.roughii_Sta1) exhibited gene flow signal to the ancestor node of all NZ2 species.

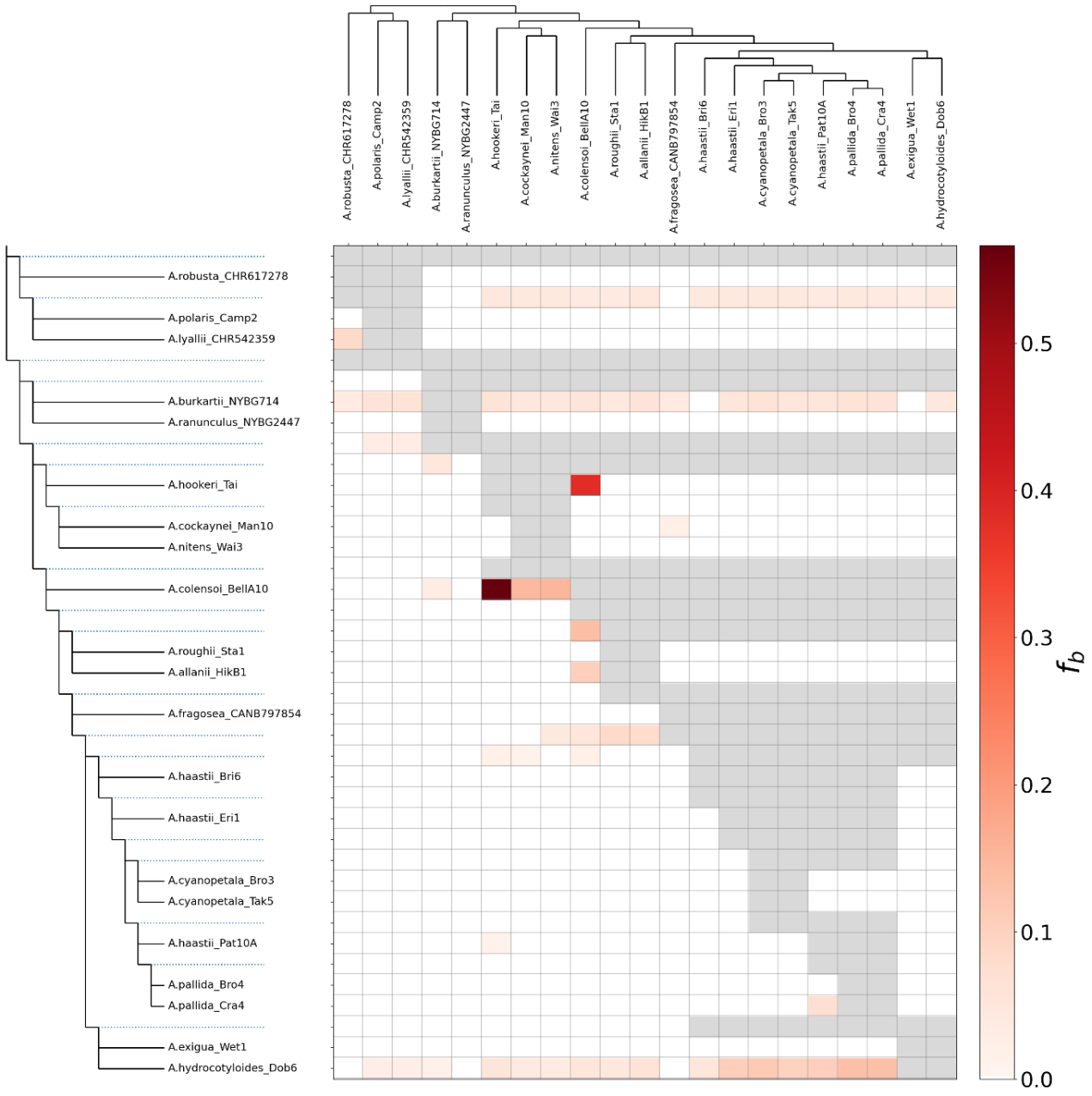


FIGURE 7 Genomic introgression signals identified by the ABBA-BABA test using genomic SNPs extracted from 22 selected *Azorella* individuals . The x-axis corresponds to the ASTRAL tree topology and the y-axis represents all pairwise correlated nodes or tips to the ASTRAL tree. The colour gradient shows f-branch values (0 to 0.5) for each species-trio combination.

2.3.7 Phylogeny and Network Estimation Using SNP Data

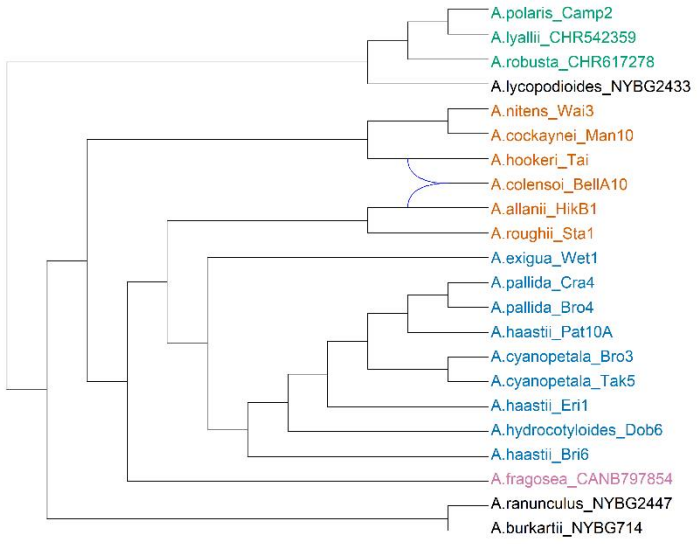
Phylogenies were constructed using 13,399 biallelic SNP sites without any missing data that were selected from 163,498 joint-called SNPs. The multi-species coalescent SVDQuartets tree (Fig. S9a) of 22 individuals showed high bootstrapping values (bs > 90%) on the backbone of the phylogeny, except for the interspecific nodes within NZ2 (bs < 70%) between *A. haastii* subsp.

haastii, *A. hydrocotyloides*, *A. haastii* subsp. *cynopetala* and *A. pallida*. In comparison to the ASTRAL tree topology for the same 22 individuals (Fig. 6a), the SVDQuartets result displayed different placements of A.fragosea_CANB797854 and the lineage comprising two NZ1 tetraploids, A.allanii_HikB1 and A.roughii_Sta1.

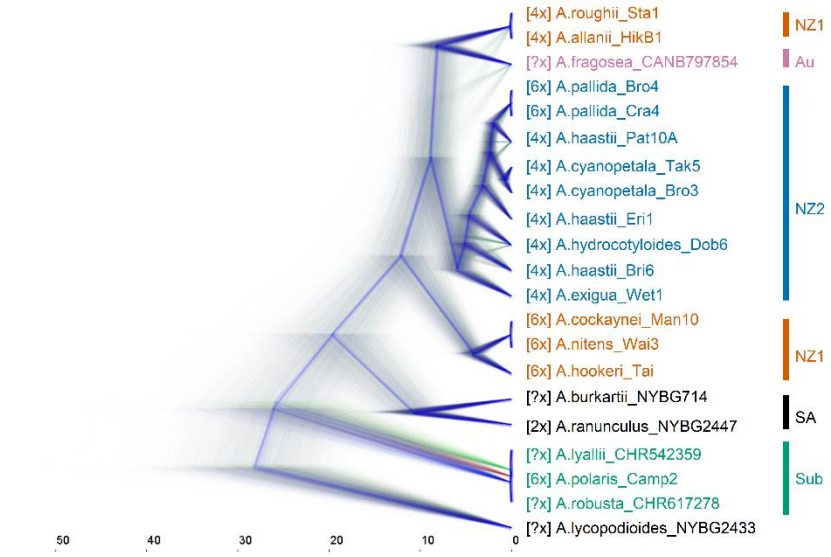
Furthermore, we estimated the Bayesian tree of 22 individuals using the SNAPP approach with the same selected 13,399 biallelic SNPs that had no missing data. The effective sample size (ESS) of the Bayesian analysis only reached 19 (ESS often > 200) after 2,609,250 iterations of Markov chain Monte Carlo (MCMC) chain length, which indicated MCMC did not reach stationarity. After removing 10% burn-in of the 10,436 estimated SNAPP trees, the consensus SNAPP tree (Fig. S9b) also showed similar relationships among the five main groups as the ASTRAL tree in Fig. 6a, but with a few differences (e.g., placement of NZ1 taxa). Although the backbone of the SNAPP species tree was supported by high posterior probabilities (> 0.90), the concordance levels among estimated SNAPP trees (Fig. S9b) showed large disagreements between two tetraploids (A.allanii_HikB1 and A.roughii_Sta1), the group of hexaploid A.hookeri_Tai and the decaploid A.colensoi_BellA10 in NZ1, and at the nodes to A.haastii_Bri6, A.hydrocotyloides_Dob6, and A.haastii_Pat10A in NZ2. Three species within the Sub group also exhibited large discordance levels.

The MCMC_BiMarkers network analysis showed different species relationships using the same SNP dataset with the similar SNAPP based approach (Fig. 8a). The SNAPP network indicated the allopolyploid origins of A.colensoi_BellA10 from A.hookeri_Tai and A.allanii_HikB1. In addition, the SNAPP network topology showed the two tetraploids in NZ1 (A.allanii_HikB1 and A.roughii_Sta1), were more closely derived from the Au lineage *A. fragosea*, which was also supported by the network relationships estimated from the 225 gene trees (Fig. 6c; Fig. 6d).

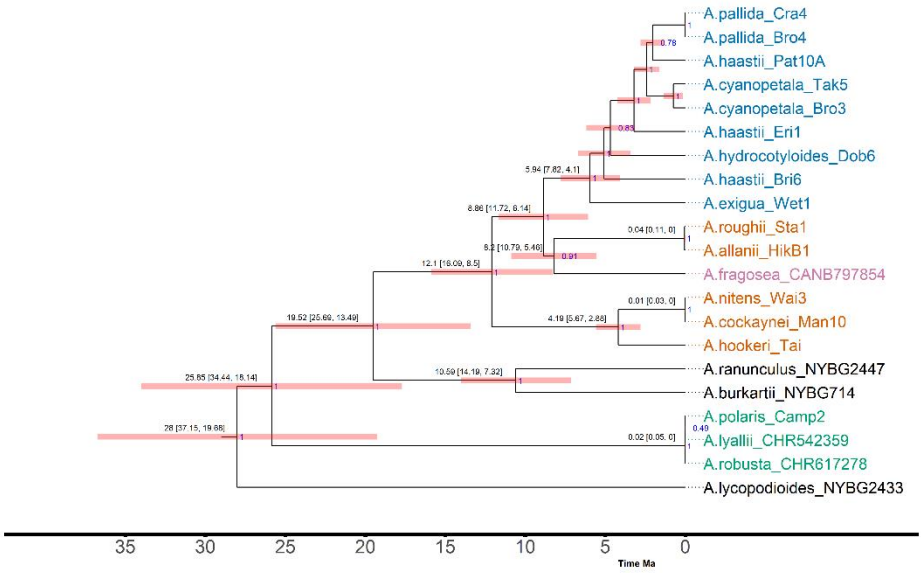
a) MCMC_BiMarkers Network



b) Cloudogram of SNAPP Trees



c) Consensus SNAPP Tree



d) Ancestral Range Reconstruction

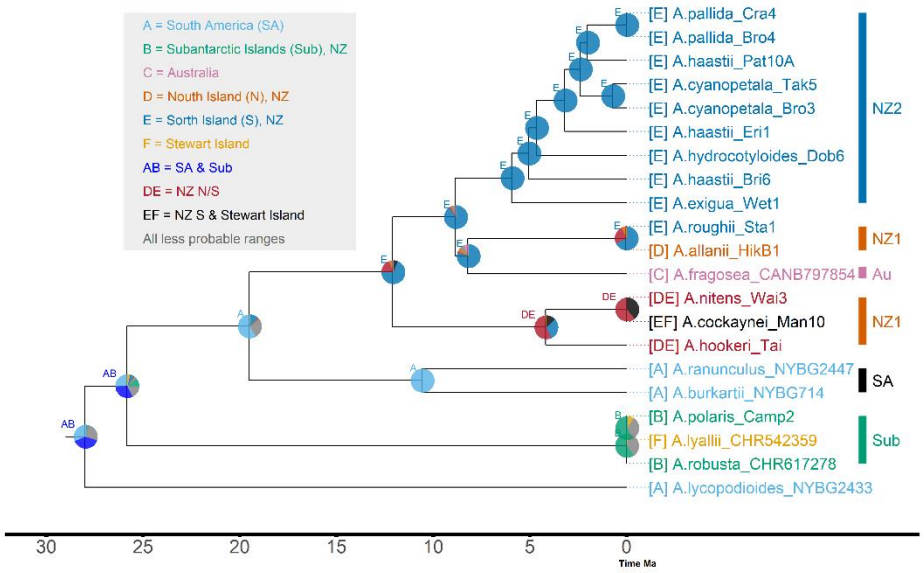


FIGURE 8 Network, phylogeny, divergence time and biogeographic history estimation of *Azorella* species with genomic SNPs. **a)** Bayesian network of selected 22 taxa with MCMC_BiMarker in PhyloNet, the blue line represents the identified hybrid node. **b)** Cloudogram of species relationships for 21 taxa (excluding the decaploid, *A. colensoi*) inferred from Bayesian phylogenetic SNAPP trees. The main consensus SNAPP tree is shown in blue with the alternative topologies represented by green or red. The x-axis shows the predicted divergence time (40 Ma to 0). The colour of individuals (tree tips) represents the previously identified genetic groups (Fig. 3). **c)** The consensus Bayesian SNAPP time-calibrated consensus phylogenetic tree of selected 21 taxa. The posterior probabilities are annotated on each node (maximum = 1), and the time scale bar is labelled on the x-axis. The red bars show the 95% highest posterior density (HPD) corresponding to the divergence time on the x-axis. **d)** Biogeographic inference of selected 21 taxa (as in c). The nodes and individuals (tree tips) are annotated and coloured with their distribution or estimated ancestral ranges before and after cladogenetic events, corresponding to the labels in the upper left box. For the genetic groups of each species, see Fig. 3. The pie charts at each node are proportional to the posterior probability of the estimated ancestral range for that node and colored based on the present species distribution ranges, with the less probable ranges represented in gray.

2.3.8 Divergence times and Biogeographical History of New Zealand Azorella Species

Because the presence of the decaploid individual A.colensoi_BellaA10 could affect the topology of the consensus SNAPP tree (Fig. 8a; Fig. S9a; Fig. S9b), we excluded this individual from the divergence time and biogeographical estimation. We reselected 13,143 biallelic SNPs from the total 163,498 joint-called SNPs for the remaining 21 individuals. The consensus SNAPP tree was extracted from 12,195 SNAPP trees after 10% burn-in. The backbone of the 21-individual SNAPP species tree was supported by high posterior probabilities (Fig. 8c), and it grouped the two tetraploids in NZ1 (A.allanii_HikB1 and A.roughii_Sta1) with Au after removing the hybrid node (Fig 8a). Furthermore, the concordance levels among SNAPP trees for 21 individuals also were improved for the three hexaploids in NZ1 (A.nitens_Wai3, A.cockaynei_Man10, A.hookeri_Tai) (Fig. 8b). However, the discordance levels between the sister species pair A.allanii_HikB1 and A.roughii_Sta1, and A.fragosea_CANB797854, exhibited a high proportion of alternative topologies in Fig. 8b, as well as discordant nodes among NZ2 and Sub.

The BioGeoBEARS results supported the DEC+J as the most likely model for biogeographical inference (Table S6). South America and subantarctic islands were the most likely origins for section *Schizeilema* (NZ1, Au, NZ2), section *Ranunculus* (SA), and section *Stilbocarpa* (Sub) (Fig. 8d). Furthermore, the model indicated two independent dispersal events and possibly two different origins of the two New Zealand *Azorella* sections. Specifically, the ancestor of section *Stilbocarpa* had more ancient evolutionary history that was diverged from South American relatives around 25.85 Ma with 95% highest posterior density (HPD) between 34.44 to 18.14. The ancestral distribution estimation of section *Stilbocarpa* showed it may originated from South America and the

Subantarctic islands, whereas the three extant megaherbs in section *Stilbocarpa* had more recent diversification which was less than 20,000 years. By contrast, the ancestor to section *Schizeilema* dispersed to New Zealand around 12.1 Ma (with 95% HPD between 16.09 to 8.5) from South America. In addition, within section *Schizeilema*, the model also supported *A. allanii* with a more recent colonization to North Island, and *A. fragosea* in Australia probably dispersed from New Zealand around 8.2 Ma (95% HPD between 10.79 to 5.46).

2.4 Discussion

Resolving relationships of polyploid-rich genera can provide insight into the origins and diversification of polyploids (Rothfels, 2021). In this study, we used the universal Angiosperms353 baits (Johnson et al., 2018) and successfully captured over 300 genome-wide single copy markers for phylogenetic, biogeographical, and network inferences of New Zealand *Azorella* polyploids (4*x*, 6*x* and 10*x*) and their close relatives. Our results also highlight the benefits of combining high copy markers (plastomes and nrDNA) with single copy nuclear genes (SCNGs) to provide a more comprehensive understanding of the origins and reticulate relationships of polyploid species.

2.4.1 Target Enrichment Analysis of Polyploid-Rich Genera

Over the last decade, the phylogenetic inferences of non-model plant genera have been significantly improved by the development of new sequencing techniques (Hibbins & Hahn, 2022; McKain et al., 2018; Rothfels, 2021). Similar to all phylogenomic approaches, a single gene tree may be influenced by ILS or missing data (Jones et al., 2013; Maddison and Knowles, 2006). Whereas a more reliable species phylogeny can be inferred when considering the phylogenetic signals (e.g., multi-species coalescent model) from genome-wide markers (Maddison and Knowles, 2006). Taking polyploid species (4*x*, 6*x* and 10*x*) in *Azorella* sections *Schizeilema* and *Stilbocarpa* as an example, we recovered over 300 single copy nuclear loci with sufficient sequencing depth from Angiosperms353 target-captured reads data. Even for individuals with limited starting material (e.g., *A. exigua*; Fig. 2) or herbarium specimens that had degraded DNA quality (Table S2), the robust laboratory and bioinformatic pipelines have recovered sufficient sequence reads for gene recovery (Fig. 3; Table S3).

Although shallow genome-skimming sequenced reads were useful to extract the complete plastome and nrDNA in this paper, Hyb-Seq approach has also shown potential to recover off-target reads from high-copy markers, which can be particularly useful when genome-skimming data are not available (de Lima Ferreira et al., 2022; Karimi et al., 2020). Moreover, we showed the

flexibilities of analysing the Hyb-Seq data via gene sequences and the genomic SNP variation (Fig. 5; Fig. 6; Fig. 8). Both approaches can be informative about the species relationships, and allowing to discover the genomic introgression signals and exploration of interspecific variation (Beck et al., 2021; Kandziora et al., 2021; Slimp et al., 2021; Leaché and Oaks, 2017).

On the other hand, Hyb-Seq sequence reads of allopolyploid or hybrid species can be particularly useful to recover all homeologous sequences for each targeted genes when the sequencing depth and coverage is sufficient (Nauheimer et al., 2021; Tiley et al., 2021). However, this is can be challenging when applying Angiosperms353 bait set to capture only the exons of the targeted genes, which may result in stitched exons from different homeologs of a polyploid (Johnson et al., 2016). In addition, phasing the Hyb-Seq reads for species with different subgenome doners often requires a high level of homeolog divergence, a low ploidy level, or the gene sequences from known diploid ancestors. New Zealand *Azorella* species lack related known diploid lineages and have a high level of allele divergence for the majority of targeted genes (Fig. 3). Phasing *Azorella* Hyb-Seq reads is limited by the short-sequenced reads, discontinuous exons within targeted genes, and especially without clear diploid ancestors as reference sequences (Nauheimer et al., 2021). Phasing polyploids *Azorella* short Hyb-Seq reads based on the sequencing depth only may generate chimeric homeologous gene sequences. By contrast, if only using the recovered genes with low allelic divergence (i.e., no paralogs) to obtain a better supported backbone of *Azorella* phylogeny may result in low or non-resolution of intra/interspecific relationships (Zhou et al., 2022).

In spite of these challenges, our results nevertheless showed the utility of Angiosperms353 SCNGs in providing resolved and meaningful species relationships within a polyploid-rich genus *Azorella* (Fig. 3; Fig. 6a; Fig. 8a), especially when comparing the network results (Fig. 6c; Fig. 6d; Fig. 8b) to the incongruent topologies based on plastome and nrDNA sequences (Fig. 4a; Fig. 4b). Therefore, it is important to compare multiple datasets to interpret hybridization signals and reconstruct polyploidization events. This means comparing biparentally-inherited nuclear markers with maternally inherited chloroplast markers, and incorporating data on chromosome counts and ploidy level, where available. Moreover, we suggest using Hyb-Seq on a long-reads sequencer (e.g., PacBio) that can capture full-length homeologous sequences and reduce conflicting signals among stitched exons. Further improvements to bioinformatic pipelines are also needed to refine the reconstruction of polyploid origins, which can increase the sensitivity of detecting paralogs among Hyb-Seq data [e.g., HybPiper V2 (Johnson et al., 2016)] and resolving the origins of homeologs when reconstructing the gene trees (e.g., Šlenker et al., 2021; Nauheimer et al., 2021).

2.4.2 Phylogeny and Biogeographical History of New Zealand Azorella

With the acknowledgment that the hybridization or polyploidization events can increase the uncertainty of final phylogeny reconstruction (Rothfels, 2021), and the limited taxon sampling (sections *Huanuca* and *Azorella* were not sampled) and uncertainty in the root position (Fig. 8c; Fig. 8d) could affect the outcome of final biogeographical inference, this paper nevertheless provides a comprehensive picture of New Zealand *Azorella* evolution and diversification patterns using robust phylogenomic methods. Compared to the previously unresolved phylogenetic relationships among sections *Schizeilema*, *Stilbocarpa* and *Ranunculus* (Andersson et al., 2006; Fernández et al., 2017; Plunkett and Nicolas, 2017), the SCNGs and SNPs showed clear sectional boundaries and closer relationships between the two smaller rhizomatous herbaceous taxa in sections *Schizeilema* and *Ranunculus* than to the megaherbs in *Stilbocarpa* (Fig. 6a; Fig. 8c). Moreover, the biogeographical inferences from the SNAPP consensus tree (Fig. 8d) suggested section *Stilbocarpa* may have a different origin compared to sections *Schizeilema* and *Ranunculus* that were diverged from ancestral lineages in South America.

The ancestral distribution showed the ancestors to megaherbs in section *Stilbocarpa* may have originated from South America and the subantarctic islands prior to the last glacial maximum (LGM; 20,000 years ago) c. 25.85 Ma (Fig. 8d). Such a dispersal pattern could be explained by species dispersal from Antarctica (e.g., Lehnebach et al., 2017; Sancho et al., 2015) to South American and subantarctic islands, before c. 14 to 3.9 Ma when ice sheet covered the whole Antarctic continent and eliminated all the flowering plants (Lewis et al., 2008; Convey et al., 2008; Sancho et al., 2015). Although, some studies have suggested that during the LGM, glaciers covered nearly most of the subantarctic islands, including all of the land mass of the Auckland Islands, a large portion of Campbell Island, and part of Stewart Island, but not Macquarie Island (McGlone, 2002; Hodgson et al., 2014; Wagstaff et al., 2011), until 15,000 years ago, the retreat of LGM glaciers provided open space for plant recolonization (Suggate, 1990; Fraser et al., 2009; McGlone, 2002). On the other hand, Rainsley et al. (2019) suggested glaciation was much more restricted on the Auckland Islands and Campbell Island during the LGM. Similarly, the megaherb genus *Pleurophyllum* (Asteraceae) also had a post-glacial dispersal history on the New Zealand subantarctic islands and subsequently survived the LGM (Wagstaff et al., 2011). Nevertheless, the model also suggested species within section *Stilbocarpa* were likely diversifying across different subantarctic islands only after LGM (Fig. 8c). Additional sampling of *Azorella* megaherbs from other subantarctic islands would be needed to confirm the finding. For example, *Azorella polaris* now is widely distributed throughout the subantarctic, including Macquarie Island (Australia), Auckland Islands, Campbell Island,

Stewart Island, and Antipodes Island. By contrast, the other two megaherbs have more restricted distributions: *A. lyallii* is endemic to Stewart Island and *A. robusta* can only be found on Snares Island.

The biogeographical model indicated the ancestor of the 14 New Zealand *Azorella* taxa in section *Schizeilema*, which was genetically close to the ancestor of the South America section *Ranunculus* lineage, dispersed to the South Island of New Zealand during the Middle Miocene, 12.1 Ma. The circumpolar westerly winds and ocean currents formed during the Miocene (23 to 5 Ma) (Barker and Burrell, 1982) provided opportunities for plant propagules to disperse from South America to New Zealand (Winkworth et al., 2002; Sanmartín and Ronquist, 2004). The diversification within section *Schizeilema* started from the Late Miocene (11 to 5 Ma) and continued throughout the Pliocene-Pleistocene (5 to 0 Ma), which also fits the predicted time of active vascular plant evolution in New Zealand, beginning in the Late Miocene and thereafter (10 to 0 Ma) (Heenan and McGlone, 2019). The diversification of the South Island endemic alpine taxa (NZ2, blue clade in Fig. 6c) may also be the result of late Pliocene-Pleistocene (5 to 0 Ma) geological events of the Southern Alps, especially their uplift and glaciation cycles (Winkworth et al., 2005). Species diversification could be additionally promoted by reticulation and polyploidization events, as has been the case in many other New Zealand plant genera, e.g., *Veronica* (Plantaginaceae) (Wagstaff and Garnock-Jones, 1998; Thomas et al., 2021), *Ranunculus* (Ranunculaceae) (Lockhart et al., 2001), *Ourisia* (Plantaginaceae) (Meudt and Simpson, 2006), *Pachycladon* (Brassicaceae) (Joly et al., 2009), *Plantago* (Plantaginaceae) (Tay et al., 2010) and *Leptinella* (Asteraceae) (Himmelreich et al., 2014).

Furthermore, the ancestor of *A. fragosea* in section *Schizeilema* dispersed from the South Island of New Zealand to Australia around 8.2 Ma (Fig. 8d). Despite the fact that strong westerly winds may limit dispersal from New Zealand to Australia, dispersal via easterly winds in the lower-atmosphere or via bird migration is still possible (Wardle, 1978; Winkworth et al., 2002). Indeed, several New Zealand plant genera show similar biogeographical patterns of South American origin and long-distance dispersal to New Zealand, where they subsequently underwent diversification and range expansion particularly in the South Island, with accompanied by subsequent dispersal events to Australia from New Zealand, e.g., *Ranunculus* (Lockhart et al., 2001), *Ourisia* (Meudt and Simpson, 2006), *Veronica* (Meudt and Bayly, 2008; Wagstaff et al., 2002), and possibly *Epilobium* (Onagraceae) (Lorimer, 2007).

The sister groups of *A. allanii* and *A. roughii*, and *A. nitens* and *A. cockaynei* both have recent divergence times (Fig. 8c). About 40,000 years ago, the southern North Island native *A. allanii*

(from the Ruahine and Tararua Ranges) diverged from the northern South Island endemic *A. roughii* (from Nelson and Marlborough). The speciation of this sister species pair and their allopatric distribution across Cook Strait may have been facilitated by two recent geologically events, i.e., a land bridge that connected southern areas of the North Island to northern parts of South Island (e.g. between the Wellington and Nelson regions) c. 1 Ma (Trewick and Bland, 2012) and the formation of the North Island mountains and alpine habitats, which began in the middle Quaternary (c 1. Ma). Similarly, the recent migration of *A. cockaynei* from the South Island to Stewart Island may also have been assisted by a land bridge that existed during the Pleistocene (Lockhart et al., 2001).

Unfortunately, *A. schizeilema* that is endemic to the New Zealand subantarctic islands (Auckland Islands and Campbell Island) was not able to be included in this study, but unpublished data from G.P. and A.N. found this species to have an ITS type that was nested between *A. fragosea* and the group of *A. allanii* and *A. roughii*, whereas its plastid marker showed more similarity to the *A. exigua* or NZ2_cp2 haplotype (Fig. 4a). Therefore, as with all other species in section *Schizeilema*, *A. schizeilema* may have originated also from ancestors in the South American (Fig. 8c). However, whether this species diverged from a mainland New Zealand ancestral lineage and dispersed to the subantarctic islands, or it had a separate colonization from South America to subantarctic islands, remains unresolved.

2.4.3 Origins of New Zealand Azorella

The networks reconstructed using the concordance factors of Hyb-Seq recovered SCNGs (Fig. 6c; Fig. 6d) or the Hyb-Seq extracted biallelic SNPs (Fig. 8a) revealed complex network relationships within and between sections *Schizeilema*, *Stilbocarpa* and *Ranunculus*. The topological incongruence between plastome, nrDNA and SCNGs, which was expected for polyploid species due to hybridization and introgression events (de Lima Ferreira et al., 2022; Kandziora et al., 2021; Rose et al., 2021), provided additional genetic evidence to support the network topology of New Zealand *Azorella* polyploids (Fig. 3; Fig. 4a; Fig. 4b).

Azorella section *Schizeilema* may have originated from the ancestors of *A. fragosea*, *A. roughii*, and *A. allanii*, whereas the ancestors of these three species also played important roles regarding the diversification of section *Schizeilema* that we discuss later. According to the plastome and nrDNA bifurcating trees (Fig. 4a, 4b), Australian *A. fragosea* had a plastome type and nrDNA cistron variation that placed it nested within the mainland New Zealand clade and close to *A. allanii* and *A. roughii*. All the Hyb-Seq network results (Fig. 6c; Fig. 6d; Fig. 8a) and the bifurcating SNAPP trees

of 21 individuals (Fig. 8b; Fig. 8c) further exhibited a consistent topology in which the lineage comprising *A. allanii* and *A. roughii* was derived from the same ancestor lineage as *A. fragosea*.

The species relationships among the two NZ1 tetraploids (*A. allanii*, *A. roughii*), three hexaploids (*A. hookeri*, *A. nitens*, and *A. cockaynei*) and the decaploid (*A. colensoi*) varied in the different bifurcating phylogenetic trees (Fig.3; Fig. 6a; Fig. 8b; Fig. S9a). By contrast, the PhyloNet result (Fig. 6d) indicated that *A. hookeri* and the sister species pair *A. nitens* and *A. cockaynei* had two independent hybrid origins: one of their parental species was related to the ancestors of *A. fragosea*, *A. allanii* and *A. roughii*, and the other was related to section *Ranunculus*. This result was also supported by plastome and nrDNA incongruence, such that the three hexaploids had two plastome types (possible maternal lineages), especially the lineage comprising *A. nitens* and *A. cockaynei*, which was sister to the group of *A. allanii* and *A. roughii* that had *A. hookeri* and *A. colensoi* nested within it (Fig. 4a). On the other hand, the nrDNA tree of the three hexaploids showed homogenized ITS types that were more closely related to South American section *Ranunculus* (possible paternal lineage) than section *Schizeilema* (Fig. 4b). It is possible that the maternal lineages of the two hybrid origins of *A. hookeri* and sister pair *A. nitens* and *A. cockaynei* had similar plastome variation, and their paternal lineage may be from the same South American taxon.

One of the consistent findings throughout our analysis was the origin of the only decaploid *A. colensoi* in New Zealand (Fig. 3; Fig. 6a; Fig. 8a). The concordance analysis of the SCNG phylogenies (Fig. 6a; Fig. S9b), network topologies (Fig. 6c; Fig. 6d; Fig. 8a), and genomic introgression test (Fig. 7) all suggested the allopolyploid origin of *A. colensoi* from a hexaploid *A. hookeri* as a maternal parent, because these two species shared nearly identical plastome sequences (Fig. 4a). By contrast, the paternal lineage of *A. colensoi* was related to the sister species pair of *A. roughii* and *A. allanii*, and is most likely to be *A. allanii* (Fig. 6d), because *A. colensoi* is also a North Island endemic whose geographic distribution overlaps with that of *A. allanii*. We did not include *A. colensoi* in the divergence time estimation due to complexities in the placement of this allodecaploid in the phylogeny (Fig. 6c, 6d). However, given that *A. colensoi* likely formed on the North Island, where volcanic eruptions and mountain uplift occurred c. 1 Ma (Trewick and Bland, 2012), and the split between *A. roughii* and *A. allanii* was 40,000 years ago, we suggest that the formation of this polyploid species was relatively recent and likely coincident with major geological events on the North Island.

The network topologies based on Hyb-Seq SCNG phylogenies showed one single hybridization origin of NZ2 taxa between the ancestors of the lineage that included current *A. fragosea*, *A. allanii*

and *A. roughii* (Fig. 6c; Fig. 6d). In addition, NZ2 showed two plastome types but only one nrDNA cistron type (Fig. 4a; Fig. 4b) that may also indicate the hybrid origin. Therefore, we hypothesize the NZ2_cp2 plastome type (Fig. 4a) was inherited and maintained via the maternal lineage that is closer to current *A. fragosea*, while the chloroplast type in NZ2_cp1 was introduced by introgression events with the paternal lineage that is more similar to current *A. roughii* and *A. allanii*. Indeed, NZ2 exhibited substantial gene flow and complex network relationships as evidenced by the discordance levels of Hyb-Seq gene trees and SNP trees (Fig. 6a; Fig. 8a Fig. S9c), as well as the reticulation signals in the network result (Fig. 6c). The single nrDNA type of NZ2 (Fig. 4b) could have also resulted from concerted evolution, accelerated by reticulation events to promote the homogenization process in NZ2 taxa (Álvarez and Wendel, 2003; Hillis and Dixon, 1991; Garcia-Jacas et al., 2009).

Azorella polaris (6*x*), *A. robusta* and *A. lyallii* in section *Stilbocarpa* (Sub) formed a clade in the SCNG, plastome and nrDNA trees that was sister to section *Ranunculus* (Fig. 3; Fig. 4a; Fig. 4b). These three megaherbs have similar growth form and were recently radiated onto different subantarctic islands (Fig. S1; Fig. 8d). The network indicated they have experienced gene flow from the ancestor of the outgroup lineage, *A. lycopodioides* (Fig. 6c; Fig. 6d) and may also have genetic contribution from the ancestor of *A. allanii* (Fig. 6d). However, the species relationships among these three megaherbs in section *Stilbocarpa* remain unresolved. This result is similar to a previous study based on ITS in which the three species formed a polytomy (Mitchell et al., 1999) Our results also showed large portions of discordant genes trees (Fig. 6a) or SNAPP trees (Fig. 8b) of Hyb-Seq data which all supported the alternative topologies among the three species need to be considered. In addition, although *A. lyallii* and *A. polaris* showed clear genetic boundaries in the SCNG ASTRAL tree (Fig. 3), which could also indicate interspecific variation, more comprehensive sampling of each species and using Angiosperms353 baits may help determine the species boundaries and population structure (e.g., Beck et al., 2021) among these subantarctic species.

Two species in section *Ranunculus*, *A. ranunculus* and *A. burkartii*, also were a monophyletic group in the single and high copy gene trees. However, in addition to the gene flow detected in this section, the networks (Fig. 6c; Fig. 6d) showed the ancestor to section *Ranunculus* may also be involved in the origin of section *Schizeilema*, especially for *A. hookeri*, sister species pair *A. nitens* and *A. cockaynei* (Fig. 4b). In the future, including additional samples of the remaining species (*A. boelckei* in section *Ranunculus*), as well as the species in sections *Azorella* and *Huanaca*, can help to fully resolve the phylogenetic complexity among these sections (Plunkett and Nicolas, 2017).

Conclusions

Target-enrichment sequencing using the Angiosperms353 bait set provides an efficient way of capturing genome-wide phylogenomic signals, which are essential when inferring the phylogeny of polyploid-rich groups that often have complex evolutionary histories involving polyploidization and reticulation (Johnson et al., 2018; Rothfels, 2021). Our results supported the South American origins of section *Schizeilema* and possible Antarctic assisted dispersal of megaherbs in section *Stilbocarpa* (Fig. 8d) (Nicolas and Plunkett, 2014). Section *Schizeilema* diverged from the ancestors of section *Ranunculus* c.12.1 Ma, while section *Stilbocarpa* diverged from ancestors of South American relatively c. 25.85 Ma .

Furthermore, phylogenetic inferences using network approaches utilized the gene concordance factors between reconstructed single copy nuclear gene trees and provided insight into the species origins of the polyploids, including hybrid origins for three hexaploid lineages *A. hookeri* , *A. nitens* and *A. cockaynei*, and an allopolyploid origin for *A. colensoi* (10*x*) from parental species *A. hookeri* (6*x*) and *A. allanii* (4*x*). The results were also able to detect gene flow between specific South Island alpine lineages that were diverged from the ancestors of *A. fragosea* and the sister species pair of *A. roughii* and *A. allanii* c 6.41 Ma, including *A. exigua* (4*x*), *A. haastii.* subsp. *haastii* (4*x*), *A. haastii.* subsp. *cyanopetala* (4*x*), *A. hydrocotyloides* (4*x*), and *A. pallida* (6*x*), which may have been facilitated by historical geological events such as mountain uplift and glaciation cycles during the late Pliocene-Pleistocene (from c. 5 Ma) in New Zealand (Winkworth et al., 2005).

Topological incongruence between bifurcating trees further highlighted the importance of including network approaches when inferring the phylogeny of polyploid-rich groups. This network framework also provides a phylogenetic hypothesis for further investigating species diversification patterns using morphological traits and ecological niche modelling. Moreover, the sufficient sequencing depth also allowed the extraction of biallelic SNPs for investigating intra- and interspecific variation (Slimp et al., 2021), or phasing the homeologous blocks to improve each genealogy reconstruction (Tiley et al., 2021; Šlenker et al., 2021). Therefore, this approach may be useful for other groups in biodiversity hotspots with high levels of polyploidy.

Data Availability

The target enrichment sequencing data and genome-skimming data are available at NCBI SRA: PRJNA885464. Scripts for conducting the analyses are available at: https://github.com/WeixuanPlant/NZAzorellaAngiosperms353

Author Contributions

W.N., H.M., W.L. and J.T. designed the study. W.N. and H.M. collected most of the field samples. A.N. and G.P. provided some of the genome-skimming data and all South American *Azorella* leaf samples. P.H. provided leaf material of two individuals *A. hookeri* for DNA extraction and species identifications. W.N. conducted the genomic laboratory work, analysed the data, and wrote the manuscript with contributions from all authors.

Acknowledgements

All field collections were made under the New Zealand Department of Conservation Global Concession CA-5160-OTH. We thank the many people who assisted with field work or collected samples, but especially John Henry, Cara-Lisa Schloots, David Glenny, Chris Ecroyd, Demet Tore, Kay Pilkington, Alex Fergus, and Antony Kusabs. A very special thanks to Xiaixiao Lin at Massey Genome Service, who made an enormous effort to polish the laboratory steps with W.N. We are grateful to staff from the following herbaria for assistance with specimen information, localities, and allowing use of the specimens for this study: Allan Herbarium (CHR), Auckland Museum (AK), Te Papa (WELT), Dame Ella Campbell Herbarium (MPN), Australian National Herbarium (CANB), and the New York Botanical Garden (NYBG). We gratefully acknowledge the New Zealand eScience Infrastructure (NeSI) for providing data storage, computing resources, and bioinformatics support. We also thank Matt Johnson, Lars Nauheimer and Andrew Crowl for their suggestions regarding the bioinformatic pipeline for Hyb-Seq data during the course of this study.

Funding

This study was conducted as part of the Royal Society of New Zealand Marsden fund (17-LCR-006) to W.L., H.M., and J.T. Additional funding from the Australasian Systematic Botany Society Hansjörg Eichler Scientific Research Fund and the Royal Society of New Zealand Hutton Fund to W.N. are gratefully acknowledged.

References Cited

Allan H.H. 1961. Flora of New Zealand. Vol. I., R.E. Owen Government Printer, Wellington, New Zealand.

Álvarez I., Wendel J.F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogen. Evol., 29:417-434.

Andersson L., Kocsis M., Eriksson R. 2006. Relationships of the genus *Azorella* (Apiaceae) and other hydrocotyloids inferred from sequence variation in three plastid markers. Taxon, 55:270-280.

Baker W.J., Bailey P., Barber V., Barker A., Bellot S., Bishop D., Botigué L.R., Brewer G., Carruthers T., Clarkson J.J. 2022. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Syst. Biol., 71:301-319.

Baldwin B.G., Sanderson M.J., Porter J.M., Wojciechowski M.F., Campbell C.S., Donoghue M.J. 1995. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann. Mo. Bot. Gard.:247-277.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol., 19:455-477.

Barker P., Burrell J. 1982. The influence upon Southern Ocean circulation, sedimentation, and climate of the opening of Drake Passage. Antarctic geoscience, Univesity Wisconsin Press, Madison, USA, p. 377-385.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon, 56:417-426.

Beck J.B., Markley M.L., Zielke M.G., Thomas J.R., Hale H.J., Williams L.D., Johnson M.G. 2021. Are Palmer's elm-leaf goldenrod and the smooth elm-leaf goldenrod real? The Angiosperms353 kit provides within-species signal in *Solidago ulmifolia* sl. Syst. Bot., 46:1107-1113.

Beuzenberg E., Hair J. 1983. Contributions to a chromosome atlas of the New Zealand flora - 25 Miscellaneous species. N. Z. J. Bot., 21:13-20.

Birky C.W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proceedings of the National Academy of Sciences, 92:11331-11338.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol., 10:e1003537.

Bouckaert R.R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics, 26:1372-1373.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25:1972-1973.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics, 30:3317-3324.

Clark J.W., Donoghue P.C.J. 2018. Whole-genome duplication and plant macroevolution. Trends Plant Sci., 23:933-945.

Convey P., Gibson J.A.E., Hillenbrand C.-D., Hodgson D.A., Pugh P.J.A., Smellie J.L., Stevens M.I. 2008. Antarctic terrestrial life – challenging the history of the frozen continent? Biological Reviews, 83:103-117.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T. 2011. The variant call format and VCFtools. Bioinformatics, 27:2156-2158.

de Lima Ferreira P., Batista R., Andermann T., Groppo M., Bacon C.D., Antonelli A. 2022. Target sequence capture of Barnadesioideae (Compositae) demonstrates the utility of low coverage loci in phylogenomic analyses. Mol. Phylogen. Evol.:107432.
Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin, Vol. 19:pp. 11-15.
Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol., 7:214.
Fernández M., Ezcurra C., Calviño C.I. 2017. Chloroplast and ITS phylogenies to understand the evolutionary history of southern South American *Azorella*, *Laretia* and *Mulinum* (Azorelloideae, Apiaceae). Mol. Phylogen. Evol., 108:1-21.
Fraser C.I., Nikula R., Spencer H.G., Waters J.M. 2009. Kelp genes reveal effects of subantarctic sea ice during the Last Glacial Maximum. Proceedings of the National Academy of Sciences, 106:3249-3253.
Garcia-Jacas N., Soltis P.S., Font M., Soltis D.E., Vilatersana R., Susanna A. 2009. The polyploid series of *Centaurea toletana*: Glacial migrations and introgression revealed by nrDNA and cpDNA sequence analyzes. Mol. Phylogen. Evol., 52:377-394.
Hair J. 1980. Contributions to a chromosome atlas of the New Zealand flora - 21 Umbelliferae (miscellaneous genera). N. Z. J. Bot., 18:559-562.
Heenan P.B., McGlone M.S. 2019. Cenozoic formation and colonisation history of the New Zealand vascular flora based on molecular clock dating of the plastid rbcL gene. N. Z. J. Bot., 57:204-226.
Hendriks K.P., Mandáková T., Hay N.M., Ly E., Hooft van Huysduynen A., Tamrakar R., Thomas S.K., Toro-Núñez O., Pires J.C., Nikolov L.A. 2021. The best of both worlds: combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. Appl. Plant Sci., 9:e11438.
Hibbins M.S., Hahn M.W. 2022. Phylogenomic approaches to detecting and characterizing introgression. Genetics, 220:iyab173.
Hillis D.M., Dixon M.T. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. The Quarterly Review of Biology 66:411-453.
Himmelreich S., Breitwieser I., Oberprieler C. 2014. Phylogenetic relationships in the extreme polyploid complex of the New Zealand genus *Leptinella* (Compositae: Anthemideae) based on AFLP data. Taxon, 63:883-898.
Hodgson D.A., Graham A.G.C., Roberts S.J., Bentley M.J., Cofaigh C.Ó., Verleyen E., Vyverman W., Jomelli V., Favier V., Brunstein D., Verfaillie D., Colhoun E.A., Saunders K.M., Selkirk P.M., Mackintosh A., Hedding D.W., Nel W., Hall K., McGlone M.S., Van der Putten N., Dickens W.A., Smith J.A. 2014. Terrestrial and submarine evidence for the extent and timing of the Last Glacial Maximum and the onset of deglaciation on the maritime-Antarctic and sub-Antarctic islands. Quat. Sci. Rev., 100:137-158.
Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol., 23:254-267.
Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature, 473:97-100.
Jin J.-J., Yu W.-B., Yang J.-B., Song Y., DePamphilis C.W., Yi T.-S., Li D.-Z. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol., 21:1-31.
Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci., 4:apps.1600016.

Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K.-s., Baker W.J., Wickett N.J. 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol., 68:594-606.

Joly S., Heenan P.B., Lockhart P.J. 2009. A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of *Pachycladon* (Brassicaceae) in New Zealand. Mol. Phylogen. Evol., 51:365-372.

Joly S., Heenan P.B., Lockhart P.J. 2014. Species radiation by niche shifts in New Zealand's rockcresses (*Pachycladon*, Brassicaceae). Syst. Biol., 63:192-202.

Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol., 62:467-478.

Junier T., Zdobnov E.M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics, 26:1669-1670.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods, 14:587-589.

Kandziora M., Sklenář P., Kolář F., Schmickl R. 2021. How to tackle phylogenetic discordance in recent and rapidly radiating groups? Developing a workflow using *Loricaria* (Asteraceae) as an example. Front. Plant Sci., 12:765719-765719.

Karbstein K., Tomasello S., Hodač L., Wagner N., Marinček P., Barke B.H., Paetzold C., Hörandl E. 2022. Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large Ranunculus auricomus species complex. New Phytol., 235:2081-2098.

Karbstein K., Tomasello S., Hodač L., Wagner N., Marinček P., Barke B.H., Paetzold C., Hörandl E. 2022. Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large *Ranunculus auricomus* species complex. New Phytol., https://doi.org/10.1111/nph.18284.

Karimi N., Grover C.E., Gallagher J.P., Wendel J.F., Ané C., Baum D.A. 2020. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; Bombacoideae; Malvaceae). Syst. Biol., 69:462-478.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol., 30:772-780.

Leaché A.D., Oaks J.R. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu. Rev. Ecol., Evol. Syst., 48:69-84.

Lehnebach C.A., Winkworth R.C., Becker M., Lockhart P.J., Hennion F. 2017. Around the pole: evolution of sub-Antarctic *Ranunculus*. J. Biogeogr., 44:875-886.

Lewis A.R., Marchant D.R., Ashworth A.C., Hedenäs L., Hemming S.R., Johnson J.V., Leng M.J., Machlus M.L., Newton A.E., Raine J.I., Willenbring J.K., Williams M., Wolfe A.P. 2008. Mid-Miocene cooling and the extinction of tundra in continental Antarctica. Proceedings of the National Academy of Sciences, 105:10676-10680.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Lockhart P.J., McLenachan P.A., Havell D., Glenny D., Huson D., Jensen U. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. Ann. Mo. Bot. Gard., 88:458-477.

Lorimer N.G. 2007. Phylogenetic reconstruction and gene tree incongruence in New Zealand *Epilobium* L.(Onagraceae). The University of Auckland.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol., 55:21-30.

Malinsky M., Matschiner M., Svardal H. 2021. Dsuite-fast D-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour., 21:584-595.

Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol., 2:1940-1955.
Matzke N. 2013. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Front. Biogeogr., 5(4):242–248.
McGlone M.S. 2002. The Late Quaternary peat, vegetation and climate history of the Southern Oceanic Islands of New Zealand. Quat. Sci. Rev., 21:683-707.
McKain M.R., Johnson M.G., Uribe-Convers S., Eaton D., Yang Y. 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci., 6:e1038.
McLay T.G.B., Birch J.L., Gunn B.F., Ning W., Tate J.A., Nauheimer L., Joyce E.M., Simpson L., Schmidt-Lebuhn A.N., Baker W.J., Forest F., Jackson C.J. 2021. New targets acquired: improving locus recovery from the Angiosperms353 probe set. Appl. Plant Sci., 9.
Meudt H.M., Albach D.C., Tanentzap A.J., Igea J., Newmarch S.C., Brandt A.J., Lee W.G., Tate J.A. 2021. Polyploidy on islands: its emergence and importance for diversification. Front. Plant Sci., 12.
Meudt H.M., Bayly M.J. 2008. Phylogeographic patterns in the Australasian genus *Chionohebe* (*Veronica* s.l., Plantaginaceae) based on AFLP and chloroplast DNA sequences. Mol. Phylogen. Evol., 47:319-338.
Meudt H.M., Simpson B.B. 2006. The biogeography of the austral, subalpine genus *Ourisia* (Plantaginaceae) based on molecular phylogenetic evidence: South American origin and dispersal to New Zealand and Tasmania. Biol. J. Linn. Soc., 87:479-513.
Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol., 37:1530-1534.
Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics, 30:i541-i548.
Mitchell A.D., Meurk C.D., Wagstaff S.J. 1999. Evolution of *Stilbocarpa*, a megaherb from New Zealand's sub-antarctic islands. N. Z. J. Bot., 37:205-211.
Nauheimer L., Weigner N., Joyce E., Crayn D., Clarke C., Nargar K. 2021. HybPhaser: a workflow for the detection and phasing of hybrids in target capture data sets. Appl. Plant Sci., 9:e11441.
Nicolas A.N., Plunkett G.M. 2012. Untangling generic limits in *Azorella*, *Laretia*, and *Mulinum* (Apiaceae: Azorelloideae): Insights from phylogenetics and biogeography. Taxon, 61:826-840.
Nicolas A.N., Plunkett G.M. 2014. Diversification times and biogeographic patterns in Apiales. Bot. Rev., 80:30-58.
Olmstead R.G., Palmer J.D. 1994. Chloroplast DNA systematics: a review of methods and data analysis. Am. J. Bot., 81:1205-1224.
Ortiz E. 2019. vcf2phylip v2. 0: convert a VCF matrix into several matrix formats for phylogenetic analysis. URL https://doi org/105281/zenodo, 2540861.
Otto S.P., Whitton J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet., 34:401-437.
Plunkett G.M., Nicolas A.N. 2017. Assessing *Azorella* (Apiaceae) and its allies: Phylogenetics and a new classification. Brittonia, 69:31-61.
Poplin R., Ruano-Rubio V., DePristo M.A., Fennell T.J., Carneiro M.O., Van der Auwera G.A., Kling D.E., Gauthier L.D., Levy-Moonshine A., Roazen D. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv:201178.
Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I., Daly M.J. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 81:559-575.

Qiu F., Baack E.J., Whitney K.D., Bock D.G., Tetreault H.M., Rieseberg L.H., Ungerer M.C. 2019. Phylogenetic trends and environmental correlates of nuclear genome size variation in *Helianthus* sunflowers. New Phytol., 221:1609-1618.
R Core Development Team. 2013. R: a language and environment for statistical computing. Vienna, Austria.
Rainsley E., Turney C.S.M., Golledge N.R., Wilmshurst J.M., McGlone M.S., Hogg A.G., Li B., Thomas Z.A., Roberts R., Jones R.T., Palmer J.G., Flett V., de Wet G., Hutchinson D.K., Lipson M.J., Fenwick P., Hines B.R., Binetti U., Fogwill C.J. 2019. Pleistocene glacial history of the New Zealand subantarctic islands. Climate of the Past, 15:423-448.
Rauscher J.T., Doyle J.J., Brown A. 2004. Multiple origins and nrDNA internal transcribed spacer homeologue evolution in the *Glycine tomentella* (Leguminosae) allopolyploid complex. Genetics, 166:987-998.
Rose J.P., Toledo C.A., Lemmon E.M., Lemmon A.R., Sytsma K.J. 2021. Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. Syst. Biol., 70:162-180.
Rothfels C.J. 2021. Polyploid phylogenetics. New Phytol., 230:66-72.
Sancho G., de Lange P.J., Donato M., Barkla J., Wagstaff S.J. 2015. Late Cenozoic diversification of the austral genus *Lagenophora* (Astereae, Asteraceae). Bot. J. Linn. Soc., 177:78-95.
Sanmartín I., Ronquist F. 2004. Southern Hemisphere Biogeography Inferred by Event-Based Models: Plant versus Animal Patterns. Syst. Biol., 53:278-298.
Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat.:461-464.
Shaw J., Lickey E.B., Schilling E.E., Small R.L. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot., 94:275-288.
Slater G.S.C., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 6:31.
Šlenker M., Kantor A., Marhold K., Schmickl R., Mandáková T., Lysak M.A., Perný M., Caboňová M., Slovák M., Zozomová-Lihová J. 2021. Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq data: a case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. Front. Plant Sci., 12.
Slimp M., Williams L.D., Hale H., Johnson M.G. 2021. On the potential of Angiosperms353 for population genomic studies. Appl. Plant Sci., 9.
Small R.L., Cronn R.C., Wendel J.F. 2004. Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot., 17:145-170.
Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol., 15:1-15.
Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol., 34:3292-3298.
Soltis D.E., Albert V.A., Leebens-Mack J., Bell C.D., Paterson A.H., Zheng C., Sankoff D., de Pamphilis C.W., Wall P.K., Soltis P.S. 2009. Polyploidy and angiosperm diversification. Am. J. Bot., 96:336-348.
Soltis D.E., Visger C.J., Soltis P.S. 2014. The polyploidy revolution then…and now: Stebbins revisited. Am. J. Bot., 101:1057-1078.
Straub S.C., Parks M., Weitemier K., Fishbein M., Cronn R.C., Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am. J. Bot., 99:349-364.
Suggate R.P. 1990. Late pliocene and quaternary glaciations of New Zealand. Quat. Sci. Rev., 9:175-197.

Sugiura N. 1978. Further analysts of the data by akaike's information criterion and the finite corrections: further analysts of the data by akaike's. Commun. Stat. Theory Methods, 7:13-26.
Swofford D.L., Sullivan J. 2003. Phylogeny inference based on parsimony and other methods using PAUP*. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny, cáp, 7:160-206.
Tay M.L., Meudt H.M., Garnock-Jones P.J., Ritchie P.A. 2010. DNA sequences from three genomes reveal multiple long-distance dispersals and non-monophyly of sections in Australasian *Plantago* (Plantaginaceae). Aust. Syst. Bot., 23:47-68.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics, 9:322.
Thomas A.E., Igea J., Meudt H.M., Albach D.C., Lee W.G., Tanentzap A.J. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. Am. J. Bot., 108:1289-1306.
Tiley G.P., Crowl A.A., Manos P.S., Sessa E.B., Solís-Lemus C., Yoder A.D., Burleigh J.G. 2021. Phasing alleles improves network inference with allopolyploids. bioRxiv, 10.1101/2021.05.04.442457:2021.2005.2004.442457.
Tillich M., Lehwark P., Pellizzer T., Ulbricht-Jones E.S., Fischer A., Bock R., Greiner S. 2017. GeSeq–versatile and accurate annotation of organelle genomes. Nucleic Acids Res., 45:W6-W11.
Trewick S.A., Bland K.J. 2012. Fire and slice: palaeogeography for biogeography at New Zealand's North Island/South Island juncture. J. R. Soc. N. Z., 42:153-183.
Tsitrone A., Kirkpatrick M., Levin D.A. 2003. A model for chloroplast capture. Evolution, 57:1776-1782.
Tung Ho L.s., Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol., 63:397-408.
Van de Peer Y., Mizrachi E., Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet., 18:411-424.
Wagstaff S.J., Bayly M.J., Garnock-Jones P.J., Albach D.C. 2002. Classification, origin, and diversification of the New Zealand hebes (Scrophulariaceae). Ann. Mo. Bot. Gard., 89:38-63.
Wagstaff S.J., Breitwieser I., Ito M. 2011. Evolution and biogeography of *Pleurophyllum* (Astereae, Asteraceae), a small genus of megaherbs endemic to the subantarctic islands. Am. J. Bot., 98:62-75.
Wagstaff S.J., Garnock-Jones P.J. 1998. Evolution and biogeography of the *Hebe* complex (Scrophulariaceae) inferred from ITS sequences. N. Z. J. Bot., 36:425-437.
Wan D., Sun Y., Zhang X., Bai X., Wang J., Wang A., Milne R. 2014. Multiple ITS copies reveal extensive hybridization within *Rheum* (Polygonaceae), a genus that has undergone rapid radiation. PLOS One, 9.
Wang X., Morton J.A., Pellicer J., Leitch I.J., Leitch A.R. 2021. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. Plant J., 107:1003-1015.
Wardle P. 1978. Origin of the New Zealand mountain flora, with special reference to trans-Tasman relationships. N. Z. J. Bot., 16:535-550.
Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci., 2:1400042.
Winkworth R.C., Wagstaff S.J., Glenny D., Lockhart P.J. 2002. Plant dispersal N.E.W.S from New Zealand. Trends Ecol. Evol., 17:514-520.
Winkworth R.C., Wagstaff S.J., Glenny D., Lockhart P.J. 2005. Evolution of the New Zealand mountain flora: origins, diversification and dispersal. Org. Divers. Evol., 5:237-247.

Xu B., Zeng X.-M., Gao X.-F., Jin D.-P., Zhang L.-B. 2017. ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). Sci. Rep., 7:1-15.
Yardeni G., Viruel J., Paris M., Hess J., Groot Crego C., de La Harpe M., Rivera N., Barfuss M.H.J., Till W., Guzmán-Jacob V., Krömer T., Lexer C., Paun O., Leroy T. 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. Mol. Ecol. Resour., 22:927-945.
Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol., 8:28-36.
Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. RECOMB international workshop on comparative genomics, Springer, p. 53-75.
Zhou W., Soghigian J., Xiang Q.-Y. 2022. A new pipeline for removing paralogs in target enrichment data. Syst. Biol., 71:410-425.

Supplementary Figures

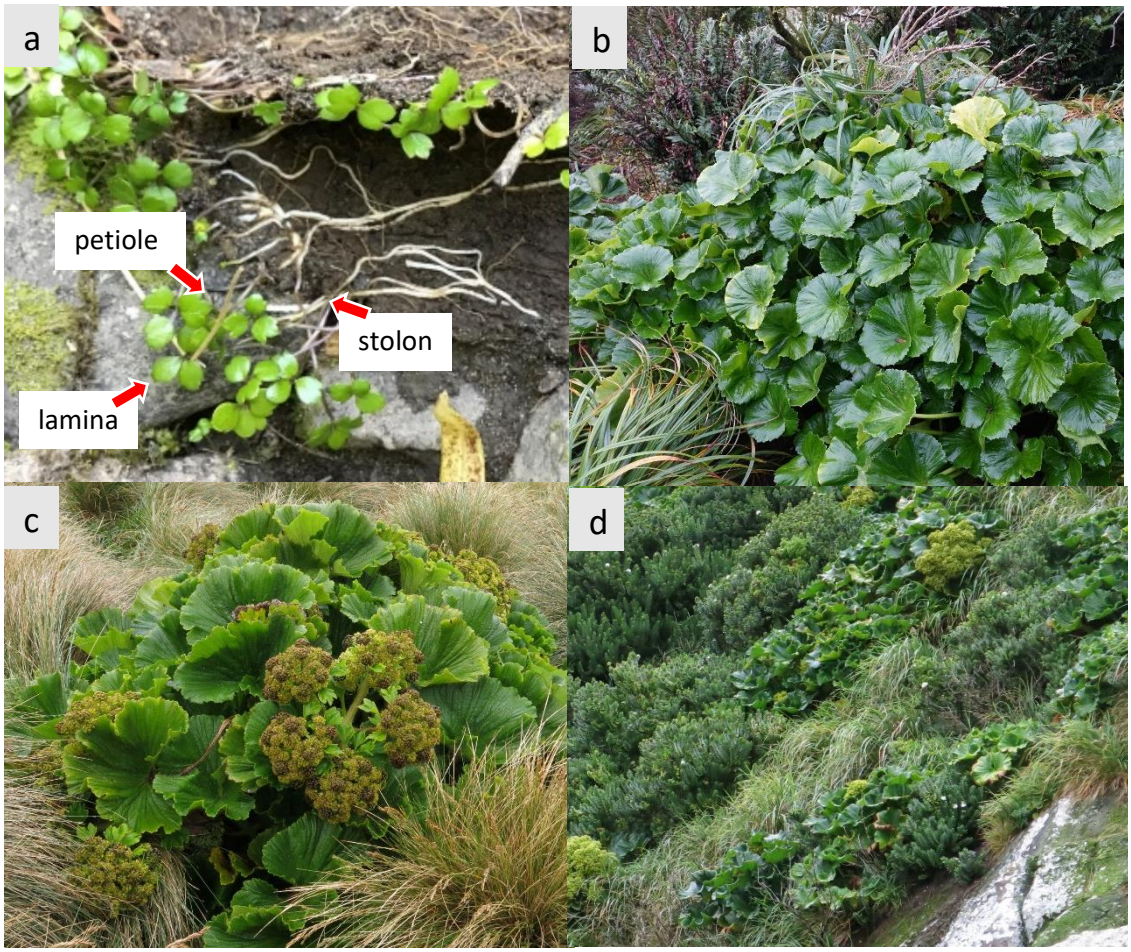


Figure S1. Habit of New Zealand *Azorella* species. a) Photo of field collected *A. nitens* by © Weixuan Ning (MPN 52525, *Azorella* section *Schizeilema*) showing the vegetative growth of the plant; b-d) Field photographs of the three species in *Azorella* section *Stilbocarpa*, i.e., b) *A. lyallii*, c) *A. polaris* and d) *A. robusta* (from iNaturalist.nz © John Barkla).

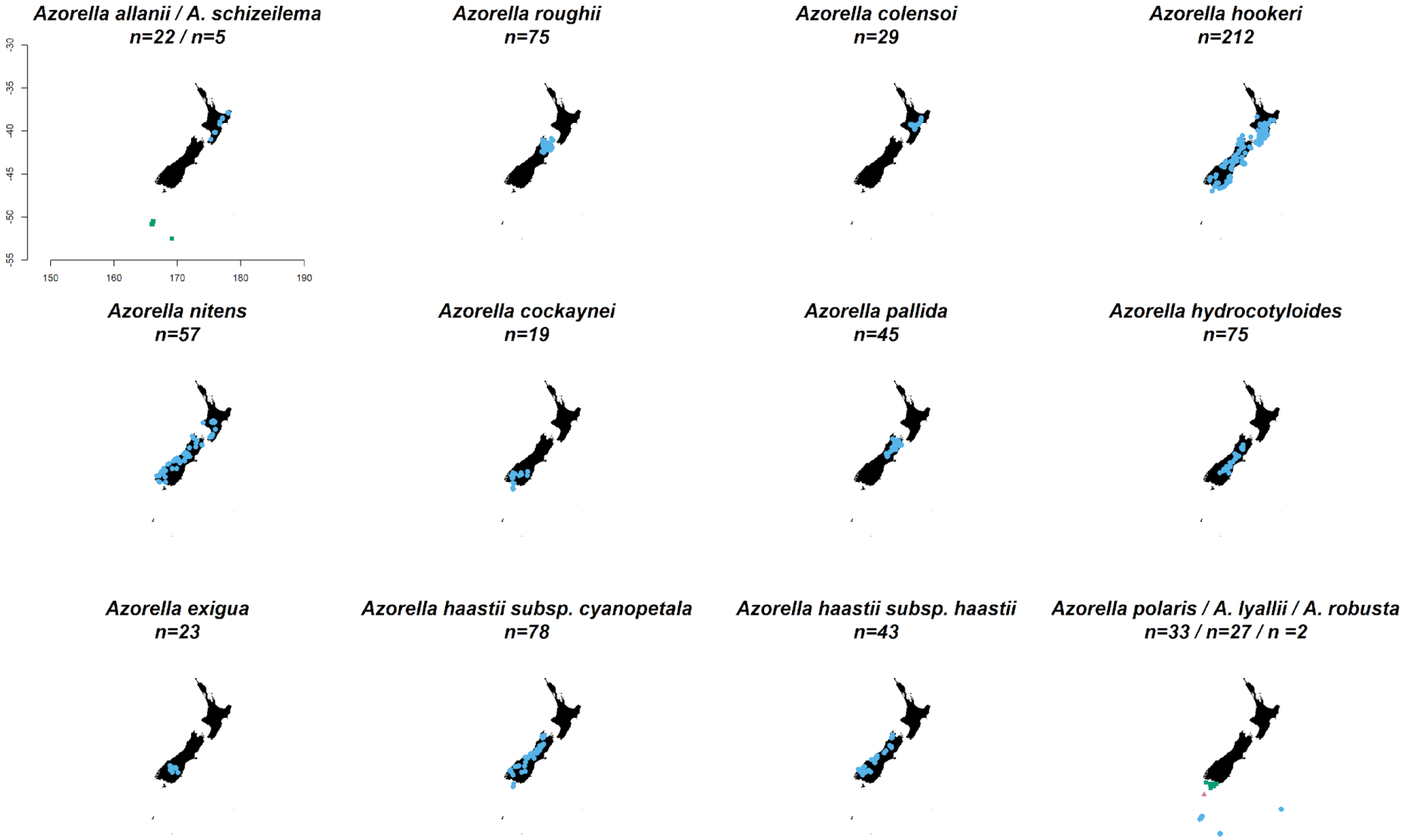


Figure S2. Distribution range of 13 *Azorella* species in section *Schizeilema* (*A. hookeri* included both varieties) and three subantarctic species in section

Stilbocarpa. The first and last map have multiple species included, each colour and shape represents one species, i.e., *A. allanii* (blue circle), *A. schizeilema* (green square), *A. polaris* (blue circle), *A. lyallii* (green square), and *A. robusta* (pink triangle). The geographical localities were collected from iNaturalist and two New Zealand herbaria, WELT and CHR (Chapter 3).

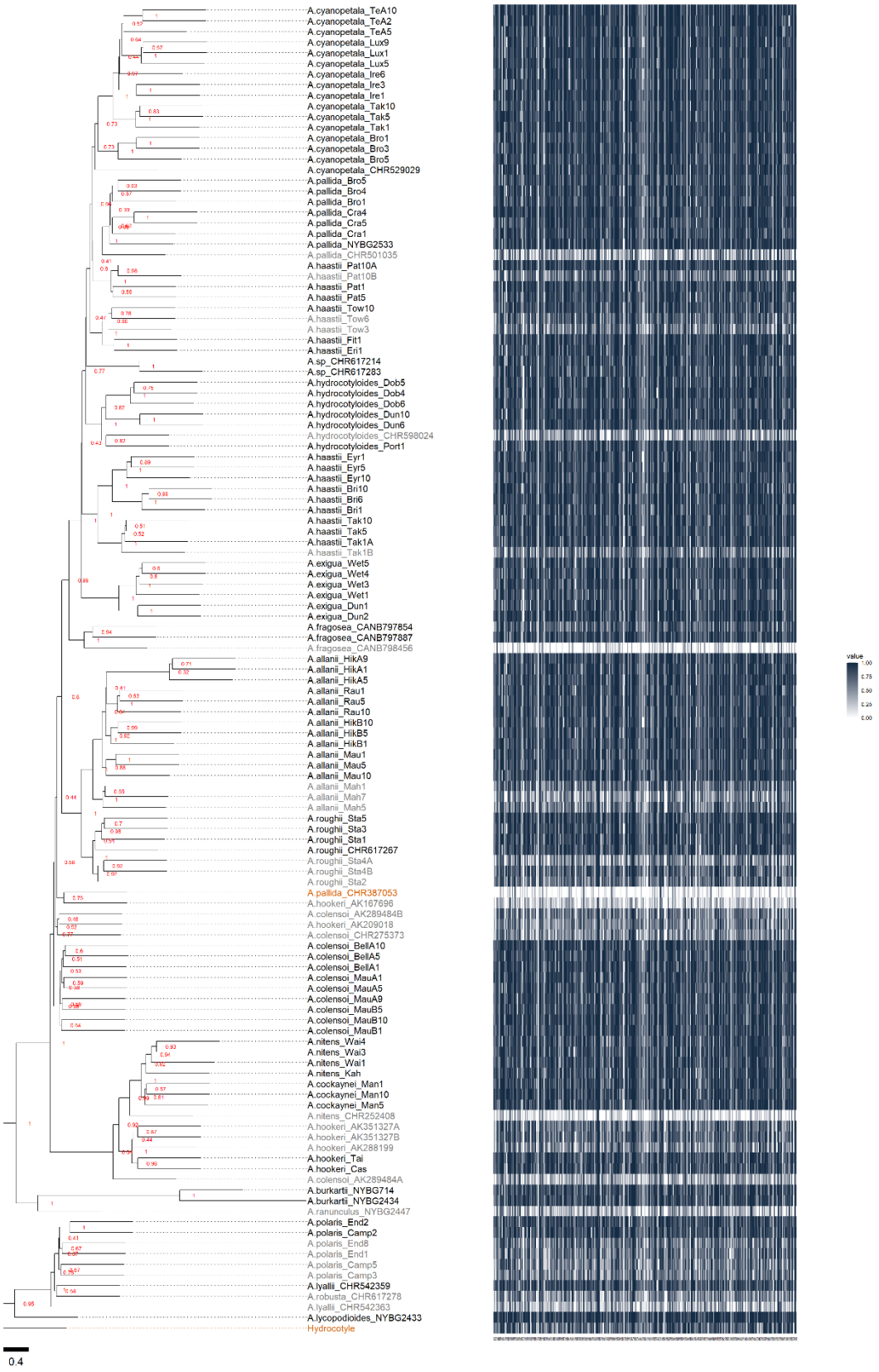


Figure S3. Phylogenomic ASTRAL tree for 125-individuals of *Azorella* constructed by 337 targeted single copy nuclear genes (SCNGs) from the Angiosperms353 bait set. Each node is annotated with a local posterior probability (max = 1). The tree tips (including the species names and collection site; species names refer to Table S2) are aligned with each row of the heatmap which shows the Hyb-Seq efficiency of the sample. The heatmap shows the recovery rate for each targeted gene with

colour gradient from 0 (white) to 1 (dark blue). Samples were sequenced by Hiseq and Miseq are annotated with black and grey colours of their ID, respectively. Two individuals (A.pallida_CHR387053 and *Hydrocotyle*) that were excluded from further analysis are annotated in orange.

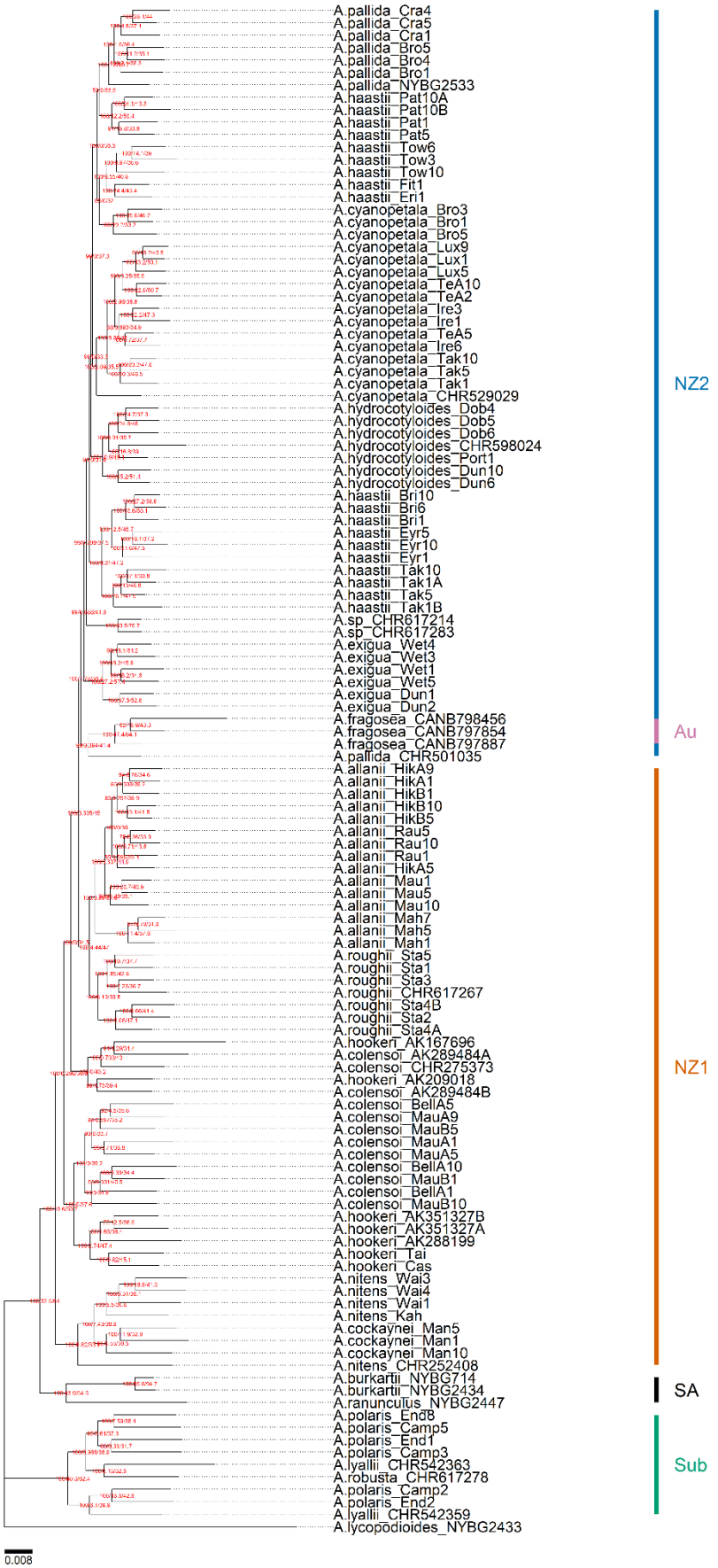


Figure S4. Phylogenetic inference of 123 *Azorella* individuals using 336 gene trees by the concatenation model in IQ-TREE2. The ID of individuals refers to the species names in Table S2.

Each node is annotated with a bootstrapping value (maximum = 100), the proportion of gene trees (maximum = 100%) and alignment sites (maximum = 100%) that are concordant with the topology of the tree. The five groups refer to identified groups in Fig 3: two New Zealand groups (NZ1 and NZ2), Australian (Au), South American (SA) and Subantarctic (Sub).

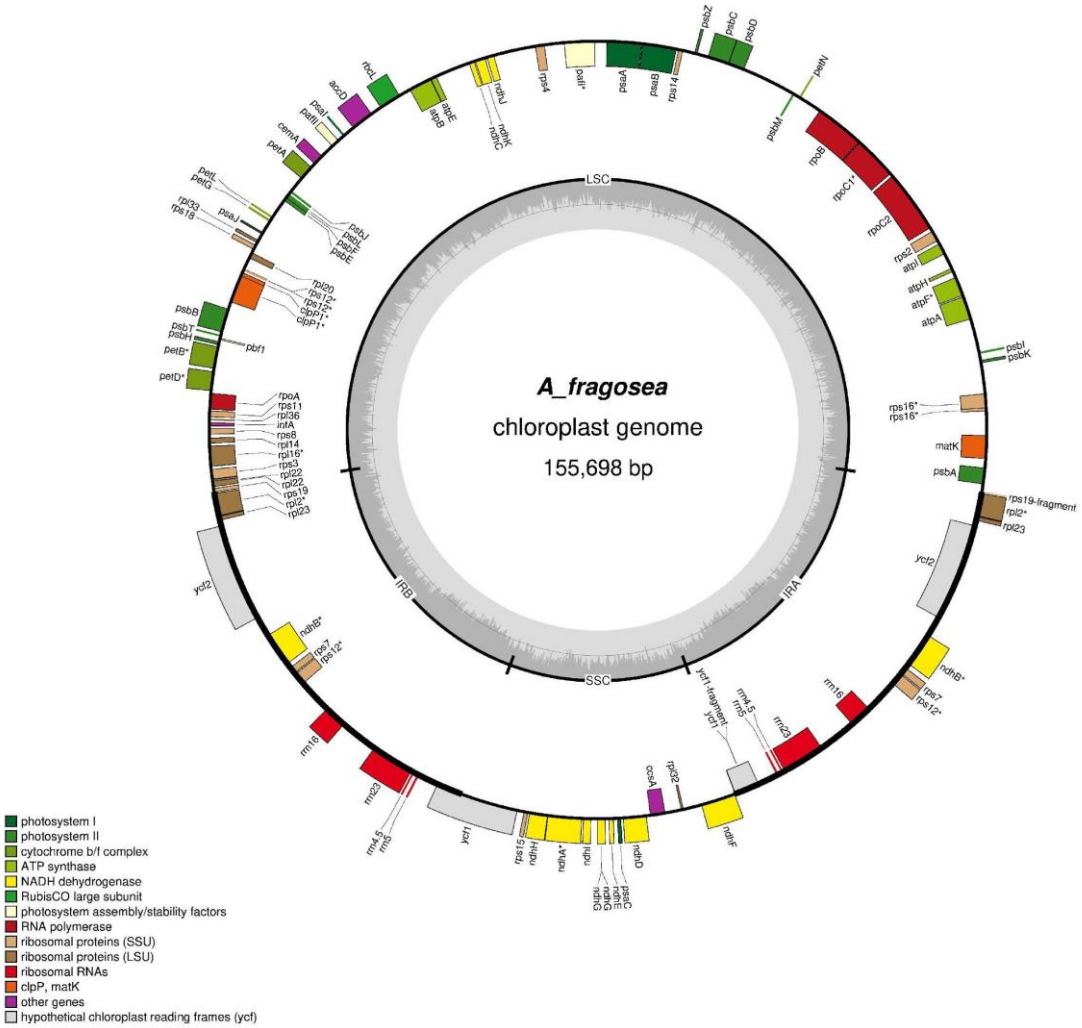


Figure S5. GeSeq annotated plastome structure of individual of the Australian endemic species, *Azorella fragosea* (CANB797887), in section *Schizeilema*, with 155,698 bp plastome extracted from genome-skimming reads. The annotations included the coding regions of genes, tRNA and rRNA among four junctions of the plastome, i.e., long single copy (LSC), short single copy (SSC), and two inverted repeats (IRA and IRB).

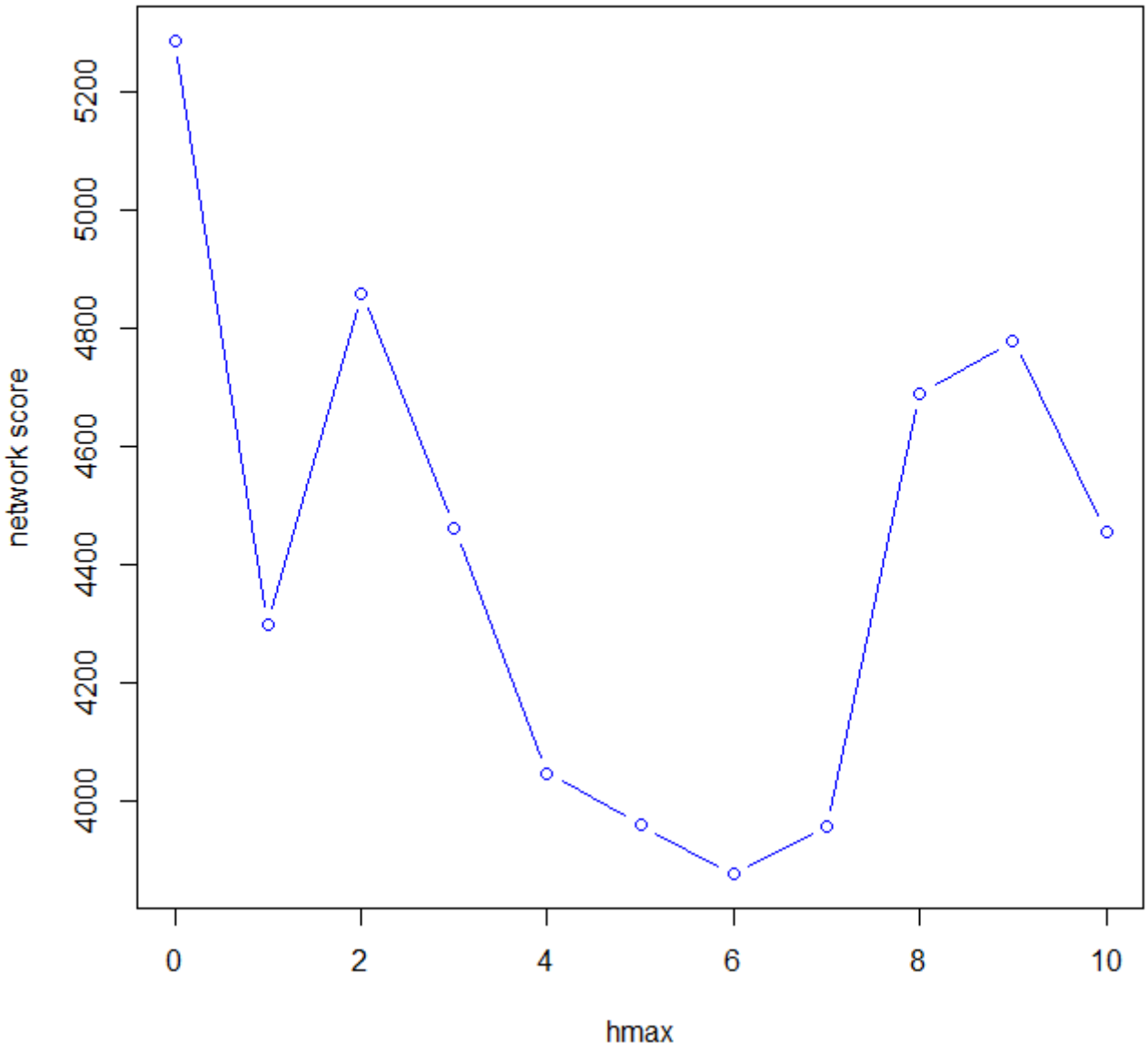


Figure S6. Selection of the best predicted network models estimated by SNaQ! based on network scores (i.e., y-axis shows the negative log-Pseudolikelihood scores). The x-axis shows ten network models that contained the hybridization events from 0 to 10 for 225 genes of 22 selected *Azorella* individuals. The model with the lowest network score was selected as the best model (i.e., hmax = 6).

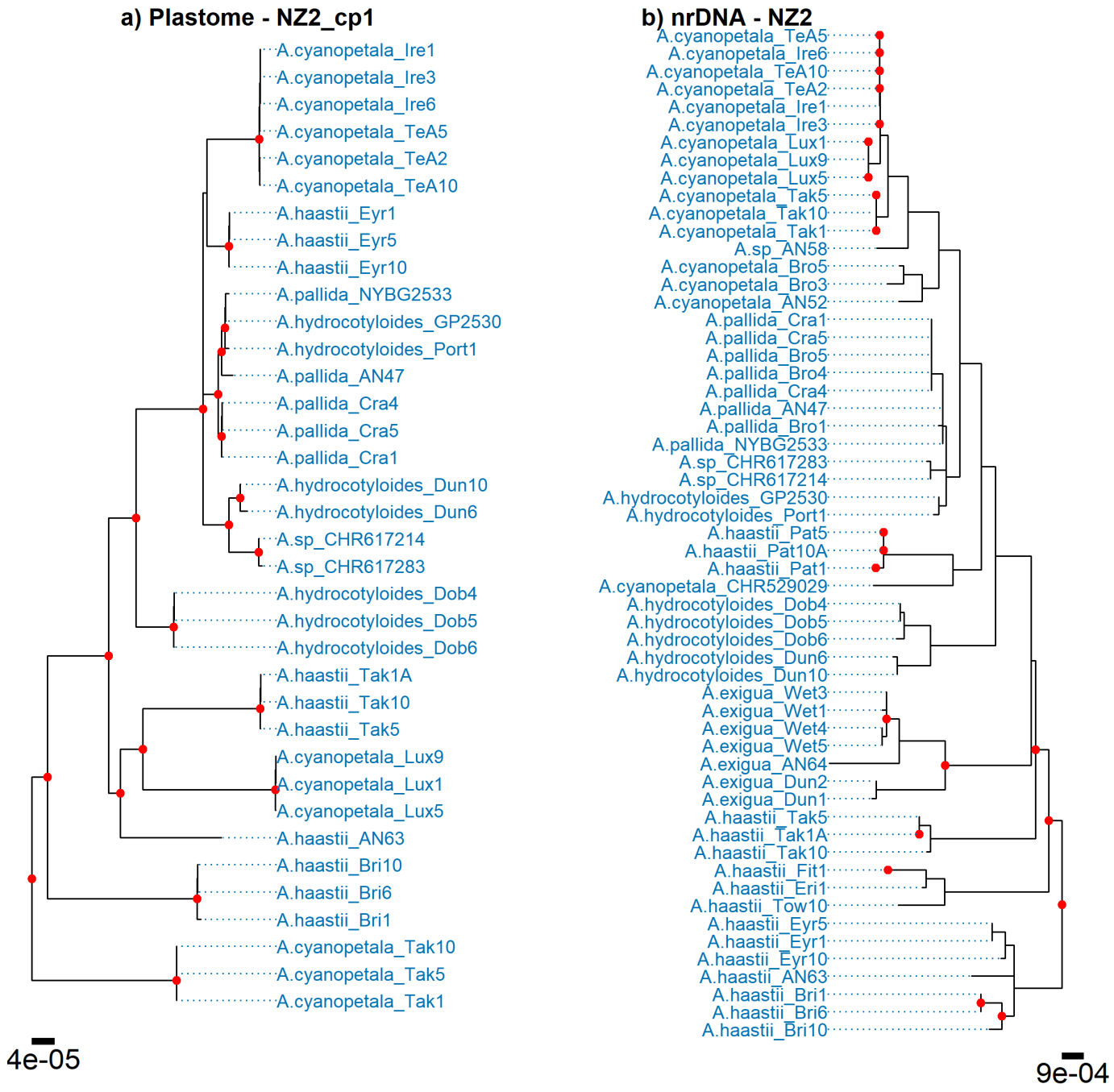


Figure S7. A subset of the plastome tree for NZ2_cp1 and nrDNA tree for NZ2 groups of New Zealand *Azorella* individuals in Fig. 4. The nodes with higher than 90% bootstrapping values are highlighted with red dots.

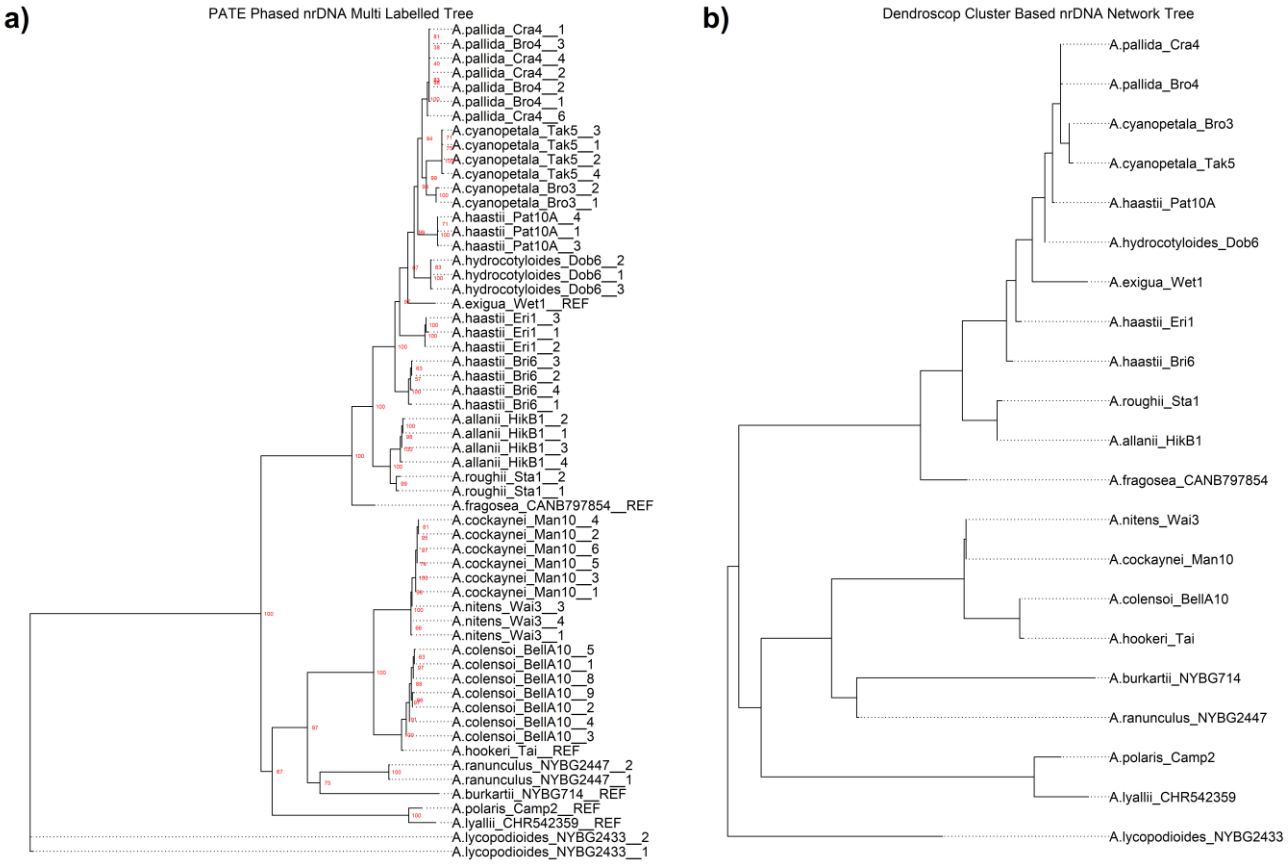


Figure S8. a) IQ-TREE2 estimated nrDNA multi-labelled species tree that was extracted and phased from genome-skimming data by GetOrganelle and PATÉ for selected 21 *Azorella* taxa. The individual sample name is followed by its extracted allele copy number variation (e.g., "__1"). Each node is annotated with a bootstrapping value (maximum = 100). b) Network inference of nrDNA multi labelled species tree for 21 individuals by Dendroscope with a cluster-based approach.

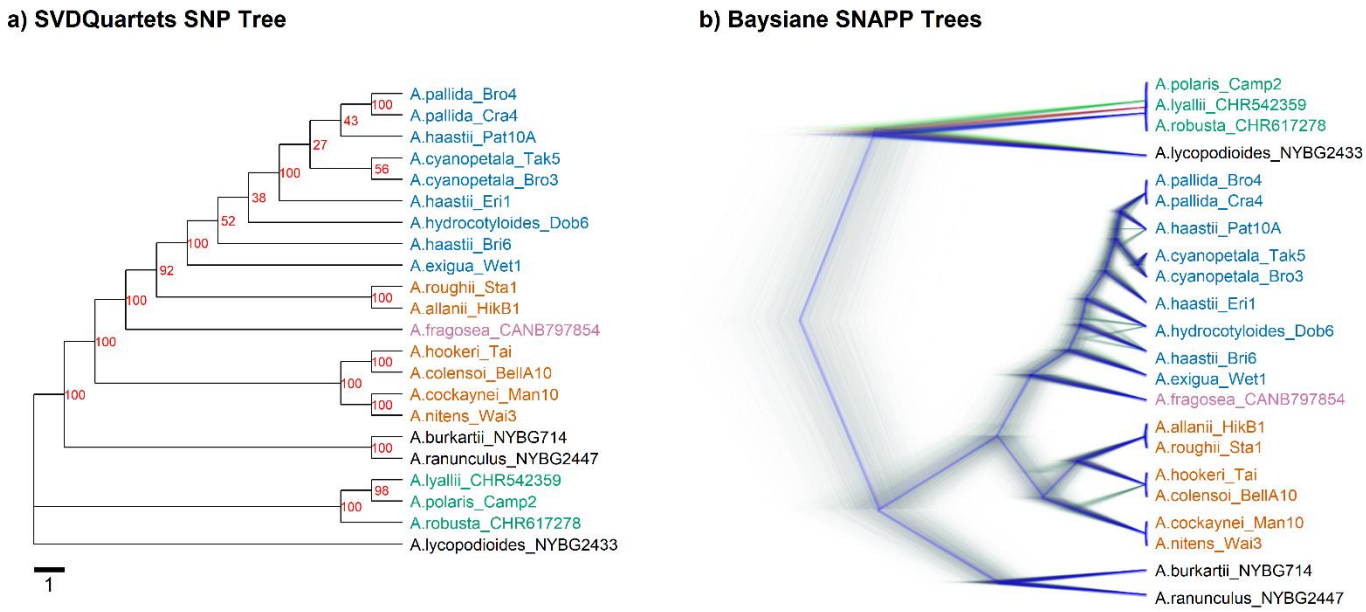


Figure S9. Phylogeny of the 22 representative individuals SVDQuartets tree and SNAPP trees for New Zealand *Azorella* and outgroups. a) Phylogenetic inference of 22 selected taxa relationships using 13,399 unlinked genomic SNPs by SVDQuartets. The nodes are labelled with bootstrapping values (maximum = 100). Individual IDs are coloured by their defined phylogenetic groups in Fig. 3. The species names refer to Table S2. b) The consensus SNAPP tree of Fig. 8a. c) The SNAPP trees of Fig 8c.

Supplementary Tables

Table S1. A summary of endemic *Azorella* section *Schizeilema* and *Stilbocarpa* species in Chile (no. 1 – 3), the subantarctic islands (no. 4 – 6), New Zealand (no. 7 – 19) and Australia (no. 20) (Plunkett & Nicolas, 2017). The 16 endemic taxa of *Azorella* in New Zealand and the subantarctic islands includes two subspecies in *A. haastii* (no. 15 – 16) and two varieties in *A. hookeri* (no. 10 – 11). Their geographic distribution (according to AK, CHR, MPN, OTA and WELT, herbarium records) and ploidal level (?*x* = unknown; from http://www.tropicos.org/project/ipcn). Taxonomy follows the New Zealand Plant Names Database (https://nzflora.landcareresearch.co.nz/).

NO.	Species	Section	Ploidal level	Chromosome number	Geographical distribution
1	*Azorella ranunculus* d'Urv.	*Ranunculus*	2*x*	16	South America
2	*Azorella burkartii* (Mathias & Constance) G.M.Plunkett &	*Ranunculus*	?*x*		South America
3	*Azorella lycopodioides* Gaud.	*Glabratae*	2*x*	16	South America
4	*Azorella polaris* (Hombr. & Jacquinot ex Hook.f.) G.M.Plunkett & A.N.Nicolas	*Stilbocarpa*	6*x*	48	Subantarctic islands
5	*Azorella robusta* (Kirk) G.M.Plunkett & A.N.Nicolas	*Stilbocarpa*	?		Subantarctic islands
6	*Azorella lyallii* (Armstr.) G.M.Plunkett & A.N.Nicolas	*Stilbocarpa*	?		Subantarctic islands
7	*Azorella roughii* (Hook.f.) Kirk	*Schizeilema*	4*x*	32	South Island
8	*Azorella allanii* (Cheeseman) G.M.Plunkett & A.N.Nicolas	*Schizeilema*	4*x*	32	North Island
9	*Azorella cockaynei* Diels	*Schizeilema*	6*x*	48	South Island
10	*Azorella hookeri* Drude var. *hookeri*	*Schizeilema*	6*x*	48	North & South Islands
11	*Azorella hookeri* var. *tripartita*	*Schizeilema*	6*x*	48	North & South Islands
12	*Azorella nitens* Petrie	*Schizeilema*	6*x*	48	North & South Islands
13	*Azorella colensoi* (Domin) G.M.Plunkett & A.N.Nicolas	*Schizeilema*	10*x*	80	North & South Islands
14	*Azorella hydrocotyloides* (Hook.f.) Kirk	*Schizeilema*	4*x*	32	South Island
15	*Azorella haastii* (Hook.f.) Drude subsp. *haastii*	*Schizeilema*	4*x*	32	South Island
16	*Azorella haastii* subsp. *cyanopetala* (Domin) G.M.Plunkett &	*Schizeilema*	4*x*	32	South Island
17	*Azorella schizeilema* G.M.Plunkett & A.N.Nicolas	*Schizeilema*	4*x*	32	Subantarctic islands

18	*Azorella exigua* (Hook.f.) Drude	*Schizeilema*	4*x*	32	South Island
19	*Azorella pallida* (Kirk) Kirk	*Schizeilema*	6*x*	48	South Island
20	*Azorella fragosea* (F.Muell.) Druce	*Schizeilema*	?		Australia

Table S2. Taxon sampling resulting of 136 individuals for target-enriched sequencing and genome-skimming sequencing for 18 *Azorella* taxa, two outgroup species, and two additional undescribed *Azorella spp.* (no. 133 and 134 collected from the same population, no. 135 was another geographically and morphologically diverged *A.* sp.). Taxon names represents the species names in Table S1, and the phylogenetic groups represents the predefined genetic groups based on phylogenomic relationships in 123-indivdiuals ASTRAL tree result (Fig. 3). The *Azorella* individual ID references to the taxon name and the sampling sources (i.e., field sites for field collected samples plus individual sequential number, and specimen ID or collection ID for herbaria sampled individuals). The individuals annotated without an asterisk represent only genome-skimming data available, or with only one asterisk (*) represent the Hyb-Seq data available (Angiosperms353 single-copy nuclear genes). Individuals labelled with two asterisks (**) represent those with both data sets available. The five individuals with biological replicated in Hyb-Seq are annotated with A and B (no. 22 & 23, 72 & 73, 76 & 77, 85 & 86, and 131 & 132). Platform refers to the sequencing platforms (Hiseq vs. Miseq) and sampling sources [field (F) vs. specimens (S)] of each individuals in Table 1, and the genome-skimming data of 13 individuals generated by G.P. and A.N. (unpubl. data) are annotated as G&A. The vouchers are deposited in MPN, NYBG, WLT, CANB and CHR herbaria.

NO	**Taxon Name**	**Individual ID**	**Phylogenetic Groups**	**Platform**	**Voucher ID**
1	*Azorella allanii*	A.allanii_AN45	N1	G&A	NYBG Nicolas45
2	*Azorella allanii*	A.allanii_HikA1**	N1	HiSeq_F	WELT SP108812
3	*Azorella allanii*	A.allanii_HikA5**	N1	HiSeq_F	WELT SP108812
4	*Azorella allanii*	A.allanii_HikA9**	N1	HiSeq_F	WELT SP108812
5	*Azorella allanii*	A.allanii_HikB1**	N1	HiSeq_F	WELT SP111286
6	*Azorella allanii*	A.allanii_HikB10**	N1	HiSeq_F	WELT SP111286
7	*Azorella allanii*	A.allanii_HikB5**	N1	HiSeq_F	WELT SP111286
8	*Azorella allanii*	A.allanii_Mah1*	N1	MiSeq_F	MPN 52524
9	*Azorella allanii*	A.allanii_Mah5*	N1	MiSeq_F	MPN 52524
10	*Azorella allanii*	A.allanii_Mah7*	N1	MiSeq_F	MPN 52524
11	*Azorella allanii*	A.allanii_Mau1**	N1	HiSeq_F	WELT SP110008

12	*Azorella allanii*	A.allanii_Mau10**	N1	HiSeq_F	WELT SP110008
13	*Azorella allanii*	A.allanii_Mau5*	N1	HiSeq_F	WELT SP110008
14	*Azorella allanii*	A.allanii_Rau1**	N1	HiSeq_F	WELT SP111287
15	*Azorella allanii*	A.allanii_Rau10**	N1	HiSeq_F	WELT SP111287
16	*Azorella allanii*	A.allanii_Rau5**	N1	HiSeq_F	WELT SP111287
17	*Azorella burkartii*	A.burkartii_NYBG2434**	SA	HiSeq_S	NYBG Plunkett2434
18	*Azorella burkartii*	A.burkartii_NYBG714**	SA	HiSeq_S	NYBG Ccalv714
19	*Azorella cockaynei*	A.cockaynei_Man1**	N1	HiSeq_F	MPN 52530
20	*Azorella cockaynei*	A.cockaynei_Man10**	N1	HiSeq_F	MPN 52530
21	*Azorella cockaynei*	A.cockaynei_Man5**	N1	HiSeq_F	MPN 52530
22	*Azorella colensoi*	A.colensoi_AK289484A*	N1	MiSeq_S	AK 289484
23	*Azorella colensoi*	A.colensoi_AK289484B*	N1	MiSeq_S	AK 289484
24	*Azorella colensoi*	A.colensoi_BellA1**	N1	HiSeq_F	WELT SP110028
25	*Azorella colensoi*	A.colensoi_BellA10**	N1	HiSeq_F	WELT SP110028
26	*Azorella colensoi*	A.colensoi_BellA5**	N1	HiSeq_F	WELT SP110028
27	*Azorella colensoi*	A.colensoi_CHR275373*	N1	MiSeq_S	CHR 275373
28	*Azorella colensoi*	A.colensoi_MauA1**	N1	HiSeq_F	WELT SP110011
29	*Azorella colensoi*	A.colensoi_MauA5**	N1	HiSeq_F	WELT SP110011
30	*Azorella colensoi*	A.colensoi_MauA9**	N1	HiSeq_F	WELT SP110011
31	*Azorella colensoi*	A.colensoi_MauB1**	N1	HiSeq_F	WELT SP110035
32	*Azorella colensoi*	A.colensoi_MauB10**	N1	HiSeq_F	WELT SP110035
33	*Azorella colensoi*	A.colensoi_MauB5**	N1	HiSeq_F	WELT SP110035
34	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_AN52	N2	G&A	NYBG Nicolas52
35	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Bro1*	N2	HiSeq_F	MPN 52529
36	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Bro3**	N2	HiSeq_F	MPN 52529
37	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Bro5**	N2	HiSeq_F	MPN 52529
38	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_CHR529029**	N2	HiSeq_S	CHR 529029
39	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Ire1**	N2	HiSeq_F	WELT SP108837
40	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Ire3**	N2	HiSeq_F	WELT SP108837
41	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Ire6**	N2	HiSeq_F	WELT SP108837
42	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Lux1**	N2	HiSeq_F	WELT SP108840
43	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Lux5**	N2	HiSeq_F	WELT SP108840

44	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Lux9**	N2	HiSeq_F	WELT SP108840
45	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Tak1**	N2	HiSeq_F	WELT SP108844
46	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Tak10**	N2	HiSeq_F	WELT SP108844
47	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_Tak5**	N2	HiSeq_F	WELT SP108844
48	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_TeA10**	N2	HiSeq_F	WELT SP108832
49	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_TeA2**	N2	HiSeq_F	WELT SP108832
50	*Azorella haastii* subsp. *cyanopetala*	A.cyanopetala_TeA5**	N2	HiSeq_F	WELT SP108832
51	*Azorella exigua*	A.exigua_AN64	N1	G&A	NYBG Nicolas64
52	*Azorella exigua*	A.exigua_Dun1**	N1	HiSeq_F	WELT SP111277
53	*Azorella exigua*	A.exigua_Dun2**	N1	HiSeq_F	WELT SP111277
54	*Azorella exigua*	A.exigua_Wet1**	N1	HiSeq_F	MPN 52528
55	*Azorella exigua*	A.exigua_Wet3**	N1	HiSeq_F	MPN 52528
56	*Azorella exigua*	A.exigua_Wet4**	N1	HiSeq_F	MPN 52528
57	*Azorella exigua*	A.exigua_Wet5**	N1	HiSeq_F	MPN 52528
58	*Azorella fragosea*	A.fragosea_CANB797854**	Au	HiSeq_S	CANB 797854
59	*Azorella fragosea*	A.fragosea_CANB797887**	Au	HiSeq_S	CANB 797887
60	*Azorella fragosea*	A.fragosea_CANB798456*	Au	MiSeq_S	CANB 798456
61	*Azorella fragosea*	A.fragosea_EI23518	Au	G&A	NYBG Eichler23518
62	*Azorella haastii* subsp. *haastii*	A.haastii_AN63	N2	G&A	NYBG Nicolas63
63	*Azorella haastii* subsp. *haastii*	A.haastii_Bri1**	N2	HiSeq_F	WELT SP111342
64	*Azorella haastii* subsp. *haastii*	A.haastii_Bri10**	N2	HiSeq_F	WELT SP111342
65	*Azorella haastii* subsp. *haastii*	A.haastii_Bri6**	N2	HiSeq_F	WELT SP111342
66	*Azorella haastii* subsp. *haastii*	A.haastii_Eri1**	N2	HiSeq_F	WELT SP107477
67	*Azorella haastii* subsp. *haastii*	A.haastii_Eyr1**	N2	HiSeq_F	WELT SP108816
68	*Azorella haastii* subsp. *haastii*	A.haastii_Eyr10**	N2	HiSeq_F	WELT SP108816
69	*Azorella haastii* subsp. *haastii*	A.haastii_Eyr5**	N2	HiSeq_F	WELT SP108816
70	*Azorella haastii* subsp. *haastii*	A.haastii_Fit1**	N2	HiSeq_F	WELT SP107484
71	*Azorella haastii* subsp. *haastii*	A.haastii_Pat1**	N2	HiSeq_F	WELT SP111326
72	*Azorella haastii* subsp. *haastii*	A.haastii_Pat10A**	N2	HiSeq_F	WELT SP111326
73	*Azorella haastii* subsp. *haastii*	A.haastii_Pat10B*	N2	MiSeq_F	WELT SP111326
74	*Azorella haastii* subsp. *haastii*	A.haastii_Pat5**	N2	HiSeq_F	WELT SP111326
75	*Azorella haastii* subsp. *haastii*	A.haastii_Tak10**	N2	HiSeq_F	WELT SP108848

76	*Azorella haastii* subsp. *haastii*	A.haastii_Tak1A**	N2	HiSeq_F	WELT SP108848
77	*Azorella haastii* subsp. *haastii*	A.haastii_Tak1B*	N2	MiSeq_F	WELT SP108848
78	*Azorella haastii* subsp. *haastii*	A.haastii_Tak5**	N2	HiSeq_F	WELT SP108848
79	*Azorella haastii* subsp. *haastii*	A.haastii_Tow10**	N2	HiSeq_F	WELT SP107453
80	*Azorella haastii* subsp. *haastii*	A.haastii_Tow3*	N2	MiSeq_F	WELT SP107453
81	*Azorella haastii* subsp. *haastii*	A.haastii_Tow6*	N2	MiSeq_F	WELT SP107453
82	*Azorella hookeri*	A.hookeri_AK167696*	N1	MiSeq_S	AK 167696
83	*Azorella hookeri*	A.hookeri_AK209018*	N1	MiSeq_S	AK 209018
84	*Azorella hookeri*	A.hookeri_AK288199*	N1	MiSeq_S	AK 288199
85	*Azorella hookeri*	A.hookeri_AK351327A*	N1	MiSeq_S	AK 351327
86	*Azorella hookeri*	A.hookeri_AK351327B*	N1	MiSeq_S	AK 351327
87	*Azorella hookeri*	A.hookeri_AN68	N1	G&A	NYBG Nicolas68
88	*Azorella hookeri*	A.hookeri_Cas**	N1	HiSeq_F	CHR 674155
89	*Azorella hookeri*	A.hookeri_Tai**	N1	HiSeq_F	MPN 52537
90	*Azorella hydrocotyloides*	A.hydrocotyloides_CHR598024*	N2	MiSeq_S	CHR 598024
91	*Azorella hydrocotyloides*	A.hydrocotyloides_Dob4**	N2	HiSeq_F	WELT SP106684
92	*Azorella hydrocotyloides*	A.hydrocotyloides_Dob5**	N2	HiSeq_F	WELT SP106684
93	*Azorella hydrocotyloides*	A.hydrocotyloides_Dob6**	N2	HiSeq_F	WELT SP106684
94	*Azorella hydrocotyloides*	A.hydrocotyloides_Dun10**	N2	HiSeq_F	WELT SP111280
95	*Azorella hydrocotyloides*	A.hydrocotyloides_Dun6**	N2	HiSeq_F	WELT SP111280
96	*Azorella hydrocotyloides*	A.hydrocotyloides_GP2530	N2	G&A	NYBG Plunkett2530
97	*Azorella hydrocotyloides*	A.hydrocotyloides_Port1**	N2	HiSeq_F	WELT SP107474
98	*Azorella lyallii*	A.lyallii_CHR542359**	Sub	HiSeq_S	CHR 542359
99	*Azorella lyallii*	A.lyallii_CHR542363*	Sub	MiSeq_S	CHR 542363
100	*Azorella lycopodioides*	A.lycopodioides_NYBG2433**	Outgroup	HiSeq_S	NYBG Plunkett2433
101	*Azorella nitens*	A.nitens_AN49	NZ1	G&A	NYBG Nicolas49
102	*Azorella nitens*	A.nitens_CHR252408*	NZ1	MiSeq_S	CHR 252408
103	*Azorella nitens*	A.nitens_Kah**	NZ1	HiSeq_F	MPN 52536
104	*Azorella nitens*	A.nitens_Wai1**	NZ1	HiSeq_F	MPN 52525
105	*Azorella nitens*	A.nitens_Wai3**	NZ1	HiSeq_F	MPN 52525
106	*Azorella nitens*	A.nitens_Wai4**	NZ1	HiSeq_F	MPN 52525
107	*Azorella pallida*	A.pallida_AN47	NZ2	G&A	NYBG Nicolas47

108	*Azorella pallida*	A.pallida_Bro1**	NZ2	HiSeq_F	MPN 52531
109	*Azorella pallida*	A.pallida_Bro4**	NZ2	HiSeq_F	MPN 52531
110	*Azorella pallida*	A.pallida_Bro5**	NZ2	HiSeq_F	MPN 52531
111	*Azorella pallida*	A.pallida_CHR387053*	NZ2	MiSeq_S	CHR 387053
112	*Azorella pallida*	A.pallida_CHR501035	NZ2	MiSeq_S	CHR 501035
113	*Azorella pallida*	A.pallida_Cra1**	NZ2	HiSeq_F	MPN 52532
114	*Azorella pallida*	A.pallida_Cra4**	NZ2	HiSeq_F	MPN 52532
115	*Azorella pallida*	A.pallida_Cra5**	NZ2	HiSeq_F	MPN 52532
116	*Azorella pallida*	A.pallida_NYBG2533**	NZ2	HiSeq_S	NYBG Plunkett2533
117	*Azorella polaris*	A.polaris_Camp2**	Sub	HiSeq_F	CHR 677157
118	*Azorella polaris*	A.polaris_Camp3*	Sub	MiSeq_F	CHR 677157
119	*Azorella polaris*	A.polaris_Camp5*	Sub	MiSeq_F	CHR 677157
120	*Azorella polaris*	A.polaris_End1*	Sub	MiSeq_F	CHR 677160
121	*Azorella polaris*	A.polaris_End2**	Sub	HiSeq_F	CHR 677160
122	*Azorella polaris*	A.polaris_End8*	Sub	MiSeq_F	CHR 677160
123	*Azorella ranunculus*	A.ranunculus_NYBG2447**	SA	MiSeq_S	NYBG Plunkett2447
124	*Azorella robusta*	A.robusta_CHR617278*	Sub	MiSeq_S	CHR 617278
125	*Azorella roughii*	A.roughii_AN67	NZ1	G&A	NYBG Nicolas67
126	*Azorella roughii*	A.roughii_CHR617267**	NZ1	HiSeq_S	CHR 617267
127	*Azorella roughii*	A.roughii_Sta1**c	NZ1	HiSeq_F	MPN 52523
128	*Azorella roughii*	A.roughii_Sta2*	NZ1	MiSeq_F	MPN 52523
129	*Azorella roughii*	A.roughii_Sta3**	NZ1	HiSeq_F	MPN 52523
130	*Azorella roughii*	A.roughii_Sta4A*	NZ1	MiSeq_F	MPN 52523
131	*Azorella roughii*	A.roughii_Sta4B*	NZ1	MiSeq_F	MPN 52523
132	*Azorella roughii*	A.roughii_Sta5**	NZ1	HiSeq_F	MPN 52523
133	*Azorella* sp.	A.sp_AN58	NZ2	G&A	NYBG Nicolas58
134	*Azorella* sp.	A.sp_CHR617214**	NZ2	HiSeq_S	CHR 617214
135	*Azorella* sp.	A.sp_CHR617283**	NZ2	HiSeq_S	CHR 617283
136	*Hydrocotyle*	Hydrocotyle*	Outgroup	MiSeq_F	WELT SP112281

Table S3. The target enrichment sequencing result of 125 *Azorella* individuals analysed by HybPiper and HybPhaser. Individual ID represented the species ID refers to Table S1. Platform refers to the sequencing platform as defined in Table1. The targeted percentage calculated by number of mapped reads divided by total number of reads produced. The average number of genes with exon assembled, the average exon recovery rate, mean exon length, and mean supercontigs length were calculated from HybPiper outputs The average allele divergence and number of genes with at least 2% allele divergence were estimated from HybPhaser output.

Individual ID	Platform	No. Reads	No. Reads Mapped	Targeted (%)	No. Gene with Exons	Average Exon Recovery Rate (%)	Average Exon length (bp)	Average Supercontigs Length (%)	Average Allele Divergence (%)	No. Gene with Allele Divergence > 2%	Paralog
A.allanii_HikA1	HiSeq_F	7,207,410	2,578,528	35.80%	341	82.45%	656	2261	2.42%	191	15
A.allanii_HikA5	HiSeq_F	9,331,499	3,355,715	36.00%	343	83.14%	670	2561	2.51%	206	15
A.allanii_HikA9	HiSeq_F	6,258,174	2,267,932	36.20%	344	81.88%	651	2333	2.48%	205	13
A.allanii_HikB1	HiSeq_F	5,694,488	2,299,540	40.40%	345	81.70%	652	2500	2.46%	201	15
A.allanii_HikB10	HiSeq_F	6,988,694	2,827,105	40.50%	344	80.29%	640	2179	2.47%	194	18
A.allanii_HikB5	HiSeq_F	6,295,934	2,311,089	36.70%	342	82.52%	659	2460	2.58%	223	15
A.allanii_Mah1	MiSeq_F	2,285,400	216,550	9.50%	322	70.29%	567	1772	1.93%	139	5
A.allanii_Mah5	MiSeq_F	1,511,917	169,571	11.20%	315	64.90%	531	1484	1.66%	107	3
A.allanii_Mah7	MiSeq_F	975,496	98,517	10.10%	285	59.43%	486	1282	1.53%	85	0
A.allanii_Mau1	HiSeq_F	5,693,172	2,288,414	40.20%	343	82.56%	653	2336	2.51%	207	13
A.allanii_Mau10	HiSeq_F	8,238,482	2,964,637	36.00%	345	82.50%	658	2450	2.61%	213	18
A.allanii_Mau5	HiSeq_F	4,242,124	1,562,11	36.80%	342	81.03%	648	2219	2.46%	201	14

			4								
A.allanii_Rau1	HiSeq_F	5,912,839	2,177,156	36.80%	343	82.30%	657	2506	2.57%	209	16
A.allanii_Rau10	HiSeq_F	8,077,465	2,863,490	35.50%	344	81.37%	649	2263	2.52%	204	15
A.allanii_Rau5	HiSeq_F	7,045,539	2,692,505	38.20%	344	82.52%	656	2358	2.51%	212	17
A.burkartii_NYBG2434	HiSeq_S	7,178,809	1,661,778	23.10%	344	80.88%	638	2227	1.01%	53	4
A.burkartii_NYBG714	HiSeq_S	5,490,255	1,278,911	23.30%	345	79.98%	634	2202	1.12%	65	3
A.cockaynei_Man1	HiSeq_F	8,399,458	2,675,257	31.90%	343	81.75%	653	2423	4.39%	307	23
A.cockaynei_Man10	HiSeq_F	9,886,992	2,442,927	24.70%	348	83.54%	662	2694	4.47%	314	16
A.cockaynei_Man5	HiSeq_F	7,795,018	2,562,764	32.90%	341	83.14%	666	2497	4.47%	304	25
A.colensoi_AK289484A	MiSeq_S	971,544	304,952	31.40%	282	46.85%	375	792	2.05%	113	1
A.colensoi_AK289484B	MiSeq_S	1,883,954	249,165	13.20%	320	62.00%	504	1273	2.94%	211	1
A.colensoi_BellA1	HiSeq_F	10,522,755	4,480,463	42.60%	345	84.18%	669	2467	4.27%	315	12
A.colensoi_BellA10	HiSeq_F	15,131,296	6,155,744	40.70%	349	83.93%	667	2581	4.42%	326	11
A.colensoi_BellA5	HiSeq_F	10,590,220	4,565,383	43.10%	346	82.48%	652	2227	4.35%	321	14
A.colensoi_CHR275373	MiSeq_S	1,402,255	321,207	22.90%	304	57.57%	467	1080	2.33%	153	1
A.colensoi_MauA1	HiSeq_F	12,239,711	4,189,401	34.20%	343	83.77%	670	2590	4.46%	320	16
A.colensoi_MauA5	HiSeq_F	7,454,459	2,990,702	40.10%	347	81.59%	650	2331	4.40%	324	14
A.colensoi_MauA9	HiSeq_F	9,185,621	3,857,572	42.00%	345	79.90%	633	2049	4.35%	320	11

A.colensoi_MauB1	HiSeq_F	7,974,598	2,619,540	32.80%	345	83.57%	661	2509	4.33%	314	18
A.colensoi_MauB10	HiSeq_F	8,438,268	3,306,634	39.20%	343	83.70%	670	2588	4.30%	315	13
A.colensoi_MauB5	HiSeq_F	6,710,088	3,230,025	48.10%	346	80.85%	642	2192	4.15%	314	11
A.cyanopetala_Bro1	HiSeq_F	6,142,717	2,329,335	37.90%	345	82.56%	654	2276	2.46%	198	9
A.cyanopetala_Bro3	HiSeq_F	9,045,119	3,109,006	34.40%	344	84.73%	672	2648	2.55%	204	12
A.cyanopetala_Bro5	HiSeq_F	7,754,509	3,133,362	40.40%	345	82.46%	659	2268	2.39%	190	4
A.cyanopetala_CHR529029	HiSeq_S	7,953,148	3,131,028	39.40%	344	85.20%	675	2568	2.37%	186	14
A.cyanopetala_Ire1	HiSeq_F	9,384,712	2,783,477	29.70%	345	86.54%	687	2939	2.44%	192	14
A.cyanopetala_Ire3	HiSeq_F	8,581,281	3,044,007	35.50%	345	83.18%	666	2552	2.43%	194	13
A.cyanopetala_Ire6	HiSeq_F	7,524,251	2,684,421	35.70%	342	84.77%	677	2707	2.47%	203	9
A.cyanopetala_Lux1	HiSeq_F	11,213,603	3,544,247	31.60%	343	84.91%	678	2736	2.57%	202	8
A.cyanopetala_Lux5	HiSeq_F	9,893,115	3,222,126	32.60%	345	85.59%	681	2853	2.50%	199	13
A.cyanopetala_Lux9	HiSeq_F	9,944,726	3,478,949	35.00%	340	85.26%	686	2734	2.49%	196	10
A.cyanopetala_Tak1	HiSeq_F	10,328,205	3,596,293	34.80%	345	86.84%	690	2783	2.59%	206	13
A.cyanopetala_Tak10	HiSeq_F	6,021,896	2,439,031	40.50%	345	83.20%	664	2359	2.50%	198	12
A.cyanopetala_Tak5	HiSeq_F	8,320,635	3,399,659	40.90%	347	84.53%	667	2507	2.56%	199	13
A.cyanopetala_TeA10	HiSeq_F	13,635,971	3,995,081	29.30%	346	87.87%	698	2982	2.58%	203	9
A.cyanopetala_TeA2	HiSeq_F	10,981,67	3,242,22	29.50%	344	88.02%	701	3010	2.56%	203	8

		4	4								
A.cyanopetala_TeA5	HiSeq_F	8,718,587	3,346,036	38.40%	344	85.26%	677	2551	2.49%	201	11
A.exigua_Dun1	HiSeq_F	6,113,540	1,849,815	30.30%	339	82.10%	661	2612	2.43%	191	12
A.exigua_Dun2	HiSeq_F	13,524,116	3,829,223	28.30%	343	85.55%	683	2907	2.54%	192	7
A.exigua_Wet1	HiSeq_F	5,333,806	2,280,699	42.80%	344	77.56%	618	1857	2.23%	168	5
A.exigua_Wet3	HiSeq_F	8,520,453	2,716,676	31.90%	343	80.87%	644	2177	2.30%	180	8
A.exigua_Wet4	HiSeq_F	5,040,266	1,730,365	34.30%	342	82.42%	654	2411	2.46%	199	12
A.exigua_Wet5	HiSeq_F	5,913,245	2,450,874	41.40%	341	80.37%	640	2148	2.34%	184	8
A.fragosea_CANB797854	HiSeq_S	2,608,092	1,570,584	60.20%	339	72.32%	580	1536	1.84%	137	2
A.fragosea_CANB797887	HiSeq_S	7,391,356	3,206,594	43.40%	344	81.48%	648	2166	2.10%	161	9
A.fragosea_CANB798456	MiSeq_S	113,129	15,323	13.50%	128	32.58%	267	499	0.59%	11	0
A.haastii_Bri1	HiSeq_F	6,646,516	2,636,149	39.70%	343	81.96%	657	2365	2.15%	172	10
A.haastii_Bri10	HiSeq_F	6,877,983	2,309,246	33.60%	344	82.61%	665	2563	2.07%	171	13
A.haastii_Bri6	HiSeq_F	8,089,561	2,473,713	30.60%	346	84.89%	675	2785	2.21%	182	10
A.haastii_Eri1	HiSeq_F	6,496,656	2,280,802	35.10%	346	83.59%	667	2584	2.47%	195	8
A.haastii_Eyr1	HiSeq_F	8,151,797	3,366,628	41.30%	341	83.08%	664	2454	2.50%	203	7
A.haastii_Eyr10	HiSeq_F	7,196,362	2,829,320	39.30%	344	81.73%	652	2275	2.38%	201	9
A.haastii_Eyr5	HiSeq_F	7,122,922	2,963,987	41.60%	345	82.49%	653	2314	2.38%	199	9

A.haastii_Fit1	HiSeq_F	8,207,037	2,617,951	31.90%	343	84.78%	671	2612	2.59%	197	8
A.haastii_Pat1	HiSeq_F	9,371,956	3,147,165	33.60%	345	84.10%	670	2620	2.52%	197	8
A.haastii_Pat10A	HiSeq_F	7,928,911	2,617,991	33.00%	346	84.98%	677	2624	2.53%	208	9
A.haastii_Pat10B	MiSeq_F	501,293	175,922	35.10%	321	67.85%	548	1586	1.82%	120	2
A.haastii_Pat5	HiSeq_F	6,253,263	2,172,658	34.70%	342	81.93%	656	2298	2.48%	191	8
A.haastii_Tak10	HiSeq_F	7,635,295	2,322,357	30.40%	345	82.92%	664	2617	2.19%	187	7
A.haastii_Tak1A	HiSeq_F	7,330,229	2,766,086	37.70%	344	83.07%	662	2462	2.34%	194	6
A.haastii_Tak1B	MiSeq_F	391,264	143,939	36.80%	319	64.25%	515	1457	1.41%	82	1
A.haastii_Tak5	HiSeq_F	7,342,767	2,480,066	33.80%	344	84.17%	671	2514	2.36%	192	10
A.haastii_Tow10	HiSeq_F	6,650,900	2,613,419	39.30%	346	79.99%	637	2182	2.52%	202	6
A.haastii_Tow3	MiSeq_F	1,264,472	147,273	11.60%	312	66.21%	537	1561	1.57%	104	1
A.haastii_Tow6	MiSeq_F	2,970,493	294,026	9.90%	334	72.54%	586	1728	1.99%	146	3
A.hookeri_AK167696	MiSeq_S	781,491	398,358	51.00%	230	36.07%	291	545	1.45%	64	0
A.hookeri_AK209018	MiSeq_S	1,532,707	296,136	19.30%	317	59.14%	473	1153	2.76%	205	1
A.hookeri_AK288199	MiSeq_S	2,170,247	946,746	43.60%	313	61.05%	489	1267	2.34%	157	3
A.hookeri_AK351327A	MiSeq_S	3,094,588	560,316	18.10%	326	67.53%	541	1550	2.87%	217	4
A.hookeri_AK351327B	MiSeq_S	3,988,492	890,013	22.30%	334	69.70%	555	1584	3.05%	226	4
A.hookeri_Cas	HiSeq_F	10,511,93	5,575,38	53.00%	344	81.03%	643	2357	3.66%	289	19

		5	7								
A.hookeri_Tai	HiSeq_F	11,322,493	4,967,948	43.90%	346	81.18%	644	2293	3.66%	291	17
A.hydrocotyloides_CHR598024	MiSeq_S	486,657	65,798	13.50%	278	53.85%	444	1046	1.26%	70	0
A.hydrocotyloides_Dob4	HiSeq_F	9,719,564	3,159,072	32.50%	343	85.69%	681	2772	2.42%	196	11
A.hydrocotyloides_Dob5	HiSeq_F	6,482,959	2,473,555	38.20%	341	84.57%	675	2556	2.29%	183	9
A.hydrocotyloides_Dob6	HiSeq_F	7,352,859	2,893,500	39.40%	346	81.24%	648	2328	2.33%	183	14
A.hydrocotyloides_Dun10	HiSeq_F	7,590,135	2,922,354	38.50%	342	85.50%	681	2575	2.36%	192	13
A.hydrocotyloides_Dun6	HiSeq_F	7,636,291	3,277,961	42.90%	345	83.67%	664	2459	2.39%	201	12
A.hydrocotyloides_Port1	HiSeq_F	9,485,310	3,875,557	40.90%	344	82.50%	655	2240	2.39%	184	10
A.lyallii_CHR542359	HiSeq_S	7,305,950	2,103,376	28.80%	347	78.91%	622	2166	3.98%	261	29
A.lyallii_CHR542363	MiSeq_S	512,655	57,702	11.30%	259	46.87%	370	821	1.80%	90	0
A.lycopodioides_NYBG2433	HiSeq_S	11,502,425	2,969,625	25.80%	349	86.73%	685	2877	0.87%	48	9
A.nitens_CHR252408	MiSeq_S	460,523	272,215	59.10%	179	33.66%	272	486	1.24%	42	0
A.nitens_Kah	HiSeq_F	7,530,675	2,628,908	34.90%	342	80.93%	646	2341	4.38%	305	20
A.nitens_Wai1	HiSeq_F	7,490,220	2,990,859	39.90%	342	81.48%	651	2305	4.35%	305	23
A.nitens_Wai3	HiSeq_F	7,278,956	2,977,604	40.90%	343	81.43%	648	2438	4.35%	305	27
A.nitens_Wai4	HiSeq_F	7,045,077	3,088,842	43.80%	343	80.18%	644	2293	4.18%	295	25
A.pallida_Bro1	HiSeq_F	6,774,339	2,663,019	39.30%	341	82.12%	657	2233	2.55%	205	7

A.pallida_Bro4	HiSeq_F	7,048,664	2,624,600	37.20%	343	84.27%	671	2668	2.50%	199	12
A.pallida_Bro5	HiSeq_F	6,692,643	2,737,344	40.90%	346	80.99%	647	2159	2.43%	196	8
A.pallida_CHR501035	MiSeq_S	418,718	60,622	14.50%	259	54.68%	452	1107	1.31%	56	0
A.pallida_Cra1	HiSeq_F	7,333,956	2,939,945	40.10%	344	82.12%	651	2235	2.55%	192	8
A.pallida_Cra4	HiSeq_F	11,064,377	4,137,466	37.40%	348	85.77%	681	2740	2.59%	207	7
A.pallida_Cra5	HiSeq_F	8,209,699	3,083,313	37.60%	343	83.32%	664	2351	2.61%	207	8
A.pallida_NYBG2533	HiSeq_S	13,913,599	5,531,719	39.80%	342	86.60%	691	2797	2.53%	200	7
A.polaris_Camp2	HiSeq_F	8,900,935	2,717,141	30.50%	344	80.36%	641	2450	3.88%	256	30
A.polaris_Camp3	MiSeq_F	2,365,826	206,165	8.70%	322	61.68%	494	1276	2.79%	189	6
A.polaris_Camp5	MiSeq_F	2,213,264	199,476	9.00%	321	62.16%	496	1294	2.87%	195	7
A.polaris_End1	MiSeq_F	1,924,716	172,751	9.00%	322	67.06%	538	1548	2.81%	189	7
A.polaris_End2	HiSeq_F	8,375,509	2,463,611	29.40%	344	81.35%	642	2407	3.88%	256	28
A.polaris_End8	MiSeq_F	1,756,915	138,181	7.90%	320	66.10%	531	1553	2.83%	183	11
A.ranunculus_NYBG2447	MiSeq_S	656,067	87,720	13.40%	249	45.80%	377	819	0.73%	33	0
A.robusta_CHR617278	MiSeq_S	1,404,254	109,492	7.80%	307	58.76%	470	1228	2.29%	154	4
A.roughii_CHR617267	HiSeq_S	5,781,052	2,530,703	43.80%	342	81.40%	653	2469	2.29%	194	11
A.roughii_Sta1	HiSeq_F	7,980,190	2,754,457	34.50%	341	83.25%	666	2489	2.62%	213	12
A.roughii_Sta2	MiSeq_	1,830,513	271,600	14.80%	318	68.19%	553	1628	1.95%	132	4

	F										
A.roughii_Sta3	HiSeq_F	8,419,131	2,674,847	31.80%	340	83.59%	668	2562	2.71%	211	16
A.roughii_Sta4A	MiSeq_F	1,006,422	130,014	12.90%	284	60.60%	496	1280	1.55%	84	3
A.roughii_Sta4B	MiSeq_F	3,582,748	447,495	12.50%	328	73.26%	593	1819	2.06%	149	7
A.roughii_Sta5	HiSeq_F	7,309,778	2,540,176	34.80%	344	82.81%	661	2495	2.65%	219	9
A.sp_CHR617214	HiSeq_S	7,953,635	2,902,981	36.50%	345	82.66%	657	2361	2.04%	168	10
A.sp_CHR617283	HiSeq_S	6,157,379	2,599,609	42.20%	342	80.85%	639	2094	2.15%	178	7
A.pallida_CHR387053	MiSeq_S	206,590	69,210	33.50%	119	28.99%	244	466			0
Hydrocotyle	MiSeq_F	1,936,646	262,445	13.60%	331	72.75%	576	1801			22

Table S4. PATE output for nrDNA phasing result. Individual IDs with one asterisk (*) represent it was selected to reconstruct the multi-labelled nrDNA tree in Fig. S8. The nrDNA cistron length (bp), number of variable sites, and the heterozygosity level are calculated for each individual. Number of blocks per phased locus represent how many blocks in the references were used for genome-skimming reads phasing (ideally all should be as 1).

Individual ID	Length (bp)	Number of variants (bp)	Heterozygosity (%)	Number of blocks per phased locus
A.allanii_AN45	6979	17	0.24%	3
A.allanii_HikA1	5731	8	0.14%	3
A.allanii_HikA5	6980	12	0.17%	2
A.allanii_HikA9	6980	11	0.16%	2
A.allanii_HikB1*	6979	11	0.16%	2
A.allanii_HikB10	6981	14	0.20%	2
A.allanii_HikB5	6979	15	0.21%	2
A.allanii_Mau1	6982	13	0.19%	2
A.allanii_Mau10	5733	3	0.05%	1
A.allanii_Rau1	6979	8	0.11%	1
A.allanii_Rau10	6979	10	0.14%	2
A.allanii_Rau5	6979	13	0.19%	3
A.burkartii_NYBG2434	6936	3	0.04%	1
A.burkartii_NYBG714*	6027	2	0.00%	0
A.cockaynei_Man1	7056	10	0.14%	3
A.cockaynei_Man10*	7056	9	0.13%	3
A.cockaynei_Man5	7056	8	0.11%	3
A.colensoi_BellA1	6868	24	0.35%	3
A.colensoi_BellA10*	6868	20	0.29%	2
A.colensoi_BellA5	6868	23	0.33%	3
A.colensoi_MauA1	6868	12	0.17%	3
A.colensoi_MauA5	6868	11	0.16%	3
A.colensoi_MauA9	6868	5	0.07%	2
A.colensoi_MauB1	6868	5	0.07%	2
A.colensoi_MauB10	6868	34	0.50%	4
A.colensoi_MauB5	6868	19	0.28%	4

A.cyanopetala_AN52	6842	15	0.22%	3
A.cyanopetala_Bro3*	6911	10	0.14%	2
A.cyanopetala_Bro5	6911	7	0.10%	1
A.cyanopetala_CHR529029	7145	6	0.08%	2
A.cyanopetala_Ire1	7039	7	0.10%	1
A.cyanopetala_Ire3	7039	7	0.10%	1
A.cyanopetala_Ire6	7039	8	0.11%	1
A.cyanopetala_Lux1	7039	13	0.18%	2
A.cyanopetala_Lux5	7039	13	0.18%	2
A.cyanopetala_Lux9	7039	14	0.20%	3
A.cyanopetala_Tak1	6991	11	0.16%	2
A.cyanopetala_Tak10	6991	9	0.13%	2
A.cyanopetala_Tak5*	6991	10	0.14%	2
A.cyanopetala_TeA10	7039	6	0.09%	1
A.cyanopetala_TeA2	7039	8	0.11%	1
A.cyanopetala_TeA5	7039	7	0.10%	2
A.exigua_AN64	6953	7	0.10%	2
A.exigua_Dun1	6967	1	0.00%	0
A.exigua_Dun2	6969	2	0.00%	0
A.exigua_Wet1*	6905	3	0.00%	0
A.exigua_Wet3	6905	6	0.09%	1
A.exigua_Wet4	6972	1	0.00%	0
A.exigua_Wet5	6972	2	0.00%	0
A.fragosea_CANB797854*	6912	3	0.00%	0
A.fragosea_CANB797887	6917	2	0.00%	0
A.fragosea_EI23518	7000	5	0.07%	1
A.haastii_AN63	6946	12	0.17%	2
A.haastii_Bri1	7241	14	0.19%	3
A.haastii_Bri10	6974	10	0.14%	3
A.haastii_Bri6*	6972	14	0.20%	2
A.haastii_Eri1*	7078	9	0.13%	3
A.haastii_Eyr1	6956	19	0.27%	3

A.haastii_Eyr10	6968	14	0.20%	3
A.haastii_Eyr5	7235	8	0.11%	2
A.haastii_Fit1	7077	19	0.27%	3
A.haastii_Pat1	7027	8	0.11%	1
A.haastii_Pat10A*	7027	9	0.13%	3
A.haastii_Pat5	7027	6	0.09%	1
A.haastii_Tak10	7144	11	0.15%	3
A.haastii_Tak1A	7143	6	0.08%	1
A.haastii_Tak5	7143	9	0.13%	3
A.haastii_Tow10	7077	4	0.06%	1
A.hookeri_AN68	6869	10	0.15%	1
A.hookeri_Cas	6869	2	0.00%	0
A.hookeri_Tai*	6836	0	0.00%	0
A.hydrocotyloides_Dob4	6907	9	0.13%	2
A.hydrocotyloides_Dob5	7044	15	0.21%	2
A.hydrocotyloides_Dob6*	7362	6	0.08%	1
A.hydrocotyloides_Dun10	6975	5	0.07%	1
A.hydrocotyloides_Dun6	7255	8	0.11%	2
A.hydrocotyloides_GP2530	7105	4	0.06%	1
A.hydrocotyloides_Port1	6498	14	0.22%	2
A.lyallii_CHR542359*	6954	1	0.00%	0
A.lycopodioides_NYBG2433*	6806	9	0.13%	2
A.nitens_AN49	7056	8	0.11%	2
A.nitens_Kah	7056	8	0.11%	2
A.nitens_Wai1	7056	7	0.10%	1
A.nitens_Wai3*	7056	4	0.06%	1
A.nitens_Wai4	7056	7	0.10%	1
A.pallida_AN47	7036	4	0.06%	1
A.pallida_Bro1	6894	2	0.00%	0
A.pallida_Bro4*	7120	2	0.03%	1
A.pallida_Bro5	7120	6	0.08%	1
A.pallida_Cra1	7120	13	0.18%	4

A.pallida_Cra4*	7120	5	0.07%	1
A.pallida_Cra5	7120	3	0.04%	1
A.pallida_NYBG2533	7120	2	0.00%	0
A.polaris_Camp2*	6952	1	0.00%	0
A.polaris_End2	6611	22	0.33%	2
A.ranunculus_NYBG2447*	6985	4	0.06%	1
A.roughii_AN67	6976	8	0.11%	1
A.roughii_CHR617267	6977	7	0.10%	1
A.roughii_Sta1*	6977	7	0.10%	1
A.roughii_Sta3	6977	9	0.13%	2
A.roughii_Sta5	6977	10	0.14%	1
A.sp_AN58	6853	14	0.20%	3
A.sp_CHR617214	7130	6	0.08%	1
A.sp_CHR617283	7130	4	0.06%	1

Table S5. Model selection for determining the most likely number of hybridization events in 225 gene trees for *Azorella* 22-individual PhyloNet.

Hybridization	**Log raito**	**EdgeNo**	**Parameter**	**NoGene**	**AICc**	**deltAIC**	**BIC**
0	-247664	42	42	225	495412	1304.565	495426.8
1	-247436	45	46	225	494964.7	857.2545	494980.9
2	-247250	48	50	225	494599.6	492.1679	494617.2
3	-247217	51	54	225	494541.1	433.6288	494560.1
4	-247156	54	58	225	494427.6	320.1744	494448.1
5	-247170	57	62	225	494463.1	355.6627	494484.9
6	-247060	60	66	225	494251.2	143.7321	494274.4
7	-247081	63	70	225	494302	194.5971	494326.7
8	-246980	66	74	225	**494107.5**	**0**	**494133.5**
9	-246996	69	78	225	494147.4	39.93793	494174.9

Abbversion: Pesudo log-likelihood ratio (Log raito); The number of predicted branch lengths (EdgeNo); The sum of edge number and perdicted number of hybridization events (Parameter); Number of input gene trees (NoGene); Akaike information criterion (AIC); Delta_AICc (deltAIC); Bayesian information criterion (BIC).

Table S6. BioGeoBears model selection based on AICc weight value.

Model	**LnL**	**numparams**	**d**	**e**	**j**	**AICc**	**AICc_wt**
DEC	-43.62	2	0.016	0.015	0	91.9	0.24
DEC+J	-41.72	3	0.0053	1.00E-12	0.032	**90.86**	**0.4**
DIVALIKE	-53.45	2	0.015	0.014	0	111.6	1.30E-05
DIVALIKE+J	-44.55	3	0.0055	1.00E-12	0.042	96.51	0.024
BAYAREALIKE	-73.07	2	0.02	0.056	0	150.8	3.80E-14
BAYAREALIKE+J	-41.87	3	0.0025	1.00E-07	0.049	91.15	0.34

Abbversion: log-likelihood (LnL) and the Akaike information criterion (AIC); N, parameters number; d, dispersion rate; e, extinction rate; J.

Chapter 3. Diversification of the polyploid-rich genus *Azorella* (Apiaceae) in New Zealand

Abstract

After polyploidization, the duplicated genomic content in a cell provides additional opportunities for diversification of newly established polyploid species. Polyploid lineages that comprise closely related species at different ploidy levels (i.e., polyploid-rich genera) can provide useful models to investigate the species' post-polyploidization macroevolutionary patterns. In New Zealand, *Azorella* section *Schizeilema* has 14 described polyploid taxa (species, subspecies, and varieties) with three varying ploidy levels (4*x*, 6*x*, and 10*x*), diverse leaf morphologies, and distinct distributional ranges. To investigate the relationships and post-polyploidization diversification of taxa in section *Schizeilema*, we first reconstructed their phylogenetic relationships and used this tree to trace the evolution of ploidy-level associated traits. Including the outgroup species, the phylogeny of 20 individuals representing 15 polyploid *Azorella* taxa was reconstructed by employing PacBio-sequenced Hyb-Seq long reads that were captured using the Angiosperms353 bait set. Data for several polyploidy-associated traits were collected, including genome size [both monoploid (1Cx) and holoploid (2C) values], stomatal guard cell length, and homeologs copy number of target-captured genes (using Hyb-Seq captured genes). The ecological divergence of New Zealand *Azorella* species was also compared using environmental niche modelling results for 11 taxa in section *Schizeilema* using ENMtools. We aimed to examine 1) the evolutionary patterns of genome size and ecological niche partitioning within a phylogenetic context; and 2) whether the measured ploidy-associated traits are correlated with chromosomal level increases. Our results showed that PacBio Hyb-Seq reads can improve the recovery of targeted loci over short-read platforms by reducing the intron gaps between exons. For taxa in section *Schizeilema*, the holoploid genome sizes (which range from 4.32 pg to 14.6 pg) and homeologous gene copy number variation provided further evidence of both polyploidization and hybridization in the evolutionary history of taxa in this section. Stomatal guard cell length was not significantly correlated to 2C genome size and the observed variation suggests that species have experienced different adaption processes. Furthermore, different taxa in *Azorella* section *Schizeilema* exhibited divergent post-polyploidization patterns in 1Cx genome sizes. The comparison of niche space for taxa within the same ploidy level showed different niche shift patterns (e.g., niche conservatism, niche novelty, niche contraction). Overall, we found the diversification of polyploid species may relate to the origins of the species (including their subgenome donors), reticulate histories, niche shift, different post-polyploidization genomic changes, and the age of polyploid species.

Keywords: *Azorella*; Angiosperms353; Apiaceae; Ecological niche; Hyb-seq; Genome size; PacBio; Polyploidy; Reticulation; Stomatal traits; New Zealand.

3.1 Introduction

Organisms that have experienced polyploidization or whole genome duplication (WGD) in their evolutionary history can contain more than two sets of genomes. All flowering plants have had at least once ancient WGD event in their evolutionary history (Jiao et al., 2011b), and especially young polyploids (neopolyploids) make up to one-third of the diversity of extant vascular plant species (Wood et al., 2009). However, understanding the diversification patterns of polyploid species remains challenging, because complex evolutionary histories of polyploids and the various abiotic and biotic factors may drive their speciation (Van de Peer et al., 2021; Soltis et al., 2015; Clark and Donoghue, 2018; Soltis et al., 2009). A group of closely related polyploid species with different chromosome numbers and thus different ploidy levels (hereafter, polyploid-rich genera) can be useful models. Comparison of the divergence of polyploidy-associated traits (e.g., genome size) and ecological conditions may provide insights into the consequences of WGD on a macroevolutionary scale (Clark and Donoghue, 2018; Meudt et al., 2021).

Genome size refers to the total amount of DNA in a nucleus (C-value = pg of 2C nucleus) and is one of the important traits that correlates with WGD (Greilhuber et al., 2005; Leitch and Bennett, 2004). For a polyploid species, genome size is expected to be additive of its parental genome sizes, regardless of whether it is an allopolyploid (with different subgenomes, i.e., homeologous copies originating from different species) or an autopolyploid (duplication of the same genome set in one species, i.e., homologous copies) (Otto and Whitton, 2000; Leitch et al., 2008). However, post-WGD genomic modification processes, including large-scale chromosome-level reorganization, the insertion or deletion of tandem duplicates and transposable elements (Dodsworth et al., 2016; Wang et al., 2021c; Zuccolo et al., 2007; Leitch and Bennett, 2004), can alter the total genome size of polyploids. Moreover, these processes can be additionally affected by environmental conditions (Van de Peer et al., 2021; Qiu et al., 2019), such as temperature or precipitation, which may eventually lead to genome size contraction or further expansion.

In flowering plants, monoploid genome size (1Cx, amount of DNA in one set of chromosomes or a 1*x* genome) decreases with increasing ploidy level, i.e., genome downsizing is evident (Greilhuber et al., 2005; Leitch and Bennett, 2004). One of the explanations for DNA loss after the formation of polyploids is the process of diploidization that may occur after each round of polyploidization (Li et al., 2021). During the diploidization process, the deletion of transposable elements or incompatible (i.e., unequal or illegitimate) recombination between duplicated genomes can promote the DNA repair process (Devos et al., 2002; Petrov, 2002), so that the genome in polyploids can function as a more stable ‘diploid’ (reviewed by Soltis et al., 2015; Wang et al., 2021c).

The duplicated genomic content in polyploid genomes can further drive phenotypic divergence (te Beest et al., 2012; Leitch and Leitch, 2008). However, morphological traits in plant tissues or organs of polyploids can be difficult to predict (Knight and Beaulieu, 2008), because these traits are regulated by more complex biological networks, which include genomic, epigenomic and transcriptomic changes (reviewed by Chen, 2007; Leitch and Leitch, 2008; te Beest et al., 2012). By contrast, many cellular level traits may be correlated with total 2C genome sizes across angiosperms (Beaulieu et al., 2008; Francis et al., 2008). Stomatal guard cell length, epidermal cell size and cell cycle have been shown to be positively correlated with genome size, whereas stomatal density may be negatively correlated with genome size (Beaulieu et al., 2008).

Compared to their diploid parental species, polyploids can have more genomic heterogeneity and genomic plasticity, which may lead to larger phenotypic plasticity and higher stress tolerance (Van de Peer et al., 2021; Leitch and Leitch, 2008). Such plasticity changes may also allow sufficient time for neopolyploids to adapt to the initial selective forces, and following the post-WGD changes, this plasticity may eventually promote new habitat colonization and ecological niche differentiation of polyploids (Stebbins, 1985). Indeed, both auto- and allopolyploid plants generally tend to show wider habitat ranges compared to their diploid ancestors (Luque et al., 2022; Paape et al., 2020; te Beest et al., 2012). However, this trend may vary in different plant lineages (Visser and Molofsky, 2015). By contrast, niche evolution (i.e., niche shifts or niche conservatism) has no clear trend when comparing polyploids to diploids (Glennon et al., 2014; Marchant et al., 2016). Nonetheless, quantifying and comparing the niche space of closely related polyploids can help to clarify the role of environmental variables to promote species diversification (e.g., Hutchinson, 1957; López-Jurado et al., 2022; Moraes et al., 2022).

Insular endemic polyploid-rich plant genera provide ideal systems for investigating whether WGD is correlated with species macroevolutionary patterns (Soltis et al., 2009; Meudt et al., 2021). The New Zealand (NZ) polyploid-rich genus *Azorella* comprises two sections, *Schizeilema* and *Stilbocarpa,* with a total of 18 taxa (species, subspecies, and varieties). *Azorella* taxa in both sections have a perennial rhizomatous growth form and can reproduce vegetatively via stolons (Plunkett and Nicolas, 2017). Three megaherbs in section *Stilbocarpa*, namely *A. polaris* (6*x*) (Beuzenberg and Hair, 1983), *A. robusta* (ploidy unknown), and *A. lyallii* (ploidy unknown), can have large leaves up to 1 m across compared to the smaller rosette leaves of the 15 taxa in section *Schizeilema,* which have leaves 1 to 3 cm wide.

Species in *Azorella* section *Schizeilema* [ploidy levels of 4*x*, 6*x* and 10*x* (*x* = 8); (Hair, 1980)] are mostly found on the three main islands of New Zealand (i.e., North Island, South Island and Stewart

Island) compared to the megaherbs of section *Stilbocarpa* which are endemic to the NZ subantarctic islands and Stewart Island. The exceptions are *Azorella fragosea* (ploidy unknown), which is endemic to New South Wales, Australia, and *A. schizeilema* (4*x*), which is endemic to Auckland Islands and Campbell Islands, NZ. The NZ mainland species in section *Schizeilema* vary in their ploidy levels: one decaploid *A. colensoi*, three hexaploids (*A. hookeri*, *A. nitens*, and *A. cockaynei*) and several tetraploids (all remaining species, probably also including *A. pallida*, which was previously reported to be hexaploid; see Results) (Table S1).

In this study, 13 polyploid *Azorella* taxa from sections *Schizeilema* and *Stilbocarpa* were selected to infer the phylogenetic relationship of NZ *Azorella* and understand their post-WGD macroevolutionary patterns. The phylogeny of these 13 polyploid species was reconstructed using a Hyb-Seq approach (Weitemier et al., 2014) that can efficiently capture the conserved exons of selected loci (often as single copy nuclear genes) to resolve their evolutionary relationships. Combining Hyb-Seq with the Angiosperms353 bait kit (Johnson et al., 2018), which can capture up to 353 single copy nuclear genes in any flowering plant lineage, we used PacBio SMRT® sequencing platform (Pacific Biosciences) to improve the recovery efficiency of all gene copies over other short-read platforms (e.g., Illumina; See Results).

Here, we aim to answer the following questions: 1) Using PacBio sequenced longer Hyb-Seq reads, can we improve the recovery efficiency of polyploid homeologous copies of the targeted loci and the resolution and support of the resulting phylogeny? 2) Are polyploidy-associated traits, including genome size variation (2C), stomatal guard cell length and homeologous copy number variation among target-enriched genes, informative about the species origins or relationships of NZ *Azorella*? And 3) Can niche comparison (via environmental niche modelling) or post-WGD genome processes (i.e., downsizing or upsizing of 1Cx genome size) be informative about the macroevolutionary history of *Azorella* polyploids?

3.2 Materials and Methods

3.2.1 Taxon Sampling & Hyb-Seq Preparation for PacBio Sequencing

We re-sampled field-collected leaves dried on silica gel and herbaria material (CANB and NYBG) that had been sequenced previously in Chapter 2 (Table S1). In total, including the outgroups *A. burkartii* and *A. ranunculus* from *Azorella* section *Ranunculus* (Plunkett and Nicolas, 2017), 20 individuals representing 15 *Azorella* taxa were sampled for phylogenetic reconstruction of sequences generated using PacBio sequenced Hyb-Seq data (Table S2). For most taxa, only one

individual was sampled, except for *A. haastii* subsp. *haastii* (three individuals representing their previously identified three paraphyletic groups), and *A. haastii* subsp. *cyanopetala* (two individuals representing two different plastome groups) (Chapter 2). In addition, for *A. hookeri* and *A. nitens*, which are distributed on both the North and South Islands, one individual of each species from each island was included.

After extracting DNA from leaf tissue using the DNeasy® Plant Mini Kit (QIAGEN), DNA integrity levels were visualized on 1% agarose gels and concentrations were measured using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific). We constructed genomic libraries using the NEBNext® UltraTM II Library Prep kit (New England Biolabs), slightly modifying the manufacturer's protocol and using a customized bead-based size selection step (Stortchevoi et al., 2020) to select DNA fragments ranging from 1 kbp to 2 kbp long (Supplementary Notes S1). Each genomic library was barcoded with a pair of NEBNext® Multiplex Oligos for Illumina and its profile was checked on a LabChip® GX Touch™ nucleic acid analyzer (PerkinElmer). Next, all the DNA libraries were pooled with equimolar volumes into one batch prior to hybridization with the Angiosperms353 universal baits (Johnson et al., 2018) following the myBaits® Kit Manual V5. The genomic library pool was purified using AMPure PB® beads to remove DNA fragments less than 500 bp, and SMRTbell library construction was accomplished using the SMRTbell Express Template Prep Kit 2.0. The final genomic library pool was sequenced on a PacBio SMRT Cell 8M (Australian Genome Research Facility, Brisbane, Australia).

3.2.2 Recovery Rates of Targeted Genes & Phylogenetic Reconstruction

PacBio circular consensus sequencing reads were demultiplexed into individual sample reads using barcoding information in LIMA v. 2.6.0 (https://lima.how/; PacBio, 2021) (Supplementary Note S2), and the read length distributions were calculated for each sample. For 18 PacBio sequenced individuals, their Illumina Hyb-Seq data were also available [from the same individual or from the same population; (Chapter 2)]. Illumina Hyb-Seq reads for the 18 replicated samples were downloaded from NCBI (PRJNA885464), including one extra sample, *A. lycopodioides*, which had only Illumina Hyb-Seq reads available and was selected as the outgroup for downstream analyses (Table S2). Next, the targeted supercontigs (exons, introns and flanking regions) were extracted using PacBio Hyb-Seq data and Illumina Hyb-Seq data, respectively, using HybPiper v. 2 (Johnson et al., 2016) (Supplementary Notes S3). Furthermore, the topologies of reconstructed ASTRAL trees (Mirarab et al., 2014) and the concordance levels of gene trees were compared for the PacBio dataset and the Illumina dataset (Notes S3). Finally, the two datasets (i.e., Illumina and PacBio

reads) were merged to increase sequencing depth and coverage for recovering single copy nuclear genes (Supplementary Notes S4).

The merged dataset (Note S4) was reanalysed using HybPiper to extract the supercontigs of target-enriched genes with adjusted parameter settings for 'paralog_min_length_percentage' set to 0.5 (instead of the default, 0.75) to increase the chances of extracting the recovered homeologous sequences. After filtering for genes without stitched exons (Notes S4), the supercontigs and their extracted homeologous sequences were aligned and trimmed using the same methods and thresholds as above (Notes S3). Gene trees of each locus were reconstructed in IQ-TREE v. 2.2.0.5 (Minh et al., 2020b), and summarised into an ASTRAL tree with ASTRAL-PRO 2 (Zhang and Mirarab, 2022) with an additional supplied mapping file (-a) to allow alleles (homeologs) to be mapped to the same individual.

Gene copy number variation that represents the number of extracted homeologs per targeted locus were calculated for 189 selected loci and averaged for each of the 21 sequenced individuals (Notes S4). In addition, to determine whether low gene copy number variation was due to an insufficient number of sequenced reads, the supercontig recovery rate of the same set of selected genes was calculated (Notes S4).

3.2.3 Genome Size & Ploidy Level Variation

Genome sizes (2C values) of 28 individuals from 13 New Zealand *Azorella* species (including one *A. lyallii* that had no genetic sequencing data available) were measured using the CyFlow® Space flow cytometer at Manaaki Whenua – Landcare Research (Christchurch, New Zealand), with a modified Otto two-step method (Otto, 1990) that substituted DAPI with Propidium Iodine (PI) (Heenan et al., 2022). Two cultivated species with different genome sizes were selected as internal standards: *Pisum sativum* (pea; 2C = 8.80 pg) and *Vicia faba* [(broadbean; 2C = 25.64 pg) (Doležel et al., 1992)].

For each sampled individual, 1 cm^2 fresh leaf tissue (field-collected or greenhouse grown) was co-chopped with one of the selected standards in 1 mL ice-cold Otto buffer I with a stainless-steel razor blade, then incubated for 2 min on ice to release the nuclei. The supernatant was filtered through a 30 μm Celltrics filter (PARTEC GmbH) to remove the cell debris, and the released nuclei were stained with 2.5 ml of Otto buffer II that contained 1 mg/mL PI and incubated for 1 min. Each sample was measured using a CyFlow® Space with a 488 nm laser as the excitation source. Each sample was measured three times, and each measurement was validated by the two clear narrow peaks on a PI density plot showed in the result, which included 2000 to 10,000 counted nuclei. The

average coefficient of variation values (CV) was calculated for the measured samples and the co-chopped standards, respectively, with a threshold below 5% considered to be acceptable. Additional genome size data from the South American endemic species *A. burkartii* and *A. lycopodioides* from Ptáček et al. (2022) were also included in downstream analyses.

Published chromosome numbers for 14 of the sampled *Azorella* taxa are listed in Table S1 (note that the chromosome numbers of *A. burkartii* and *A. fragosea* are unknown). For New Zealand *Azorella* taxa, we confirmed the identification of the CHR herbarium chromosome vouchers of Hair (1980), and estimated the ploidy level of the samples by comparing the genome size measured here with their published chromosome numbers (see Results).

3.2.4 Genome Size Evolution

Phylogenetic signal (Pagel's λ based on Pagel, 1999) between the reconstructed ASTRAL-PRO tree and the measured genome sizes was calculated using 'phylosig' in the R package phytools (Revell, 2012). Pagel's λ can range between $0 < \lambda < 1$ representing different degrees of phylogenetic signal. When $\lambda = 0$ (no phylogenetic signal) the tested variable evolves independently from the phylogeny, whereas when $\lambda = 1$, a pure Brownian motion model is indicated (Felsenstein, 1973)] and the tested variable is highly dependent upon the phylogenetic relationships.

Ancestral monoploid genome size was estimated using maximum likelihood for a continuous trait using the 'ace.ml' function in the R package phytools with the Brownian motion evolution model. Tree tips that did not have genome size or ploidy level data were pruned from the ASTRAL-PRO tree prior to analyzing and plotting the estimated ancestral genome size values using the ggtree package in R (Yu et al., 2017).

3.2.5 Stomatal Guard Cell Length

The lower leaf epidermis from 50 silica-gel stored leaf samples and nine herbarium specimens were photographed using scanning electron microscopy (SEM) representing 18 *Azorella* taxa (including two taxa, *A. schizeilema* and *A. lyallii,* for which no sequencing data were available). The stomatal guard cell length was chosen for analysis because this a character that shows sensitivity to genome size changes (Beaulieu et al., 2008), and especially to WGD events (Masterson, 1994; Lattier et al., 2019). The measurements were taken on each image using the software ImageJ v.1.52 (Schindelin et al., 2012). Prior to pooling all measurements together for each species, the guard cell length of five individuals from five *Azorella* taxa that had both silica-gel stored and herbarium

specimen samples were compared to determine whether measurement results varied between the different sampling types (see Results).

One to three fully expanded leaves from different individuals of each population were selected and photographed. For each individual, and depending on the leaf size, images were taken at 200x magnification of one to four different flat areas on the leaf with evenly distributed stomata. Only *A. lycopodioides*, which has curled leaves, was photographed at 1000x. Multiple (20 to 30) guard cells were measured in each image, and these were averaged for each population and each taxon to allow comparison among individuals and taxa.

3.2.6 Extracting Georeferenced Records and Bioclimate Data

Georeferenced records of 16 *Azorella* taxa (Table S1) were extracted and downloaded from GBIF (https://www.gbif.org/) using the R package 'rgbif' (Chamberlain et al., 2022) in RStudio (RStudio Team, 2020; R Core Development Team, 2013). For New Zealand endemic *Azorella* taxa, additional herbarium specimen records from WELT and CHR herbarium databases and research-grade observations from iNaturalist (https://inaturalist.nz/) were downloaded. After removing duplicated records, the remaining records were filtered through the R package CoordinateCleaner (Zizka et al., 2019) to remove any that mapped to erroneous locations (e.g., in the ocean or an urban location). For New Zealand *Azorella* taxa, records were filtered for each taxon by their known distribution range (Plunkett and Nicolas, 2017) (Table S1) and collection year (removing those collected prior to 1800).

Bioclimate data in 30 arc seconds resolution (approx. 1 km^2) of 19 environmental layers and elevation (in meters) were downloaded from WorldClim 2 (Fick and Hijmans, 2017) for three regions: South America (Chile and Argentina), Australia (New South Wales), and New Zealand (including the New Zealand subantarctic islands), respectively. The elevation and 19 bioclimate data for the records of all 16 *Azorella* taxa were extracted from the downloaded environmental layers using the R package 'raster' (Hijmans and van Etten, 2012). We removed any georeferenced occurrence records that did not have any extracted climate data prior to averaging the remaining values for each taxon for downstream analysis.

3.2.7 Environmental (Ecological) Niche Modelling

Environmental niche models (ENM) can predict a species' potential geographic distribution based on the abiotic and biotic conditions from its known georeferenced occurrence data (reviewed in Bates and Bertelsmeier, 2021; Kearney and Porter, 2009). We performed ENM on the 11 taxa in

Azorella section *Schizeilema* endemic to mainland New Zealand (North, South and Steward Island). The Australian species, *A. fragosea*, and the more distantly related species in sections *Stilbocarpa* (from the New Zealand subantarctic islands) and *Ranunculus* (from South America), were not included because their chromosome numbers and ploidy levels are unknown (see Results). In addition, bioclimate data are not well resolved for the subantarctic island archipelagoes (e.g., Snares, Auckland, and Campbell Islands).

To perform ENM, first the correlation matrix of environmental layers (19 bioclimate layers plus elevation) was calculated using the 'raster.cor.matrix' function in the R package ENMtools (Warren et al., 2021), and correlated layers with a cutoff threshold of 0.75 were identified and removed using 'findCorrelation' as implemented in the R caret (Kuhn, 2012). Ecological niches were modelled using the MaxEnt model (i.e., maximum entropy modelling) in ENMtools using seven selected environmental layers (see Results). We set 70% of the georeferenced records as training data to construct the model and the remaining 30% were used as test data to assess the performance of the predicted model for each taxon. The area under the curve (AUC) of the receiver operating characteristic was calculated to measure the accuracy of the predicted species distribution results (Peterson et al., 2008). AUC ranges from 0.5, where the model is no better than random, to 1, where the model can efficiently predict the species distribution. The R function 'enmtools.vip' in ENMtools was used to calculate the importance of the seven variables as environmental predictors in niche modelling for each taxon.

Niche breadth (B2 values based on Levins, 1968) was calculated to represent the breadth of suitable climate conditions of a taxon, which can range from 0 (highly specialized species) to 1 (generalist species with higher environmental tolerances) (Sexton et al., 2017). Niche overlap (Schoener's *D* based on Schoener, 1968) was calculated to evaluate the similarity of a pair of modelled niches and can range from 0 (no similarity) to 1 (identical projection). We calculated the niche breadth of predicted species distribution range using 'raster.breadth' for each taxon and performed pairwise niche overlap analysis using 'raster.overlap' in ENMTools, respectively.

3.2.8 Correlations Between Measured Traits

To investigate the divergence of polyploidy-associated traits, NZ *Azorella* ploidy levels were compared to 2C genome sizes and the number of homeologous copies, to determine 1) the total genome size changes after WGD; and 2) if higher polyploids have more homeologous copies per targeted gene due to polyploidization (using the formula: Homeologous copy number = Ploidy level + Supercontig recovery rate; see Notes S4). In addition, the correlation between 2C genome and

stomatal guard cell length was calculated to explore 3) the cell upscaling effect that has been shown to occur in plants with larger genome sizes (Beaulieu et al., 2008).

Furthermore, NZ *Azorella* ploidy levels were also compared to post-WGD associated traits (e.g., 1Cx genome sizes, 19 bioclimate plus elevation layers extracted environmental data, and niche breadth) of taxa in section *Schizeilema* to reveal 1) post-polyploidization genomic modification processes, such as genome downsizing (Leitch and Bennett, 2004); 2) environmental divergence, which may be correlated with polyploidization; and 3) niche evolution patterns between different ploidy levels (Christian and Olivier, 2016).

We used least-squares linear regression (LM) that excluded phylogenetic signal, and phylogenetic generalized least-squares (PGLS) regression (Orme et al., 2013) that included phylogenetic signal (estimated λ = "ML") to perform the correlation tests in the R package CAPER. The phylogeny of the clade comprising section *Schizeilema* was selected from the ASTRAL-PRO tree for the PGLS modelling. The adjusted *R* square values (R^2) and *P* values were calculated for each model.

3.3 Results

3.3.1 Comparison of PacBio and Illumina Platforms for Hyb-Seq Sequencing

In total, we generated 3,327,814 PacBio HiFi reads with quality scores of Q20 or greater for 20 *Azorella* individuals. Of these, 3,178,842 (95.52%) reads were successfully demultiplexed for each sampled individual using their Illumina dual-barcoded sequences. On average, 70% of an individuals' sequenced reads were 500 bp to 1.5 kbp long (Table S5). Given the library profile differences, the individual Azranunculus_NYBG2447_pacbio had the fewest sequenced reads (10,930) compared to individual Azhookeri_Cas_pacbio which had the most reads (360,801).

When the same settings were used to extract supercontigs and homeologs in HybPiper2 for sequences from the two different platforms, the 18 individuals had a similar number of genes with exons assembled (average: 273) in both Illumina Hyb-Seq and PacBio Hyb-Seq datasets (Table S6). However, compared to the Illumina dataset, the PacBio dataset had on average 3.5 times the number of genes with homeolog by length warning (35 vs. 10) and twice the number of genes without stitched exons (205 vs. 101). In spite of these differences, ASTRAL trees reconstructed from 304 and 284 genes trees using IQTREE2 for Illumina and PacBio datasets, respectively, exhibited similar topologies and gene concordance levels (Fig. S1a; Fig. S1b).

When analysing the supercontigs from both datasets together (i.e., combing the extracted supercontigs from two datasets in one), the reconstructed ASTRAL tree based on 281 filtered gene trees showed high concordance levels between the same individuals that were sequenced by the two different platforms (Fig. S2), except for the hexaploid, *A. hookeri*, and the sister species, *A. nitens* and *A. cockaynei*. Given the similar and highly congruent phylogenetic results based on the PacBio and Illumina datasets in separate and combined analyses (Fig. S2; Table S6), we merged the Illumina and PacBio reads for each of the 18 individuals to improve the extraction of supercontigs and homeologs (see details in Note S4).

3.3.2 Phylogenetic Analysis

All 21 individuals, including two samples that had only PacBio sequence reads (individuals Azhookeri_Mon_pacbio and Aznitens_Boy_pacbio), one individual that had only Illumina sequence reads (Azlycopodioides_NYBG2433_illumina) and the remaining 18 individuals that had sequence reads from both platforms, were reanalysed in HybPiper2 with modified homeologs detection parameters to extract supercontigs and homeologs. The results (Table S7) show that each individual had on average 308 genes with exons assembled, which included 209 genes without stitched exons, 90 genes with homeologs detected by length (i.e., multiple homeologs had a length longer than 50% of reference gene sequences), and 115 genes with homeologs detected by depth (i.e., multiple homeologs were detected regardless their *de novo* assembled length compared to the reference gene). The final dataset comprised 227 targeted genes (see Notes S4), which included both extracted supercontigs and homeologs.

Each trimmed gene alignment was on average 679 bp in length and had on average 52 informative sites. The multi-labelled gene trees reconstructed in IQTREE2 were summarized into a species tree using ASTRAL-PRO. The backbone of this tree was supported by high local posterior probabilities and showed that section *Stilbocarpa* (Sub) was sister to the clades of section *Schizeilema* (NZ1, NZ2 and Au) and section *Ranunculus* (SA) (Fig. 1). Within section *Schizeilema*, the NZ1 group contained species in three ploidy levels: the sister tetraploid species, *A. roughii* and *A. allanii*, which are closely related to the Australian *A. fragosea*, plus a grade of the sister hexaploid species, *A. nitens* and *A. cockaynei*, which is sister to the hexaploid *A. hookeri* and the decaploid *A. colensoi*. The NZ2 group contained mostly tetraploid species, including *A. hydrocotyloides, A. exigua, A. pallida* and the two non-monophyletic subspecies, *A. haastii* subsp. *haastii* and *A. haastii* subsp. *cyanopetala*. Regarding species relationships within section *Schizeilema*, the ASTRAL-PRO tree exhibited mostly low local posterior probability values for

many species and in some cases short branch lengths (e.g., *A. fragosea*, the sister species of *A. roughii* and *A. allanii*, and the species in NZ2).

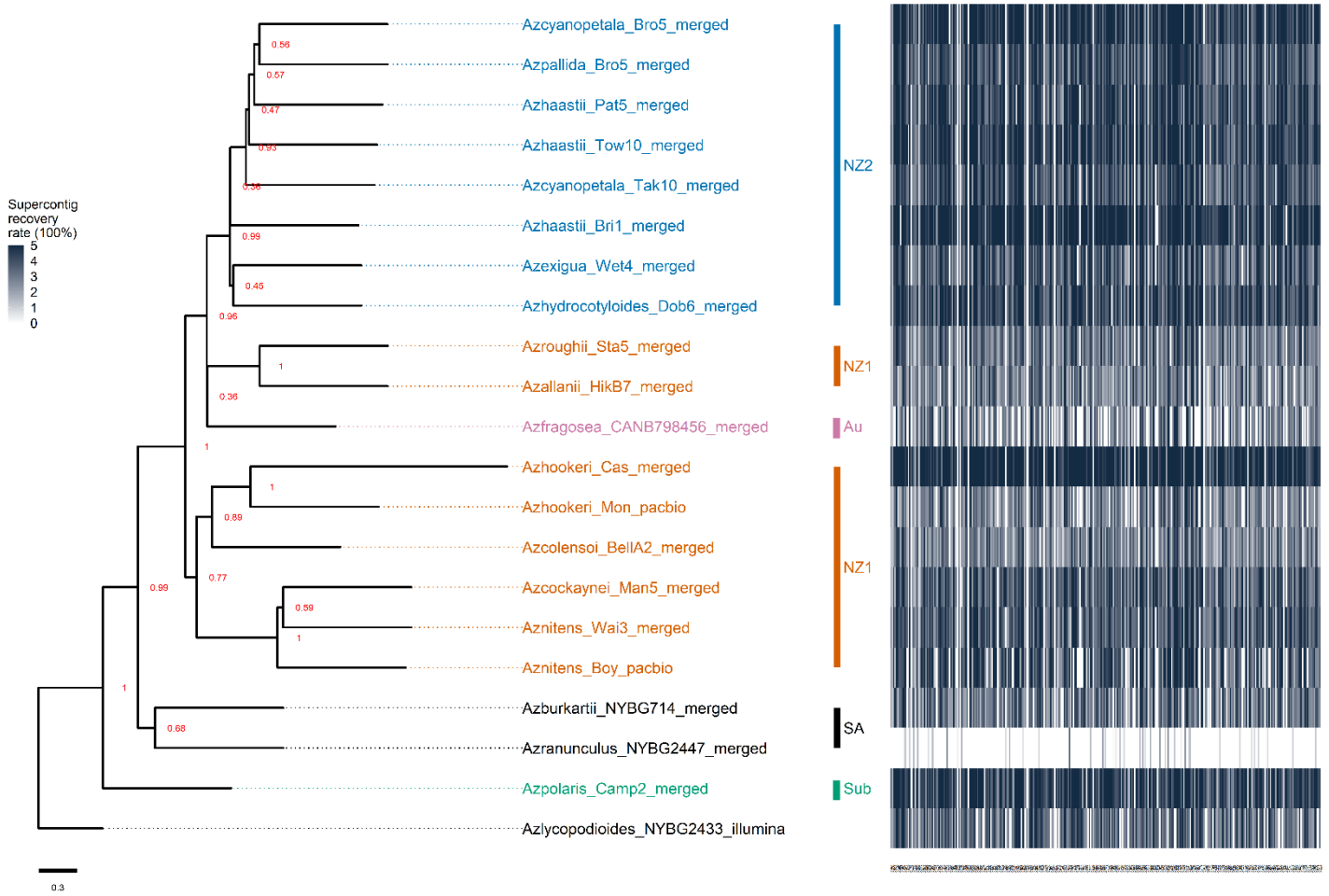


Fig. 1 Reconstructed phylogenetic ASTRAL-PRO tree for 21 individuals representing 13 New Zealand and Australian *Azorella* taxa, as well as three South American taxa (for information about sample names, see Table S2). Each node is labelled with its local posterior probability. The assigned genetic groups: New Zealand (NZ1, NZ2), Australia (Au), South America (SA) and subantarctic islands (Sub), are labelled at the tips of the tree. The heatmap shows the average supercontig recovery rates for each sample, which ranges from 0 to 500% [100 % = complete recovery of exons for the

length of the HybPiper selected reference. The rates higher than 100%, and up to 500 % = supercontigs (including all reference exons, targeted gene introns or flanking regions) were recovered].

3.3.3 Analysis of Genome Size and Ploidy Level

The genome size of 28 individuals from 14 New Zealand *Azorella* taxa were measured using flow cytometry. The measured 2C DNA content showed that ploidy does not vary within species for which multiple individuals were measured (Table S8). Therefore, the average 2C values of each taxon were calculated to represent total genome size (Table 1; Fig. 2).

The 2C genome size *of Azorella* section *Schizeilema* showed a 3.4-fold difference from the smallest (4.32 pg in *A. hydrocotyloides;* $2n = 4x = 32$) to the largest (14.6 pg in *A. colensoi;* $2n = 10x = 80$). The tetraploids showed the largest genome size variation, from 4.32 pg (*A. hydrocotyloides*) to 8.09 pg (*A. roughii*). Although *A. pallida* was previously reported as a hexaploid ($2n = 6x = 48$) (Hair, 1980), the genome size measured here is 4.54 pg, which is more similar to that of the tetraploids. By contrast, the three mainland New Zealand hexaploids, *A. hookeri*, *A. nitens* and *A. cockaynei,* had a similar average genome size of 8.59 pg. Similarly, the subantarctic island megaherbs *A. lyallii* (?*x*) and *A. polaris* (6*x*) had measured genome sizes of 7.83 pg and 8.09 pg, respectively. The average genome size of the sole New Zealand decaploid, *A. colensoi,* was 14.60 pg. Genome size (2C) of sampled *Azorella* showed strong phylogenetic signal ($\lambda = 1$).

The identifications of the New Zealand *Azorella* voucher specimens (CHR) for the previously published chromosome numbers Hair (1980) were confirmed, except for *A. pallida*, whose specimen listed in the publication could not be located. Given the close relationships and similar genome sizes of *A. pallida* and the other tetraploid species in NZ2, we consider the sampled individual of *A. pallida* (2C 4.54pg) to be tetraploid ($2n = 4x = 32$) for downstream analyses.

Monoploid genome size (1Cx) in New Zealand *Azorella* ranged from 1.08 pg to 2.02 pg (Table 1). Three hexaploids (*A. nitens*, *A. cockaynei* and *A. hookeri*) had similar 1Cx values (c. 1.42 pg) and seven tetraploids exhibited variation of 1Cx genome sizes to different degrees (ranging from 1.08 pg to 2.02 pg). By contrast, the decaploid *A. colensoi* also had a 1Cx similar to hexaploid *A. hookeri* (1.46 pg). The estimated monoploid ancestral genome size of all three *Azorella* sections (*Schilzeilema*, *Stilbocarpa*, and *Ranunculus*) was 1.93 pg (95% confidence interval 1.75 – 2.11) (Fig. 2). Only *A. roughii* (2.02 pg) had a monoploid genome size that was higher than the estimated ancestral 1Cx genome size, whereas all remaining taxa showed various levels of genome downsizing.

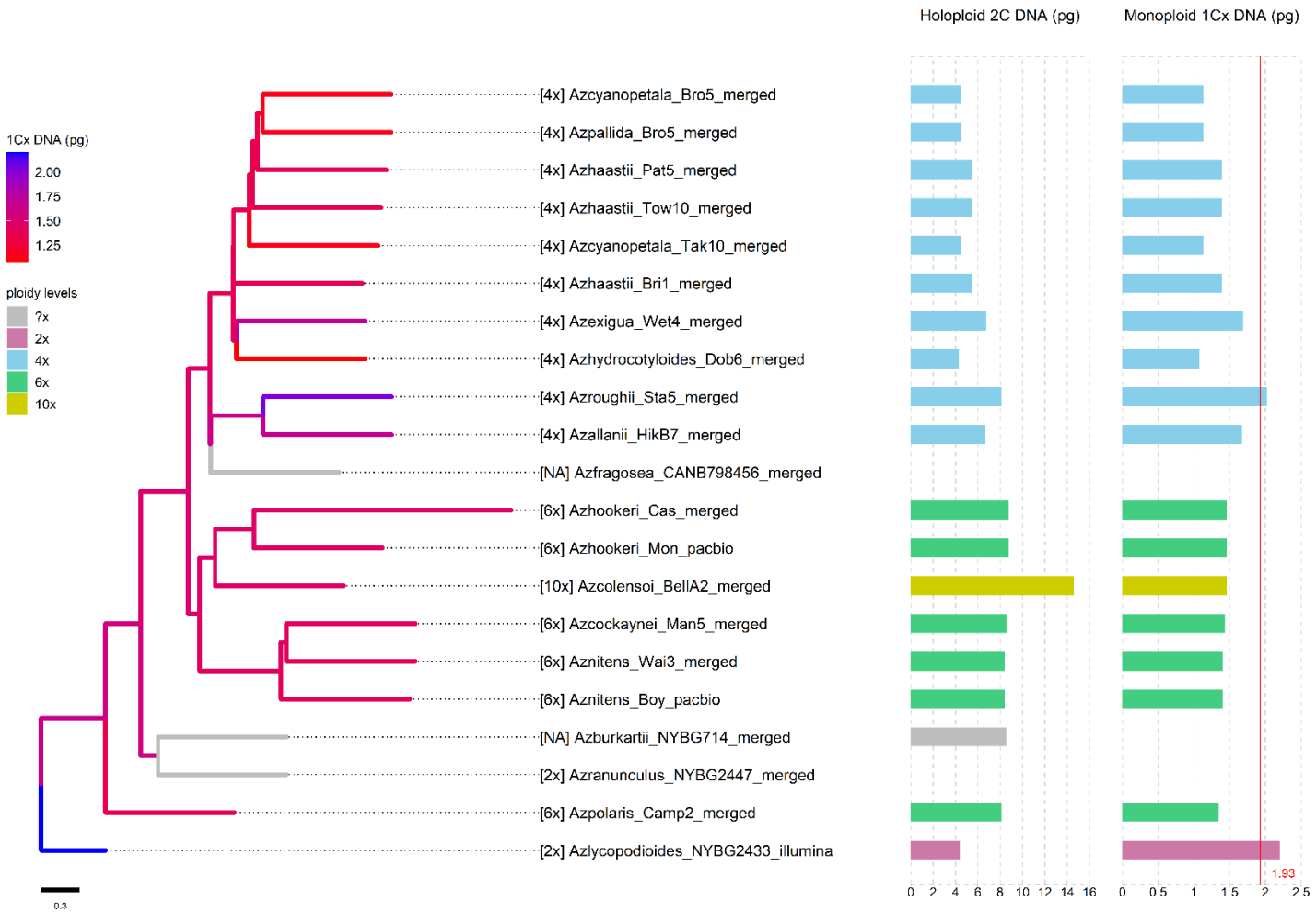


Fig. 2 Estimated ancestral monoploid genome sizes (1Cx) in *Azorella* section *Schilzeilema* mapped onto the ASTRAL-PRO phylogenetic tree (see Fig. 1). Branches are coloured by the estimated ancestral 1Cx values (scale bar on the left shows 1Cx range from 1 to 2.25 pg), and the three individuals

with missing data (Table 2) are grey. Genome size values (2C and 1Cx) are plotted to the right of the tree, and the estimated ancestral genome size (1Cx) 1.93 pg is represented by the red line.

Table 1 Genome size (2C) estimation of New Zealand *Azorella* polyploid species using flow cytometry with two standards, pea (2C = 8.08 pg) and broadbean (2C = 25.64 pg) (see Methods). Groups represent phylogenetic clades and/or geographical distributions (see Fig. 1). Chromosome numbers were from Hair (1980) for New Zealand *Azorella* (NZ1 and NZ2), by Beuzenberg and Hair (1983) for subantarctic island megaherbs (Sub), by (Moore, 1967) for South America outgroups (SA). *n* refers to the number of individuals included in the genome size study for each species or subspecies. Mean values of CV and SD were calculated for tested samples and selected standards. Genome sizes for the South American species were estimated by Ptáček et al. (2022). The monoploid 1Cx genome size was calculated by dividing the holoploid 2C genome size by ploidy level. Mean guard cell length was measured and calculated in the current study as stated in the Methods. New Zealand *Azorella* taxa (with 2C measured; above the line) are ordered by the chromosome number within each genetic group.

Species	**Chromosome No.**	**Ploidy level**	**Group**	***n***	**Mean sample CV (%) ± SD**	**Mean Standard CV (%) ± SD**	**2C (pg) ± SD**	**1Cx (pg)**	**Standard**	**Mean guard cell length (µm) ± SD**
A. allanii	32	4*x*	NZ1	3	3.47 ± 0.16	3.32 ± 0.32	6.68 ± 0.09	1.67	Pea	19.71 ± 2.12
A. roughii	32	4*x*	NZ1	2	4.82 ± 0.28	3.10 ± 0.73	8.09 ± 0.20	2.02	Broadbean	23.27 ± 3.12
A. cockaynei	48	6*x*	NZ1	2	5.15 ± 1.10	3.48 ± 0.65	8.59 ± 0.31	1.43	Broadbean	17.58 ± 1.94
A. hookeri	48	6*x*	NZ1	3	4.46 ± 0.98	3.30 ± 0.33	8.77 ± 0.07	1.46	Broadbean	18.32 ± 1.75
A. nitens	48	6*x*	NZ1	4	5.04 ± 0.67	3.80 ± 0.32	8.42 ± 0.25	1.4	Broadbean	17.81 ± 1.83
A. colensoi	80	10*x*	NZ1	3	2.61 ± 0.30	3.33 ± 0.53	14.60 ± 0.46	1.46	Pea	19.74 ± 1.57
A. exigua	32	4*x*	NZ2	1	4.32	3.45	6.76	1.69	Pea	15.07 ± 1.63
A. haastii subsp. *cyanopetala*	32	4*x*	NZ2	1	4.71	4.66	4.53	1.13	Pea	15.35 ± 1.79
A. haastii subsp. *haastii*	32	4*x*	NZ2	4	4.62 ± 0.85	3.42 ± 0.29	5.55 ± 0.09	1.39	Pea	18.12 ± 2.34
A. hydrocotyloides	32	4*x*	NZ2	1	4.52	3.26	4.32	1.08	Pea	22.31 ± 2.6
A. pallida	32	4*x*	NZ2	2	5.97 ± 0.81	4.43 ± 0.88	4.54 ± 0.01	1.14	Pea	17.5 ± 1.85
A. polaris	48	6*x*	Sub	1	3.78	2.55	8.09	1.35	Broadbean	15.46 ± 1.79
A. lyallii		?*x*	Sub	1	5.45	2.86	7.83		Broadbean	17.71 ± 1.29
A. schizeilema	32	4*x*	NZ	-	-	-	-	-	-	13.82 ± 1.37
A. fragosea		?x	Au	-	-	-	-	-	-	14.66 ± 1.54
A. burkartii		?*x*	SA	-	-	-	8.53	-	-	18.66 ± 1.65
A. ranunculus	16	2*x*	SA	-	-	-	NA	-	-	17.39 ± 2.55
A. lycopodioides	16	2*x*	Outgroup	-	-	-	4.4	2.2	-	13.92 ± 1.48

3.3.4 Analysis of Guard Cell Length

The five individuals tested using two different drying processes showed herbarium specimen measured stomatal guard cell are slightly larger than silica gel-dried samples (0.28 μm to 0.91 μm longer, or 1-5% longer) (Table S3). However, the Student's *t*-Test showed no significant difference was reported between two different drying processes measured results (*P*-value = 0.7645). Of the 18 *Azorella* taxa measured (Table S4), including two additional taxa (i.e., *A. schizeilema* and *A. lyallii*, for which no sequence data were available), the tetraploids showed the largest variation in average measured guard cell length, ranging from 13.82 μm (*A. schizeilema*; 4*x*) to 23.38 μm (*A. roughii*; 4*x*) (Table 1). *Azorella haastii* subsp. *haastii* exhibited the largest intraspecific guard cell length variation (Fig. S3).

Among South American diploid taxa, *A. lycopodioides* (2*x*) showed the smallest length (13.92 μm), whereas *A. ranunculus* (2*x*) and *A. burkartii* (?*x*) had lengths of 17.39 μm and 18.67 μm, respectively. The three mainland New Zealand hexaploids (*A. nitens*, *A. hookeri* and *A. cockaynei*) showed similar average results c. 17.91 μm. The subantarctic island endemic *A. polaris* (6*x*) and *A. lyallii* (?*x*) averaged 15.47 and 17.71 μm, respectively. By contrast, the decaploid *A. colensoi* (10*x*) averaged 19.74 μm.

3.3.5 Copy Number Variation of Homeologs

The average homeologs copy number for the 189 selected loci varied from 0.24 in *A. ranunculus* (Azranunculus_NYBG2447_merged) to 2.06 in *A. hookeri* (Azhookeri_Cas_merged) (Table 2). Overall, New Zealand hexaploids (*A. nitens*, *A. hookeri*. *A. cockaynei* and *A. polaris*) had high numbers of homeologs which were above 1.75, except for Azhookeri_Mon_pacbio with 1.53, which may be due to its lower supercontig recovery rate (3.13). The sole decaploid sample *A. colensoi* (Azcolensoi_BellA2_merged) had the second lowest supercontig recovery rate (2.85), yet it still had relatively large numbers of gene copies (1.68). All tetraploids had homeolog copies per targeted locus ranging from 1.26 in *A. allanii* (Azallanii_HikB7_merged) to 1.69 in *A. haastii* subsp. *cyanopetala* (Azcyanopetala_Bro5_merged). By contrast, excluding *A. ranunculus* (Azranunculus_NYBG2447_merged) that had few homeologs due to insufficient sequence reads (Table 2), the average number of homeolog copies for South American and Australian taxa approached 1.

Table 2 The average extracted gene copy number and mean supercontig recovery rate for 186 selected target-enriched loci for each sequenced individual of New Zealand *Azorella* included in the PacBio and Illumina sequencing (Note S4). The individual ID represents the sampled individual (e.g., Azhookeri_Cas population of *Azorella hookeri*) and sequenced reads sourced (e.g., Illumina, PacBio or the merged reads from both sequencing platforms). For ploidy level, see Table 1. Individuals are listed in descending order of mean gene copy number.

Individual ID	Ploidy level	Mean gene copy number	Mean supercontig recovery rates
Azhookeri_Cas_merged	6*x*	2.06	6.19
Aznitens_Boy_pacbio	6*x*	1.79	4.33
Aznitens_Wai3_merged	6*x*	1.79	4.09
Azpolaris_Camp2_merged	6*x*	1.77	5.01
Azcockaynei_Man5_merged	6x	1.76	3.99
Azcyanopetala_Bro5_merged	4*x*	1.69	5.4
Azcolensoi_BellA2_merged	10*x*	1.68	2.85
Azhaastii_Bri1_merged	4*x*	1.58	5.73
Azpallida_Bro5_merged	4*x*	1.57	4.25
Azhaastii_Tow10_merged	4*x*	1.55	5.06
Azhydrocotyloides_Dob6_merged	4*x*	1.54	4.65
Azhookeri_Mon_pacbio	6*x*	1.53	3.13
Azhaastii_Pat5_merged	4*x*	1.52	4.62
Azcyanopetala_Tak10_merged	4*x*	1.51	4.24
Azexigua_Wet4_merged	4*x*	1.45	4.09
Azroughii_Sta5_merged	4*x*	1.42	3.34
Azallanii_HikB7_merged	4*x*	1.26	3.13
Azfragosea_CANB798456_merged	?*x*	1.14	3.36
Azburkartii_NYBG714_merged	?*x*	1.08	3.49
Azlycopodioides_NYBG2433_illumina	2*x*	0.97	3
Azranunculus_NYBG2447_merged	2*x*	0.24	0.3

3.3.6 Environmental Niche Modelling

Environmental niche modelling of 11 New Zealand mainland *Azorella* taxa (Fig. 2) was performed using 1,120 filtered georeferenced records and 7 selected environmental layers (i.e., BIO2, BIO3, BIO5, BIO8, BIO9, BIO13 and BIO15) (Fig. S4; Table 5). The predicted geographic distribution range for the 11 *Azorella* taxa covered most of their known species' distribution (Fig. 3). All the taxa had at least 20 presence records for niche modelling, except for *A. allanii* and *A. cockaynei* which had only 13 and 16, respectively (Table 3). In addition, AUC values for all training models as calculated by ENM tools (Table 3) showed that two widely distributed hexaploids, *A. nitens* and *A. hookeri* were classified as good models (0.8 to 0.9), whereas all remaining taxa were excellent (0.9 to 1).

The environmental predictor importance analysis (Table S9) showed BIO8 (Mean Temperature of Wettest Quarter) was important for the North Island endemic species *A. allanii* and *A. colensoi*, whereas BIO5 (Max Temperature of Warmest Month) was important for all remaining South Island endemic taxa (including those occurring on both the North and South Islands, *A. hookeri* and *A. nitens*), except for *A. exigua* (BIO2 = Mean Diurnal Range) and *A. pallida* (no particular high importance values for any predictor).

The three hexaploids had high niche breadth values (Table 3): *A. nitens* (0.72), *A. hookeri* (0.70) and *A. cockaynei* (0.62). Two NZ1 tetraploids had medium values, *A. allanii* (0.46) and *A. roughii* (0.34). The decaploid *A. colensoi* (0.29) had the lowest niche breadth among NZ1 taxa. All NZ2 tetraploids showed niche breadth values between 0.20 to 0.27. The mean niche overlap value (Schoener's *D*) across all species pairs was 0.52. We considered niche overlap scores (Schoener's *D*) above 0.65 to be highly overlapped niche pairs (Table 5). Niche overlap analysis (Table 5) showed the following taxon pairs in NZ1 had high similarity values: *A. allanii* and *A. colensoi*, *A. allanii* and *A. nitens*, and *A. nitens* and *A. hookeri*. Whereas within NZ2, similar niche taxon pairs were mostly between four taxa including: *A. haastii* subsp. *cyanopetala*, *A. haastii* subsp. *haastii*, *A. hydrocotyloides* and *A. pallida*. By contrast, *A. exigua* showed the lowest niche overlap compared to all species pairs (average across all species pairs Schoener's $D = 0.39$) (Table 3).

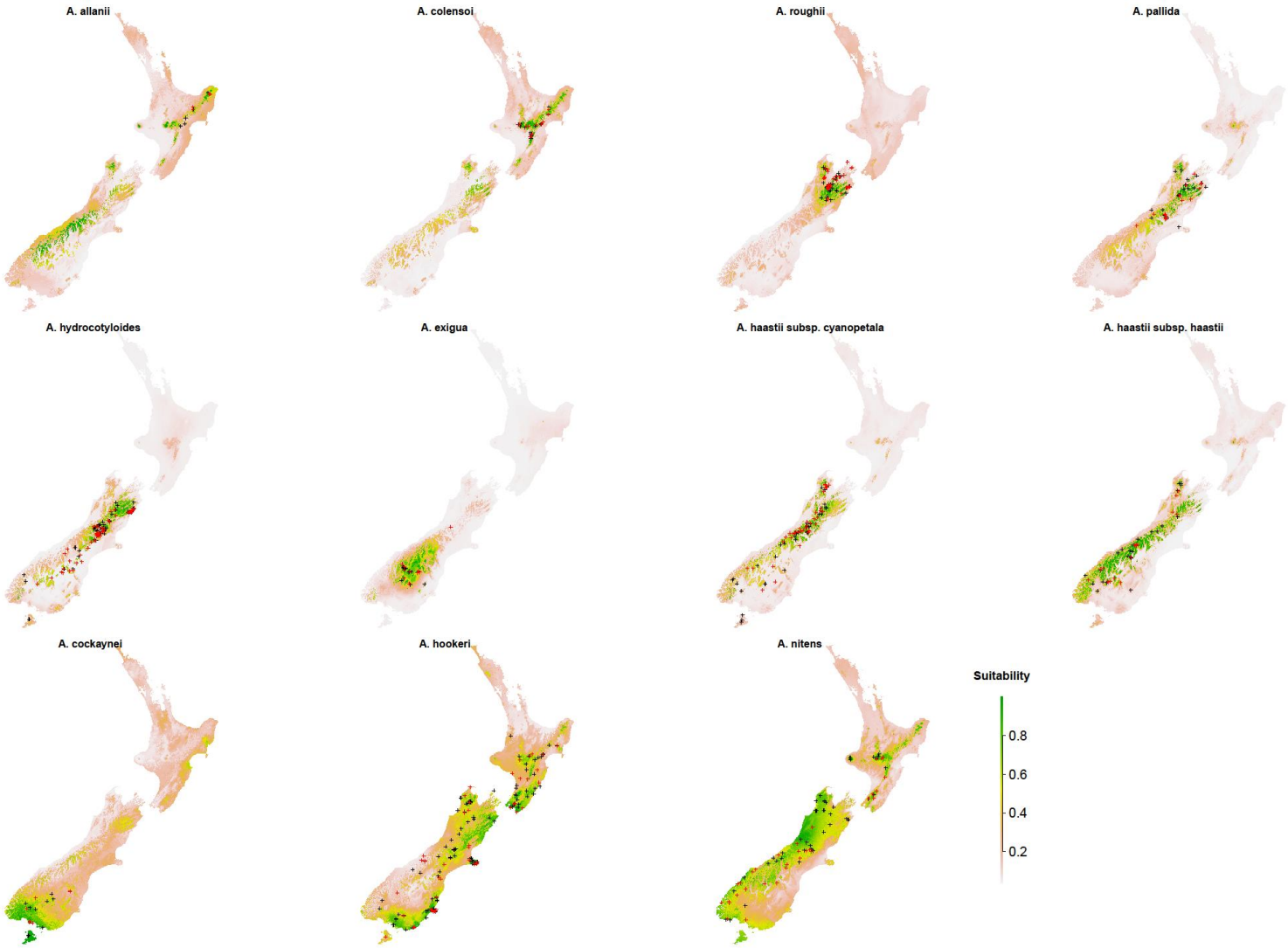


Fig. 3 Ecological niche modelling outcomes for 11 New Zealand mainland endemic *Azorella* taxa. Species names are labelled above each predicted suitability plot starting with North Island endemic species in the upper left to South Island endemic species in the lower right. The training individuals

used to build each niche model in MaxEnt are labelled with a black cross, whereas the testing individuals are labelled with a red cross. The scale bar shows the suitability score, which ranges from 0 (not suitable) to 1 (highly suitable).

Table 3 ENMtools niche modelling results for each mainland New Zealand endemic *Azorella* species and subspecies in section *Schizeilema*. The number of individuals for building the model (training samples) and for testing the model (testing samples) are given, as well as the AUC values used to evaluate niche model performance. Niche breadth was estimated using the predicted suitability range of each taxon. Species are in descending order of niche breadth within each ploidy level.

Species	**Ploidy level**	**Group**	**no. training sample**	**no. testing sample**	**Training evaluation AUC**	**Niche breadth**
A. exigua	4*x*	NZ2	25	12	0.97	0.20
A. hydrocotyloides	4*x*	NZ2	168	73	0.98	0.21
A. haastii subsp. *cyanopetala*	4*x*	NZ2	104	45	0.95	0.25
A. haastii subsp. *haastii*	4*x*	NZ2	54	24	0.94	0.25
A. pallida	4*x*	NZ2	49	22	0.95	0.27
A. roughii	4*x*	NZ1	98	42	0.98	0.34
A. allanii	4*x*	NZ1	13	6	0.94	0.46
A. cockaynei	6*x*	NZ1	16	8	0.93	0.62
A. hookeri	6*x*	NZ1	136	59	0.84	0.70
A. nitens	6*x*	NZ1	81	35	0.82	0.72
A. colensoi	10*x*	NZ1	35	15	0.97	0.29

Table 4 Pairwise niche overlap comparison using Schoener's *D* scores for species and subspecies of New Zealand *Azorella* section *Schizeilema*. The colour gradients indicate the overlap (Schoener's *D*), which ranges from 0 to 1, and the values (Schoener's *D*) higher than 0.65 are shown in bold text (see main text).

	A. allanii	*A. cockaynei*	*A. colensoi*	*A. exigua*	*A. haastii* subsp. *cyanopetala*	*A. haastii* subsp. *haastii*	*A. hookeri*	*A. hydrocotyloides*	*A. nitens*	*A. pallida*
A. cockaynei	0.57									
A. colensoi	**0.68**	0.44								
A. exigua	0.44	0.41	0.32							
A. haastii subsp. *cyanopetala*	0.61	0.48	0.49	0.44						
A. haastii subsp. *haastii*	0.59	0.43	0.42	0.50	**0.74**					
A. hookeri	0.54	0.62	0.46	0.29	0.42	0.35				
A. hydrocotyloides	0.47	0.42	0.44	0.39	**0.69**	0.57	0.41			
A. nitens	**0.65**	0.63	0.49	0.40	0.55	0.52	**0.72**	0.48		
A. pallida	0.60	0.44	0.51	0.44	**0.77**	**0.68**	0.48	**0.69**	0.59	
A. roughii	0.56	0.52	0.59	0.29	0.61	0.47	0.59	0.61	0.58	0.64

3.3.7 Correlation between WGD Associated Traits

For taxa in section *Schizeilema*, the correlation analyses between polyploidy-associated traits showed: 1) higher ploidy levels were significantly positively correlated with 2C genome size ($P < 0.001$; Fig. 4a); 2) 2C genome size values exhibited an increasing trend with measured guard cell length, however, no significant correlation was detected ($P > 0.1$; not significant; Fig. 4b); 3) the correlation between ploidy level and number of homeologous copies was significantly correlated ($P < 0.001$; PGLS $R^2 = 0.69$; LM $R^2 = 0.75$).

Among post-WGD divergence associated traits, New Zealand *Azorella* ploidy levels showed no correlation with 1Cx genome size ($P > 0.1$; Fig. 4c), nor with elevation ($P > 0.1$; Fig. 4d) or niche breadth ($P > 0.1$; Table 5). Only five bioclimate variables (BIO1 = Annual Mean Temperature, BIO5 = Max Temperature of Warmest Month, BIO6 = Min Temperature of Coldest Month, BIO10 = Mean Temperature of Warmest Quarter, and BIO11 = Mean Temperature of Coldest Quarter) were significantly correlated with higher ploidy levels in the results using LM models (Table 5).

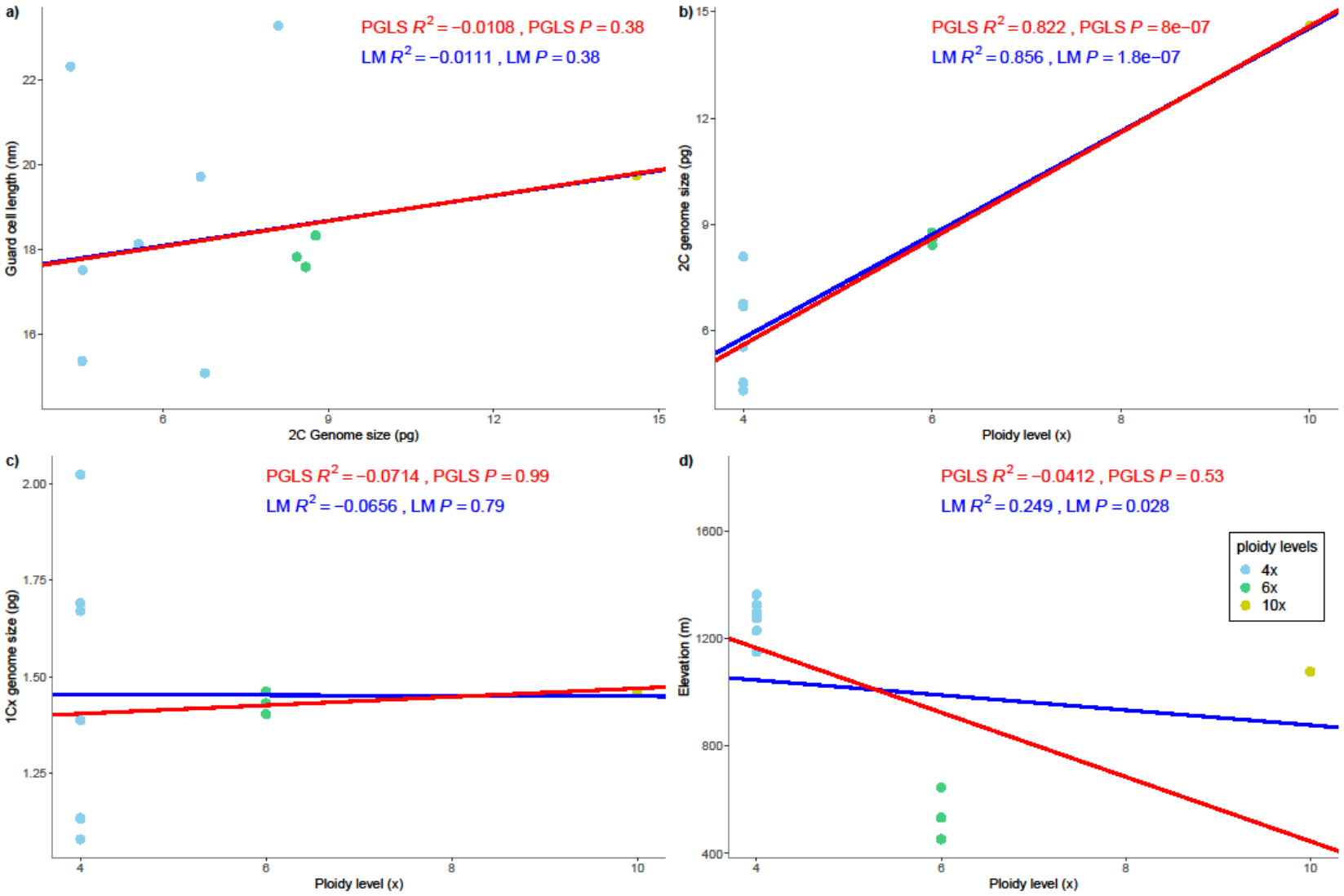

a)
PGLS R^2 = −0.0108 , PGLS P = 0.38
LM R^2 = −0.0111 , LM P = 0.38
Guard cell length (nm)
2C Genome size (pg)
b)
PGLS R^2 = 0.822 , PGLS P = 8e−07
LM R^2 = 0.856 , LM P = 1.8e−07
2C genome size (pg)
Ploidy level (x)
c)
PGLS R^2 = −0.0714 , PGLS P = 0.99
LM R^2 = −0.0656 , LM P = 0.79
1Cx genome size (pg)
Ploidy level (x)
d)
PGLS R^2 = −0.0412 , PGLS P = 0.53
LM R^2 = 0.249 , LM P = 0.028
ploidy levels
4x
6x
10x
Elevation (m)
Ploidy level (x)

Fig. 4 Correlation tests of polyploidy-associated traits by linear model (LM; blue) and phylogenetic generalized least-squares (PGLS; red) regressions. The R^2 and *P* values of each model are shown. a) Correlation between 2C genome size (pg) and stomatal guard cell length (μm). Correlations between ploidy leve (*x*) and b) 2C genome sizes (pg), c) 1Cx genome size (pg), and d) elevation (m), respectively.

Table 5 Correlations tested using a linear model (LM) and phylogenetic generalized least-squares (PGLS) model between ploidy level and 19 different bioclimate layers (for all species of New Zealand *Azorella*) and also niche breadth (for taxa in *Azorella* section *Schizeilema* only). The calculated slope, R^2 and *P* value of each test are shown. *P* value significance levels are represented by asterisks, i.e. * $P < 0.01$; ** $P < 0.001$.

	Slope (LM)	R^2 **(LM)**	***P*** **(LM)**	**Slope (PGLS)**	R^2 **(PGLS)**	***P*** **(PGLS)**
BIO1 = Annual Mean Temperature	7.32	0.389	0.00584 **	3.79	0.121	0.102
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))	-0.139	-0.0683	0.841	-0.126	-0.0686	0.85
BIO3 = Isothermality (BIO2/BIO7) (×100)	0.406	0.311	0.0145	0.294	0.163	0.068
BIO4 = Temperature Seasonality (standard deviation ×100)	-65.4	0.184	0.0552	-33.2	-0.00406	0.349
BIO5 = Max Temperature of Warmest Month	5.81	0.421	0.00391 **	3.42	0.153	0.0746
BIO6 = Min Temperature of Coldest Month	7.86	0.374	0.00698 **	4.31	0.125	0.0982
BIO7 = Temperature Annual Range (BIO5-BIO6)	-2.04	0.056	0.191	-1.92	0.0496	0.203
BIO8 = Mean Temperature of Wettest Quarter	4.54	0.115	0.108	-0.0967	-0.0713	0.969
BIO9 = Mean Temperature of Driest Quarter	6.7	0.114	0.109	6.78	0.135	0.0894
BIO10 = Mean Temperature of Warmest Quarter	6.57	0.401	0.00502 **	3.46	0.131	0.0928
BIO11 = Mean Temperature of Coldest Quarter	8.35	0.381	0.00643 **	4.46	0.123	0.0998
BIO12 = Annual Precipitation	-87.2	0.00166	0.329	-73.1	-0.0131	0.384
BIO13 = Precipitation of Wettest Month	-6.29	-0.0353	0.496	-4.88	-0.0469	0.576
BIO14 = Precipitation of Driest Month	-3.31	-0.0367	0.505	-2.62	-0.0469	0.576
BIO15 = Precipitation Seasonality (Coefficient of Variation)	0.553	0.0882	0.14	0.481	0.0392	0.225
BIO16 = Precipitation of Wettest Quarter	-17.1	-0.0398	0.524	-12.9	-0.0512	0.612
BIO17 = Precipitation of Driest Quarter	-19.5	0.014	0.289	-16.7	-0.00039	0.336
BIO18 = Precipitation of Warmest Quarter	-20.3	0.014	0.289	-17.7	0.00223	0.327
BIO19 = Precipitation of Coldest Quarter	-10.8	-0.0583	0.683	-6.07	-0.0667	0.807
Niche breadth	0.0545	0.122	0.101	-0.00125	-0.0713	0.965

3.4 Discussion

The combination of the PacBio platform with a Hyb-Seq approach and the Angiosperms353 bait set (Johnson et al., 2018) improved the target gene recovery and their homeologous copy extraction by increasing the length of *de novo* assembled contigs and reducing the intron gaps between target-enriched exons. For New Zealand polyploid *Azorella* in sections *Schizeilema* and *Stilbocarpa*, the comparison of polyploidy-associated traits (i.e., holoploid 2C genome size, the number of extracted homeologous copies, and stomatal guard cell length) provided further evidence of species' polyploid evolutionary histories and post-WGD divergence. Furthermore, the haploid 1Cx genome size variation and niche space comparison of 11 taxa in *Azorella* section *Schizeilema* showed different post-polyploidization diversification patterns, which were mostly driven by the phylogenetic relationships (i.e., genome content), different genomic modification processes (1Cx downsizing or upsizing), reticulate histories, niche shift patterns, and the age of the polyploid species.

3.4.1 Hyb-Seq for Polyploid Species

Next-gene sequencing of the loci in the Angiosperms353 bait set via Hyb-Seq approach (Johnson et al., 2018) has been applied to resolve taxonomic relationships in studies ranging from high-level family classification (Baker et al., 2022) to low-level population genetic variation (Slimp et al., 2021). Using concordance levels of hundreds of captured single copy nuclear genes, Hyb-Seq can be particularly useful to identify the lineages that are involved in recent polyploidization and reticulation events of polyploid-rich groups (Thomas et al., 2021; Nauheimer et al., 2021; Tiley et al., 2021). However, to fully resolve the origins of polyploids and to understand their reticulate relationships, including additional homeologous gene copy of target-enriched loci is necessary (Nauheimer et al., 2021; Tiley et al., 2021; Šlenker et al., 2021; Gardner et al., 2021).

Because the Angiosperms353 bait set was based on the exons of reference sequences, using this bait set can only target most of the exons and limited introns from the enriched reads (Johnson et al., 2018). Recovery of such discontinuous exons presents bioinformatic challenges for phasing homeologous genes, especially when these have low sequence divergence (between the contributing parental genomes) or there are further reticulations among the lineages (Tiley et al., 2021; McKain et al., 2018). As a novel approach, we increased the fragment length of genomic libraries for Hyb-Seq and combined this with sequencing on a third-generation PacBio sequencer to increase the probability of intron recovery, as well as homeolog extraction. We overcame the limitation of

PacBio genomic libraries, which typically cannot be used for hybridization with Angiosperms353 baits due to the looped structure of their HiFi reads (Wenger et al., 2019), and successfully demonstrated the possibility of sequencing and demultiplexing the Illumina genomic library reads from a PacBio sequencer.

Our results showed PacBio long-read Hyb-Seq data can provide a similarly reliable outcome as traditional Illumina short paired-end Hyb-Seq data (Fig. S1). In addition, PacBio Hyb-Seq data can increase the extracted supercontig lengths (Fig. S2) and reduce the number of intron gaps between recovered exons (Table S7). This method can capture the conserved genomic regions in a few kbp length for multiple individuals across different taxonomic levels in one sequence pool. However, PacBio Hyb-Seq can be more sensitive to the DNA quality of starting material compared to Illumina Hyb-Seq (Brewer et al., 2019). For example, the highly degraded DNA sample of Azranunculus_NYBG2447 showed a significant difference in the number of genes with exons recovered between two sequence platforms, i.e., only 14 on PacBio vs. 204 on Illumina platforms (Table S7).

Although individual gene trees were likely to be affected by incomplete lineage sorting (Maddison and Knowles, 2006), for taxa in section *Schizeilema*, the repeated patterns of different homeologous copies formed divergent genetic groups among reconstructed multi-labelled gene trees, highlighting the complex evolutionary histories of New Zealand *Azorella* (Chapter 2). Moreover, by summarizing all multi-labelled gene trees, the reconstructed ASTRAL-PRO tree of the merged dataset (using sequencing reads from both Illumina and PacBio) showed alternative interspecific relationships among taxa in section *Schizeilema* (Fig. 1) compared to ASTRAL trees that had no homeologous genes (Fig. S1). The topological incongruence may indicate the polyploidization and hybridization evolutionary histories of taxa in *Azorella* section *Schizeilema* (Chapter 2). Nevertheless, a more efficient network approach that can remove the homologous gene copy from the extracted homeologs, summarize the discordance levels between homeologous genes, and allow for multiple origins of polyploid species, which remain as a bioinformatics challenge, may in future provide further insights into the species relationships (Hibbins and Hahn, 2022; Rothfels, 2021).

3.4.2a Genome Size Variation (2C) and Species Relationships

In a previous study, New Zealand *Azorella* sections *Schizeilema* and *Stilbocarpa* were found to be sister to South American sections *Huanaca, Ranunculus and Azorella*, but their phylogenetic relationships remain unresolved (Plunkett and Nicolas, 2017). The strong phylogenetic signal ($\lambda = 1$)

that was detected in the 2C genome sizes of taxa among *Azorella* sections *Schizeilema, Ranunculus* and *Stilbocarpa* shows that total genome size is significantly correlated to phylogenetic divergence (reviewed in Revell et al., 2008). Therefore, the similarities between 2C genome sizes of *Azorella* in New Zealand and South America may provide additional insight into species relationships. In particular, the New Zealand hexaploids (*A. hookeri*, *A. nitens*, *A. cockaynei*, *A. polaris*), as well as *A. lyallii* (ploidy unknown), in sections *Schizeilema* and *Stilbocarpa* (Table 1) have similar genome sizes (c. 8.0 to 8.5 pg) to their South American relatives in sections *Huanaca*, *Ranunculus* and *Azorella* (2C ranging from 8.13 pg in *A. andina* to 8.53 pg in *A. burkartii*) (Ptáček et al., 2022). These similar genome sizes suggest that, to resolve the origins of section *Schizeilema* in New Zealand, future studies should include taxa from sections *Huanaca* and *Azorella*. Although the largest genome size of tetraploid *A. roughii* (2C = 8.09 pg) was similar to that of the hexaploids, it is more likely the result of genome expansion (Fig. 2). On the other hand, *A. colensoi* was previously identified as an allopolyploid that originated between maternal hexaploid *A. hookeri* and paternal tetraploid *A. allanii* (Chapter 2), the additive holoploid 2C genome size of *A. colensoi* (2C = 14.60 pg) from *A. hookeri* (2C = 8.77 pg) and *A. allanii* (2C = 6.68 pg) may provide additional evidence of its allopolyploid origin.

3.4.2b Allele Variation Among 353 Single Copy Nuclear Genes

Polyploidization and reticulate evolutionary histories can increase genomic complexity, as well as the copy number variation of assembled alleles among targeted single copy nuclear genes (Karbstein et al., 2022b; Nauheimer et al., 2021; Johnson et al., 2016). On the other hand, duplicated gene copies (homologs or homeologs) resulting from WGD can have different fates, e.g., gene loss, subfunctionalization or neofunctionalization (reviewed by Prince and Pickett, 2002; Comai, 2005). For 353 single copy nuclear loci, the number of extracted genes with the allele variation can vary between different plant groups, e.g., ranging from 1 to 41 in eight sunflower (Asteraceae) subfamilies (Siniscalchi et al., 2021), c. 42 in *Veronica* (Plantaginaceae) (Thomas et al., 2021) and c. 123 in *Pogonolepis* (Asteraceae: Gnaphalieae) (Schmidt-Lebuhn, 2022). However, our results show that both DNA quality and downstream bioinformatic settings can affect allele variation detection (Table S6; Table S7), and since these factors varied among different studies, this may have led to the observed of targeted genes with allele variation mentioned above.

Nevertheless, the correlation between gene copy number variation and ploidy level of New Zealand *Azorella* suggested the higher polyploids with sufficient recovered supercontigs were expected to contain more homeologs copies. All hexaploids (*A. hookeri*, *A. nitens*, *A. cockaynei* and

A. polaris) and the decaploid (*A. colensoi*) showed a large number of genes with homeologous copies, which indicate their subgenome donors may be from more distantly related species (Table 2). By contrast, the high homeologous copy number seen in NZ2 tetraploids (*A. haastii* subsp. *haastii*, *A. haastii* subsp. *cyanopetala*, *A. hydrocotyloides* and *A. pallida*) compared to *A. exigua* (4*x*; NZ2 group) may result from high levels of gene flow between polyploid species (e.g., Schmickl and Yant, 2021), which was also evident in their genetic data when analyzed with a network approach (Chapter 2).

3.4.2c Cell Size Variation of Polyploids

Stomatal guard cell length is expected to be positively correlated with larger genome size (e.g., Lattier et al., 2019; Beaulieu et al., 2008). However, stomatal traits are important for plant leaf physiology functions (e.g., water usage efficiency or photosynthetic rates) (Aasamaa et al., 2001), and variation (i.e., plasticity changes) of the stomatal size may also be affected by different environmental conditions and species adaptation process (reviewed in Hetherington and Woodward, 2003; Hodgson et al., 2010). For naturally diverged polyploid species, guard cell length or physiological and functional traits of stomata may not correlate with their 2C genome size (Wong and Murray, 2012; Wei et al., 2019; Knight and Beaulieu, 2008).

Our results did not show a significant positive correlation between stomatal guard cell length and 2C genome size of taxa in section *Schizeilema* (Fig. 4b), which may be due to species of different ploidy levels occupying different habitats (Fig. 3). The three hexaploids (*A. hookeri*, *A. nitens*, *A. cockaynei*) are found mostly at lower elevations (c. 400 m to 600 m) compared to tetraploids, which are all alpine (c. 1,200 m to 1,500 m) (Fig. 4d). Among the NZ2 tetraploids that have similar genomic content, stomatal guard cell length variation (Fig. S3) may indicate different adaptation processes in the New Zealand mountains in different regions (Fig. 3) (e.g., Joly et al., 2014).

3.4.3a Post-WGD Genome Size Variation (1Cx)

Most post-polyploidization diversification is related to the process of diploidization (Li et al., 2021; Clark and Donoghue, 2018; Simonin and Roddy, 2018). Indeed, diploidization or genome downsizing is evident in many polyploid-rich plant lineages (Leitch and Bennett, 2004), e.g., *Veronica* (Plantaginaceae) (Meudt et al., 2015), *Allium* (Amaryllidaceae) (Wang et al., 2021a). However, this trend may vary in different plant groups (e.g., 1Cx upsizing or no correlation between ploidy and 1Cx), e.g., *Chenopodium* (Amaranthaceae) (Mandák et al., 2016), *Nicotiana* (Solanaceae) (Leitch et al., 2008) and *Plantago* (Plantaginaceae) (Wong and Murray, 2012).

Azorella section *Schizeilema* showed post-polyploidization diversification patterns in 1Cx genome size variation that differed among taxa in three ploidy levels (Fig. 4c). On the one hand, in comparison to the estimated ancestral 1Cx genome size of their South American ancestors, New Zealand polyploid taxa in *Azorella* sections *Schizeilema* (excluding *A. roughii*) and *Stilbocarpa* (only hexaploid *A. polaris* included) showed a downward trend (Fig. 2). However, 1Cx genome sizes within section *Schizeilema* showed no negative correlation with ploidy level (Fig. 4c). This pattern could be explained by the multiple origins (i.e., different subgenome donors) of tetraploids and hexaploids in section *Schizeilema* (Chapter 2). The additional combination of dissimilar genomes via allopolyploidization or hybridization may buffer the downsizing trend or even increase the 1Cx value in polyploids. Furthermore, the decaploid *A. colensoi* has a similar 1Cx as the hexaploids (Fig. 2), which may be due in part to the young age of *A. colensoi* (less than 50,000 years ago) (Chapter 2). Indeed, Leitch et al. (2008) showed the diploidization process may relate to the divergence time of the allotetraploids in *Nicotiana*. Hanna et al. (2006) also indicated the neopolyploids in *Orobanche* may need longer evolutionary time to show genomic reduction. By contrast, the NZ2 tetraploids (*A. haastii* subsp. *haastii*, *A. haastii* subsp. *cyanopetala*, *A. hydrocotyloides* and *A. pallida*) exhibited clear genomic downsizing in section *Schizeilema* (Fig. 2). Such a trend may result from adaptive evolution of divergent habitat conditions (Fig. 3) (e.g., Moraes et al., 2022; Luque et al., 2022).

Although the evolutionary history of genome size may vary in different lineages, the diploidization process is nevertheless an important process for the long-term survival of polyploid species (Faizullah et al., 2021; Petrov, 2002; Wang et al., 2021c). This is because the accumulated larger genome sizes in polyploids will have higher nutrition requirements (e.g., N and P) to maintain the integrity of the duplicated genome (e.g., for DNA repair) (Novák et al., 2020), which may limit nutrition to support cellular activities (e.g., photosynthesis and growth) (Kang et al., 2015; Faizullah et al., 2021). In addition, polyploids may benefit from WGD in the short-term via selective advantages by masking the expression of recessive deleterious mutations due to gene redundancy (Comai, 2005; Conover and Wendel, 2022). However, genome upsizing can increase the accumulation of deleterious mutations and eventually have a negative impact on the fitness of polyploids (Arrigo and Barker, 2012; Otto and Whitton, 2000). In the future, with the improvements of inferring polyploid species phylogenetic relationships using the advanced sequencing platforms, the post-polyploidization genomic evolution in polyploids may begin to show a more detailed pattern (Rothfels, 2021; Wang et al., 2021c).

3.4.3b Post-WGD Niche Evolution

Niche shifting after WGD is shown to be an important factor that can promote the divergence of polyploids from their diploid progenitors (Marchant et al., 2016), as well as the divergence between closely related polyploids (Joly et al., 2014; Padilla-García et al., 2022). On the one hand, environmental conditions had no clear pattern with the ploidy level changes in section *Schizeilema* (Table 5). By contrast, using niche overlap and niche breadth (e.g., Marchant et al., 2016), different niche evolution patterns (i.e., niche contraction, niche novelty, and niche conservatism) have been observed in *Azorella* section *Schizeilema* taxa with different ploidy levels.

The similar 1Cx values of the three closely related hexaploids (*A. nitens*, *A. hookeri* and *A. cockaynei*) indicated similar post-polyploidization genomic modification processes for these taxa (Fig. 2). Highly overlapped niche space and similar niche characterization (i.e., niche breadth) between *A. nitens* and *A. hookeri* suggested niche conservatism (Table 3; Table 4). The niche space of *A. cockaynei* was also highly overlapped with *A. nitens* and *A. hookeri*, but by contrast, its smaller niche breadth may indicate niche contraction (Table 3). Allo-decaploid *A. colensoi* showed highly overlapped niche space with its parental species *A. allanii*. However, it has a smaller niche breadth compared to its parental species, *A. hookeri* and *A. allanii* (Table 3), which may indicate the niche contraction or the young age of this neopolyploid ,i.e. it may be too early in the evolutionary history of this species to detect its evolutionary change.

The tetraploids (*A. allanii* and *A. roughii*) in the NZ1 group and *A. exigua* in NZ2 had higher 1Cx genome size compared to the remaining NZ2 tetraploids (*A. haastii* subsp. *haastii*, *A. haastii* subsp. *cyanopetala*, *A. hydrocotyloides* and *A. pallida*) (Chapter 2). The divergence between sister species *A. allanii* and *A. roughii* was most likely due to different post-WGD genomic modification processes (i.e., 1Cx expansion in *A. roughii*) (Fig. 2) (e.g., Qiu et al., 2019; Kang et al., 2014), as well as niche divergence that showed intermediate niche overlap (Schoener's $D = 0.56$) and smaller niche breadth in *A. roughii* (Table 3; Table 4).

Although *A. exigua* has a similar 1Cx value to *A. allanii*, it showed overall the lowest niche overlap among all species pairs (average Schoener's $D = 0.39$) (Table 4) and the smallest niche breadth, which may suggest it has a novel niche compared to all other tetraploids in this group. Indeed, *A. exigua* also showed diverged leaf morphology and different populations were always monophyletic in the phylogenetic trees using different genetic markers (i.e., single copy nuclear genes, nrDNA, and plastome; Chapter2). Hence, *A. exigua* showed as a distinct lineage compared to all other NZ *Azorella* taxa. By contrast, the other NZ2 tetraploids showed high niche overlap (average Schoener's $D = 0.69$) and similar niche breadth values, which suggest niche conservatism.

Compared to the radiation of monophyletic tetraploids in the New Zealand South Island endemic genus *Pachycladon* (Brassicaceae) (Joly et al., 2014), which showed limited gene flow and clear niche shift signals among species, the niche conservatism of NZ2 *Azorella* tetraploids may instead point to interspecific gene flow.

3.5 Conclusion

Polyploidization and hybridization events contributed to the origin of several lineages in *Azorella* section *Schizeilema*. Each lineage showed unique post-polyploidization diversification patterns. Our results highlighted the importance of comparing polyploidy-associated traits and niche spaces among polyploid species to understand their divergence patterns. In addition, compared to a diploid-tetraploid plant system, this insular polyploid-rich genus proved to be a more complex evolutionary model for exploring the consequences of WGD on a macroevolutionary scale. For example, including events such as gene flow between polyploids, additional mixture of dissimilar genomes via allopolyploidization or hybridization, and species diversification related to geological events or habitat differences. Moreover, the increasing ability to recover homeologous sequences via third-generation platforms for phylogenetic inference of polyploid species can provide further insights into their species relationships and diversification patterns.

Data Availability

The PacBio Hyb-Seq reads are available at NCBI BioProject ID:PRJNA911442. The Scripts for conducting the analyses are available at: https://github.com/WeixuanPlant/NZAzorellaAngiosperms353

Acknowledgements

We thank Caroline Mitchell, who helped W.N. with the flow cytometry protocol to estimate genome size of *Azorella* at Manaaki Whenua - Landcare Research, Lincoln. A very special thanks to Xiaoxiao Lin at Massey Genome Service, who made an enormous effort to polish the laboratory steps for PacBio sequencing with W.N. We are grateful to all staff from Massey University Manawatū Microscopy and Imaging Centre for assisting W.N. with stomatal guard cell length measurement using SEM. Thanks to the New Zealand eScience Infrastructure (NeSI) for providing high throughput computing resources.

Funding

This study was conducted as part of the Royal Society of New Zealand Marsden fund (17-LCR-006) to W.L., H.M., and J.T. Additional funding from the Miss E. L. Hellaby Indigenous Grasslands Research and Royal Society of New Zealand Hutton Fund to W.N. is gratefully acknowledged.

References Cited

Aasamaa K, Sõber A, Rahi M. 2001. Leaf anatomical characteristics associated with shoot hydraulic conductance, stomatal conductance and stomatal sensitivity to changes of leaf water status in temperate deciduous trees. *Functional Plant Biology* 28(8): 765-774.

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* 15(2): 140-146.

Baker WJ, Bailey P, Barber V, Barker A, Bellot S, Bishop D, Botigué LR, Brewer G, Carruthers T, Clarkson JJ. 2022. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology* 71(2): 301-319.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5): 455-477.

Bates OK, Bertelsmeier C. 2021. Climatic niche shifts in introduced species. *Current Biology* 31(19): R1252-R1266.

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist* 179(4): 975-986.

Beuzenberg E, Hair J. 1983. Contributions to a chromosome atlas of the New Zealand flora - 25 Miscellaneous species. *New Zealand Journal of Botany* 21(1): 13-20.

Brewer GE, Clarkson JJ, Maurin O, Zuntini AR, Barber V, Bellot S, Biggs N, Cowan RS, Davies NMJ, Dodsworth S, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15): 1972-1973.

Chamberlain S, Oldoni D, Waller J 2022. rgbif: Interface to the Global Biodiversity Information Facility API.

Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology* 58(1): 377-406.

Christian P, Olivier B. 2016. Towards unified hypotheses of the impact of polyploidy on ecological niches. *New Phytologist* 212(3): 540-542.

Clark JW, Donoghue PCJ. 2018. Whole-genome duplication and plant macroevolution. *Trends in Plant Science* 23(10): 933-945.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6(11): 836-846.

Conover JL, Wendel JF. 2022. Deleterious mutations accumulate faster in allopolyploid than diploid cotton (*Gossypium*) and unequally between subgenomes. *Molecular Biology and Evolution* 39(2): msac024.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research* 12(7): 1075-1079.

Dodsworth S, Chase MW, Leitch AR. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180(1): 1-5.

Doležel J, Sgorbati S, Lucretti S. 1992. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiologia Plantarum* 85(4): 625-631.

Faizullah L, Morton JA, Hersch-Green EI, Walczyk AM, Leitch AR, Leitch IJ. 2021. Exploring environmental selection on genome size in angiosperms. *Trends in Plant Science* 26(10): 1039-1049.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25(5): 471-492.
Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12): 4302-4315.
Francis D, Davies MS, Barlow PW. 2008. A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Annals of Botany* 101(6): 747-757.
Gardner EM, Johnson MG, Pereira JT, Puad ASA, Arifiani D, Sahromi, Wickett NJ, Zerega NJC. 2021. Paralogs and off-target sequences improve phylogenetic resolution in a densely sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Systematic Biology* 70(3): 558-575.
Glennon KL, Ritchie ME, Segraves KA. 2014. Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecology Letters* 17(5): 574-582.
Greilhuber J, DoleŽEl J, LysÁK MA, Bennett MD. 2005. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-Value' to describe nuclear DNA contents. *Annals of Botany* 95(1): 255-260.
Hair J. 1980. Contributions to a chromosome atlas of the New Zealand flora - 21 Umbelliferae (miscellaneous genera). *New Zealand Journal of Botany* 18(4): 559-562.
Hanna W-S, Johann G, Schneeweiss GM. 2006. Genome size evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. *American Journal of Botany* 93(1): 148-156.
Heenan PB, Cheeseman DF, Mitchell CM, Dawson MI, Smith LA, Houliston GJ. 2022. Genetic diversity of *Tradescantia fluminensis* complex (Commelinaceae) naturalised in Australia, New Zealand and South Africa. *New Zealand Journal of Botany*: 1-15.
Hetherington AM, Woodward FI. 2003. The role of stomata in sensing and driving environmental change. *Nature* 424(6951): 901-908.
Hibbins MS, Hahn MW. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* 220(2): iyab173.
Hijmans RJ, van Etten J 2012. raster: geographic analysis and modeling with raster data.
Hodgson JG, Sharafi M, Jalili A, Díaz S, Montserrat-Martí G, Palmer C, Cerabolini B, Pierce S, Hamzehee B, Asri Y, et al. 2010. Stomatal vs. genome size in angiosperms: the somatic tail wagging the genomic dog? *Annals of Botany* 105(4): 573-584.
Hutchinson GE 1957. Concluding remarks. population studies: animal ecology and demography. *Cold Spring Harbor Symposia on Quantative Biology*. 415-427.
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345): 97-100.
Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4(7): apps.1600016.
Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, et al. 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68(4): 594-606.
Joly S, Heenan PB, Lockhart PJ. 2014. Species radiation by niche shifts in New Zealand's rockcresses (*Pachycladon*, Brassicaceae). *Systematic Biology* 63(2): 192-202.
Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* 26(13): 1669-1670.
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14(6): 587-589.

Kang M, Tao J, Wang J, Ren C, Qi Q, Xiang Q-Y, Huang H. 2014. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. *New Phytologist* 202(4): 1371-1381.
Kang M, Wang J, Huang H. 2015. Nitrogen limitation as a driver of genome size evolution in a group of karst plants. *Scientific Reports* 5(1): 11636.
Karbstein K, Tomasello S, Hodač L, Wagner N, Marinček P, Barke BH, Paetzold C, Hörandl E. 2022. Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large *Ranunculus auricomus* species complex. *New Phytologist*.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30(4): 772-780.
Kearney M, Porter W. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters* 12(4): 334-350.
Knight CA, Beaulieu JM. 2008. Genome size scaling through phenotype space. *Annals of Botany* 101(6): 759-766.
Kuhn M. 2012. The caret package. *Journal of Statistical Software* 28.
Lattier JD, Chen H, Contreras RN. 2019. Variation in genome size, ploidy, stomata, and rDNA signals in Althea. *Journal of the American Society for Horticultural Science* 144(2): 130-140.
Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* 320(5875): 481-483.
Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82(4): 651-663.
Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Annals of Botany* 101(6): 805-814.
Levins R. 1968. *Evolution in changing environments*: Princeton University Press, Princeton, NJ.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
Li Z, McKibben MTW, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72(1): 387-410.
López-Jurado J, Mateos-Naranjo E, Balao F. 2022. Polyploidy promotes divergent evolution across the leaf economics spectrum and plant edaphic niche in the *Dianthus* broteri complex. *Journal of Ecology* 110(3): 605-618.
Luque JMR, Moreno EMS, Kovalsky IE, Seijo JG, SolÍs Neffa VG. 2022. Polyploidy, genome size variation and diversification in an autopolyploid complex: the case of *Turnera sidoides* (Passifloraceae, Turneroideae). *Systematics and biodiversity* 20(1): 1-18.
Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55(1): 21-30.
Mandák B, Krak K, Vít P, Pavlíková Z, Lomonosova MN, Habibi F, Wang L, Jellen EN, Douda J. 2016. How genome size variation is linked with evolution within *Chenopodium* sensu lato. *Perspectives in Plant Ecology, Evolution and Systematics* 23: 18-32.
Marchant BD, Soltis DE, Soltis PS. 2016. Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytologist* 212(3): 708-718.
Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264(5157): 421-424.
McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6(3): e1038.
McLay TGB, Birch JL, Gunn BF, Ning W, Tate JA, Nauheimer L, Joyce EM, Simpson L, Schmidt-Lebuhn AN, Baker WJ, et al. 2021. New targets acquired: improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9(7).

Meudt HM, Albach DC, Tanentzap AJ, Igea J, Newmarch SC, Brandt AJ, Lee WG, Tate JA. 2021. Polyploidy on islands: its emergence and importance for diversification. *Frontiers in Plant Science* 12(336).
Meudt HM, Rojas-Andrés BM, Prebble JM, Low E, Garnock-Jones PJ, Albach DC. 2015. Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Botanical Journal of the Linnean Society* 178(2): 243-266.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37(5): 1530-1534.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17): i541-i548.
Moore D. 1967. Chromosome numbers of Falkland Islands angiosperms. *British Antarctic Survey Bulletin* 14: 69-82.
Moraes AP, Engel TBJ, Forni-Martins ER, de Barros F, Felix LP, Cabral JS. 2022. Are chromosome number and genome size associated with habit and environmental niche variables? Insights from the Neotropical orchids. *Annals of Botany* 130(1): 11-25.
Nauheimer L, Weigner N, Joyce E, Crayn D, Clarke C, Nargar K. 2021. HybPhaser: a workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences* 9(7): e11441.
Novák P, Guignard MS, Neumann P, Kelly LJ, Mlinarec J, Koblížková A, Dodsworth S, Kovařík A, Pellicer J, Wang W, et al. 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants* 6(11): 1325-1329.
Orme D, Freckleton RP, Thomas GH, Petzoldt T, Fritz SA, Isaac N. 2013. CAPER: comparative analyses of phylogenetics and evolution in R. *Methods in Ecology and Evolution* 3: 145-151.
Otto F. 1990. DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. *Methods Cell Biology* 33: 105-110.
Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34(1): 401-437.
Paape T, Akiyama R, Cereghetti T, Onda Y, Hirao AS, Kenta T, Shimizu KK. 2020. Experimental and field data support range expansion in an allopolyploid *Arabidopsis* owing to parental legacy of heavy metal hyperaccumulation. *Frontiers in Genetics* 11.
Padilla-García N, Šrámková G, Záveská E, Šlenker M, Clo J, Zeisek V, Lučanová M, Rurane I, Kolář F, Marhold K. 2022. The importance of considering the evolutionary history of polyploids when assessing climatic niche evolution. *Journal of Biogeography* n/a(n/a).
Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401(6756): 877-884.
Peterson AT, Papeş M, Soberón J. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling* 213(1): 63-72.
Petrov DA. 2002. Mutational equilibrium model of genome size evolution. *Theoretical Population Biology* 61(4): 531-544.
Plunkett GM, Nicolas AN. 2017. Assessing *Azorella* (Apiaceae) and its allies: Phylogenetics and a new classification. *Brittonia* 69(1): 31-61.
Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics* 3(11): 827-837.
Ptáček J, Sklenář P, Pinc J, Urfusová R, Calviño CI, Urfus T. 2022. A pentaploid endosperm and a Penaea-type embryo sac are likely synapomorphies of *Azorella* (Apiaceae, Azorelloideae). *Plant Systematics and Evolution* 308(6): 40.
Qiu F, Baack EJ, Whitney KD, Bock DG, Tetreault HM, Rieseberg LH, Ungerer MC. 2019. Phylogenetic trends and environmental correlates of nuclear genome size variation in *Helianthus* sunflowers. *New Phytologist* 221(3): 1609-1618.

R Core Development Team 2013. R: a language and environment for statistical computing. Vienna, Austria.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3(2): 217-223.
Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* 57(4): 591-601.
Rothfels CJ. 2021. Polyploid phylogenetics. *New Phytologist* 230(1): 66-72.
RStudio Team 2020. RStudio: integrated development for R. *Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com*.
Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, et al. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9(7): 676-682.
Schmickl R, Yant L. 2021. Adaptive introgression: how polyploidy reshapes gene flow landscapes. *New Phytologist* 230(2): 457-461.
Schmidt-Lebuhn AN. 2022. Sequence capture data support the taxonomy of *Pogonolepis* (Asteraceae: Gnaphalieae) and show unexpected genetic structure. *Australian systematic botany* 35(4): 317-325.
Schoener TW. 1968. The anolis lizards of bimini: resource partitioning in a complex fauna. *Ecology* 49(4): 704-726.
Sexton JP, Montiel J, Shay JE, Stephens MR, Slatyer RA. 2017. Evolution of ecological niche breadth. *Annual Review of Ecology, Evolution, and Systematics* 48(1): 183-206.
Simonin KA, Roddy AB. 2018. Genome downsizing, physiological novelty, and the global dominance of flowering plants. *PLoS Biology* 16(1): e2003706.
Siniscalchi CM, Hidalgo O, Palazzesi L, Pellicer J, Pokorny L, Maurin O, Leitch IJ, Forest F, Baker WJ, Mandel JR. 2021. Lineage-specific vs. universal: a comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* 9(7).
Šlenker M, Kantor A, Marhold K, Schmickl R, Mandáková T, Lysak MA, Perný M, Caboňová M, Slovák M, Zozomová-Lihová J. 2021. Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq data: a case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. *Frontiers in Plant Science* 12.
Slimp M, Williams LD, Hale H, Johnson MG. 2021. On the potential of Angiosperms353 for population genomic studies. *Applications in Plant Sciences* 9(7).
Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15(1): 1-15.
Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, de Pamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96(1): 336-348.
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* 35: 119-125.
Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. *Annals of the Missouri botanical garden* 72(4): 824-832.
Stortchevoi A, Kamelamela N, Levine SS. 2020. SPRI beads-based size selection in the range of 2-10kb. *Journal of Biomolecular Techniques* 31(1): 7-10.
te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubešová M, Pyšek P. 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany* 109(1): 19-45.

Thomas AE, Igea J, Meudt HM, Albach DC, Lee WG, Tanentzap AJ. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *American Journal of Botany* 108(7): 1289-1306.
Tiley GP, Crowl AA, Manos PS, Sessa EB, Solís-Lemus C, Yoder AD, Burleigh JG. 2021. Phasing alleles improves network inference with allopolyploids. *bioRxiv*: 2021.2005.2004.442457.
Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* 33(1): 11-26.
Visser V, Molofsky J. 2015. Ecological niche differentiation of polyploidization is not supported by environmental differences among species in a cosmopolitan grass genus. *American Journal of Botany* 102(1): 36-49.
Wang G, Zhou N, Chen Q, Yang Y, Yang Y, Duan Y. 2021a. Gradual genome size evolution and polyploidy in *Allium* from the Qinghai–Tibetan Plateau. *Annals of Botany*: 109–122.
Wang X, Morton JA, Pellicer J, Leitch IJ, Leitch AR. 2021b. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal* 107(4): 1003-1015.
Warren DL, Matzke NJ, Cardillo M, Baumgartner JB, Beaumont LJ, Turelli M, Glor RE, Huron NA, Simões M, Iglesias TL, et al. 2021. ENMTools 1.0: an R package for comparative ecological biogeography. *Ecography* 44(4): 504-511.
Wei N, Cronn R, Liston A, Ashman T-L. 2019. Functional trait divergence and trait plasticity confer polyploid advantage in heterogeneous environments. *New Phytologist* 221(4): 2286-2297.
Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.
Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 37(10): 1155-1162.
Wong C, Murray BG. 2012. Variable changes in genome size associated with different polyploid events in *Plantago* (Plantaginaceae). *Journal of Heredity* 103(5): 711-719.
Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106(33): 13875-13879.
Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8(1): 28-36.
Zhang C, Mirarab S. 2022. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*: btac620.
Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30(5): 614-620.
Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, et al. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10(5): 744-751.
Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA. 2007. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evolutionary Biology* 7(1): 152.

Supplementary Notes

Note S1 Modified genomic library construction steps

The NEBNext® UltraTM II Library Prep kit protocol comprises four main steps, i.e. 1) DNA fragmentation, 2) adaptor ligation, 3) size selection, and 4) barcoding via PCR amplification. However, the standard NEB protocol can yield a maximum genomic library size of 1 kbp, because the binding buffer of the AMPure XP Beads (Beckman Coulter) used in the size selection step can only select fragments from 150 to 800 bp. Therefore, we followed the instructions in Stortchevoi et al. (2020), changing the concentration of NaCl in the binding buffer to increase the molecular weight capacity and to capture larger DNA fragments. The new buffer is made up of 20% PEG 8000 + 0.7 M NaCl + 10 mM Tris–HCl + 1 mM EDTA + 0.05% Tween 20, pH 8.0 @ 25 °C.

We started with c. 300 ng input DNA with a high molecular weight band and reduced the fragmentation from 10 min to 2-3 min to increase the portion of larger DNA fragments retained. After the standard adaptor ligation steps, a 1x ratio AMPure XP Beads in the original binding buffer was added to perform a size selection and pull out fragments larger than 200 bp. The supernatant was incubated for 5 min, and all the liquids were removed using a pipettor, such that only the pellet beads bound with selected DNA fragments were left on the magnetic rack. The beads were washed twice using 200 μl freshly prepared 80% ethanol, and air dried for 3 min at room temperature.

The pelleted beads were resuspended in 35 μl 1x TE buffer to release the bound DNA fragments. Next, a 0.75x ratio of new binding buffer was added and pipetted up and down 10 times to allow the desired DNA fragments to reattach to the beads. After 10 min incubation, the undesired DNA fragments (< 500 bp) in the supernatant were removed, and the pellet beads were washed twice using 200 μl freshly prepared 80% ethanol. Then the beads were eluted in 0.1x TE up to a total volume of 10 ul to continue the standard NEB protocol, including barcoding via PCR amplification.

Note S2. LIMA parameter settings

1) PacBio-sequenced Illumina genomic library reads are expected to be in the following structure:

[Illumina adaptor] + [Barcode1] + [NEB adaptor] + [Insertion between 500 bp ~ 2 kbp] + [NEB adaptor] + [Barcode2] + [Illumina adaptor]

2) We first generated a barcode fasta file (i.e., PacBio_barcode.lima.fasta) that included the sample names and the barcode pair as below:

> Sample1.1

[Illumina adaptor] + [Barcode1] + [NEB adaptor]

> Sample1.2

[Illumina adaptor] + [Barcode2] + [NEB adaptor]

3) Then we used the following command to demultiplex the PacBio reads from the compressed Lib22_HiFi.fastq file:

```
./lima --peek-guess --hifi-preset ASYMMETRIC --store-unbarcoded Lib22_HiFi.fastq \
PacBio_barcode.lima.fasta output_lima.fastq --split-named
```

Note S3. Phylogenetic reconstruction of PacBio- and Illumina-sequenced datasets

The supercontigs of each individual's targeted single copy nuclear genes were extracted using HybPiper v2 (Johnson et al., 2016). The target enrichment mega fasta reference file from McLay et al. (2021) was selected as a reference. For the *de novo* assembling steps, BWA (Li, 2013) was selected for reads mapping to the reference files for both PacBio and Illumina datasets. The intron and supercontig extractions were completed using the command '--run_intronerate'. In addition, for the Illumina dataset, we used the option '--merged' to allow paired-end reads to be merged when extracting exons in SPAde (Bankevich et al., 2012). By contrast, PacBio reads were treated as single-end reads in Hybpiper.

Extracted supercontigs for both datasets were initially analysed separately to compare the topologies of the resulting phylogenetic tree. Specifically, the supercontigs were aligned using MAFFT v.7.429 (Katoh and Standley, 2013) with '--auto' option, then trimmed in trimAl v.1.4 (Capella-Gutiérrez et al., 2009) to remove gap sites present in 30% or more of the sequences (-gt 0.7). Individual gene trees were reconstructed using IQ-TREE2 v. 2.2.0.5 (Minh et al., 2020b) with automatic model selection using ModelFinder (Kalyaanamoorthy et al., 2017), and a bootstrapping analysis was run on each gene tree 1000 times (-B 1000). All the nodes with bootstrapping values less than 30% in each gene tree were collapsed using the commands 'i & b<=30' in 'nw_ed' (https://github.com/tjunier/newick_utils) (Junier and Zdobnov, 2010).

Node-collapsed gene trees were rooted to the selected outgroup *A. lycopodioides* using 'reroot_trees.py' (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts). Next, gene trees that contained the outgroup were selected and summarized into ASTRAL trees using ASTRAL v.5.7.7 (Mirarab et al., 2014) for the PacBio dataset and Illumina datasets, respectively. The concordance levels among selected genes were calculated in PhyParts (https://bitbucket.org/blackrim/phyparts) (Smith et al., 2015) and 'phypartspiecharts.py' (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts). We plotted the concordance levels of gene trees using pie chart using the ggtree package (Yu et al., 2017) in R v.4.0.1 (R Core Development Team, 2013) (Fig. S1). In addition, we combined the supercontigs extracted from both Illumina and PacBio datasets for each gene. Using similar steps and thresholds as above, we reconstructed the phylogeny that included individuals that had been sequenced on both platforms (Fig. S2).

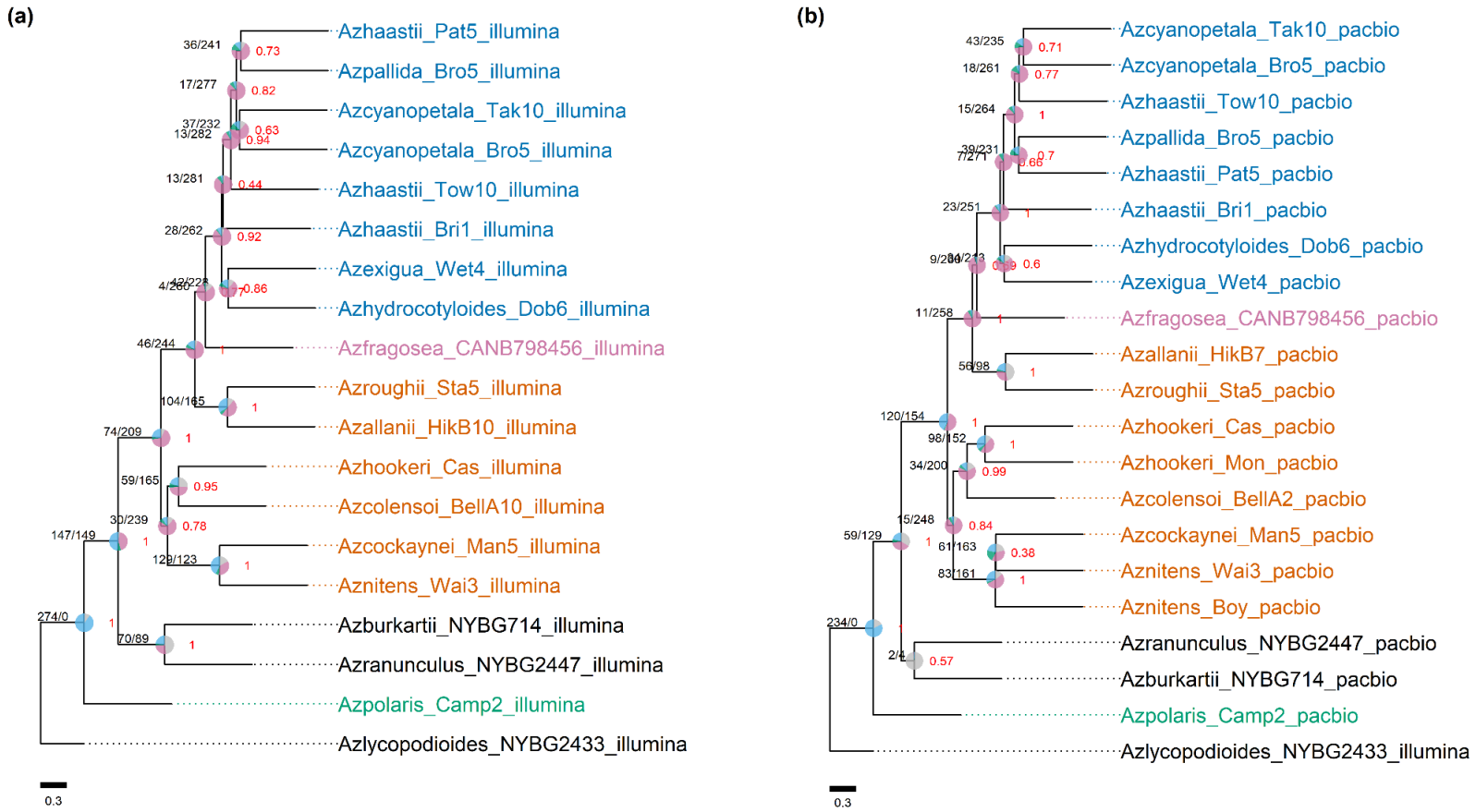
Note S4. Phylogenetic reconstruction of New Zealand *Azorella* individuals using merged data from both the Illumina and PacBio datasets

PacBio and Illumina sequence reads were merged with the following settings to improve the extraction of HybPiper supercontigs : the merged Illumina paired-end reads using PEAR v.0.9.11 (Zhang et al., 2014) were added to the PacBio reads. Two individuals (i.e., Azlycopodioides_NYBG2433_illumina and Azranunculus_NYBG2447_merged) were excluded from the next filtering step due to large missing data or only Illumina Hyb-Seq data was available.

We selected 227 genes without stitched exons and each gene has at least four taxa with extracted supercontigs, a minimal requirement for reconstructing a gene tree with bootstrap value, using the HybPiper output of the gene recovery rates for only 19 individuals (Table S7). Using 'paralog_retriever' in HybPiper with the default settings, the supercontigs and their extracted homeologs of 227 gene trees (i.e., multi-labelled gene trees) were reconstructed in IQ-TREE2 v. 2.2.0.5 (Minh et al., 2020b) and summarised into a species tree in ASTRAL-PRO (Zhang and Mirarab, 2022) (for details see Methods).

Furthermore, we calculated the average number of homeologous copies per targeted locus for 19 individuals, to test if higher-level (i.e., 6*x*, 10*x*) polyploids would have more homeologs per targeted locus. In addition, to reduce the effects of missing data, we selected 189 genes out of 227 genes using following thresholds: 1) each gene had all 19 individuals present (i.e., individual number ≥ 19); 2) there was at least one gene with allele variation detected (i.e., mean homeologous copy number per gene > 1); 3) a gene can only have a maximum of 5 homeologous copies (i.e., only two loci had values higher than 5 and may contain assembly errors). Therefore, 189 genes without stitched exons and with an average number of extracted homeologs between 1 to 5 per targeted locus were selected.

The average supercontig recovery rates (i.e., the length of recovered supercontigs / reference exon length) for the same selected 189 genes were calculated to rule out the possibility that insufficient sequence reads were the cause of a low number of homeolog copies being extracted (Table 2).



Supplementary Figures

Figure S1 ASTRAL topologies and gene concordance level comparison of phylogenetic trees of New Zealand *Azorella* based on a) Illumina and b) PacBio datasets. The colour of the text of the sampled individual names represents their phylogenetic groups in Fig. 1. Each node is labelled with a local posterior probability in red (0 to 1), two numbers in black (the number of gene trees that are concordant with that node/the number of gene trees

that are discordant), and a pie chart showing the concordance level (blue = concordant, green = discordant with a main alternative, pink = discordant with all remaining alternatives, and grey = uninformative).

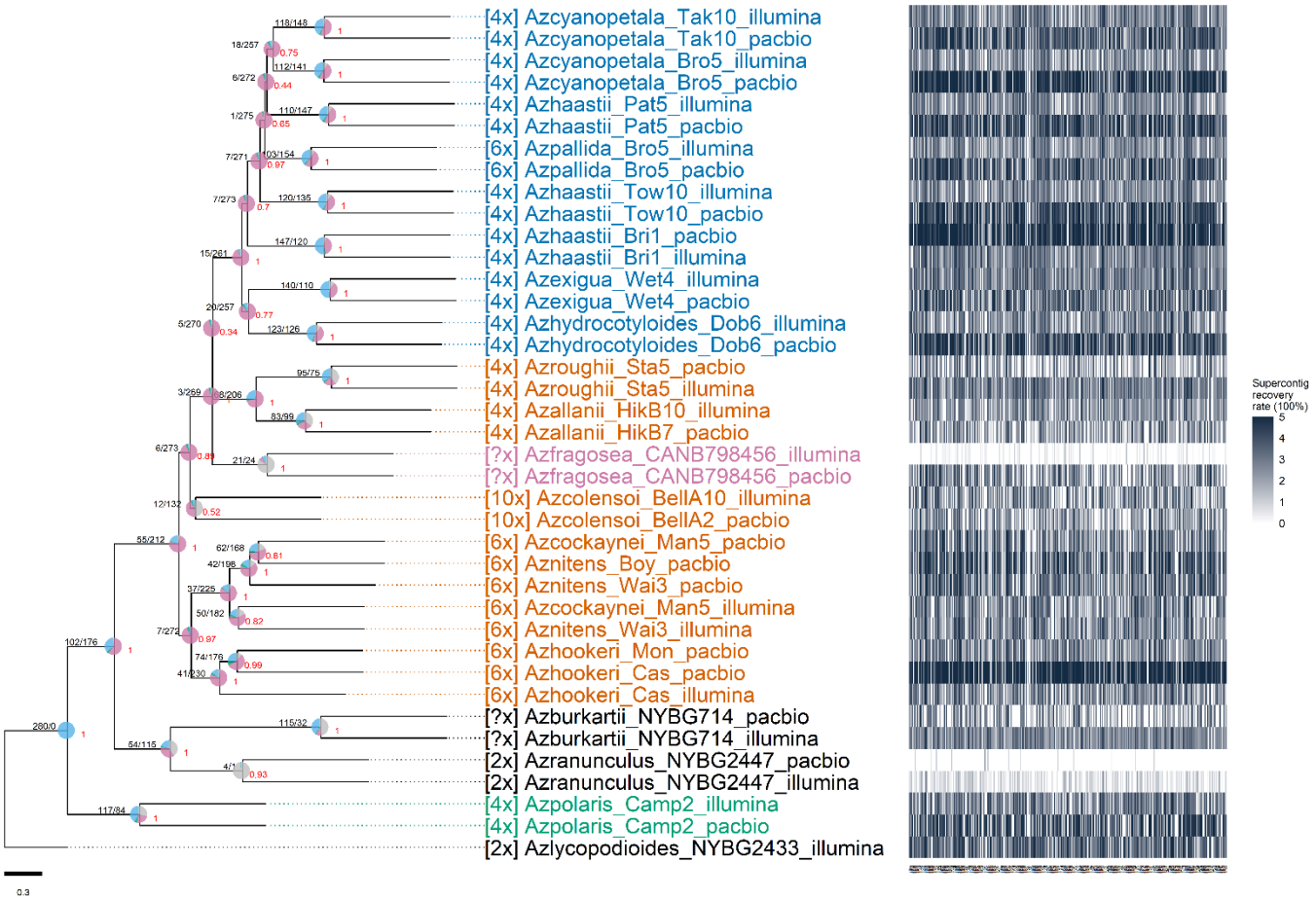


Figure S2 ASTRAL phylogenetic tree of combined data for 21 individuals of New Zealand and Australian *Azorella* and supercontig recovery rate heatmap (ranging from 0 to 500%, 100% = fully recovered genes of the reference exon length). For information about the numbers and pie charts at the tree nodes, and the heatmap, see Fig. S1.

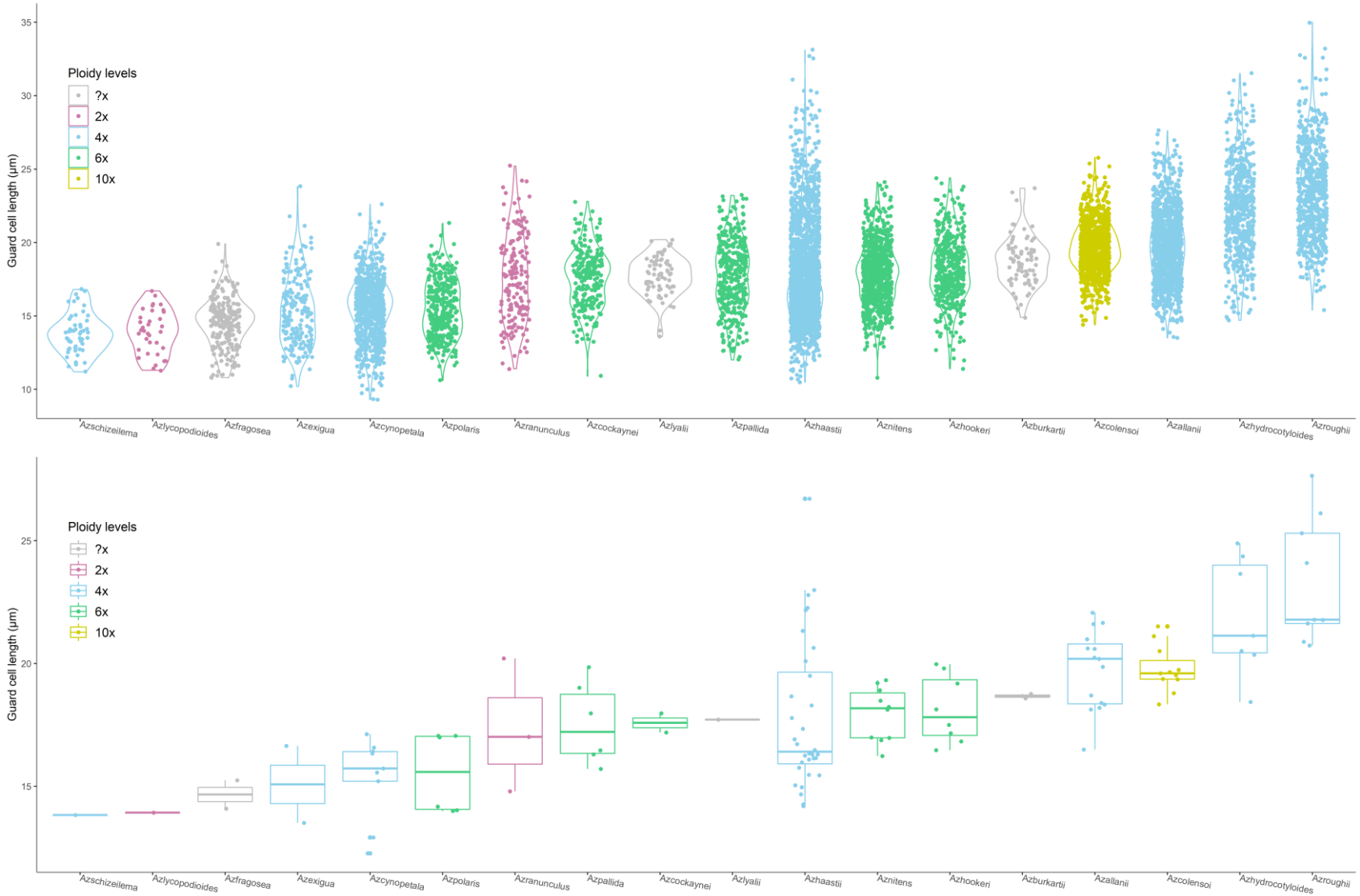


Figure S3 Measured stomatal guard cell length for 18 *Azorella* taxa. Each dot represents one measured guard cell in the upper violin plot, and each dot represents one individual in the lower box plot. Taxa are coloured by their ploidy levels.

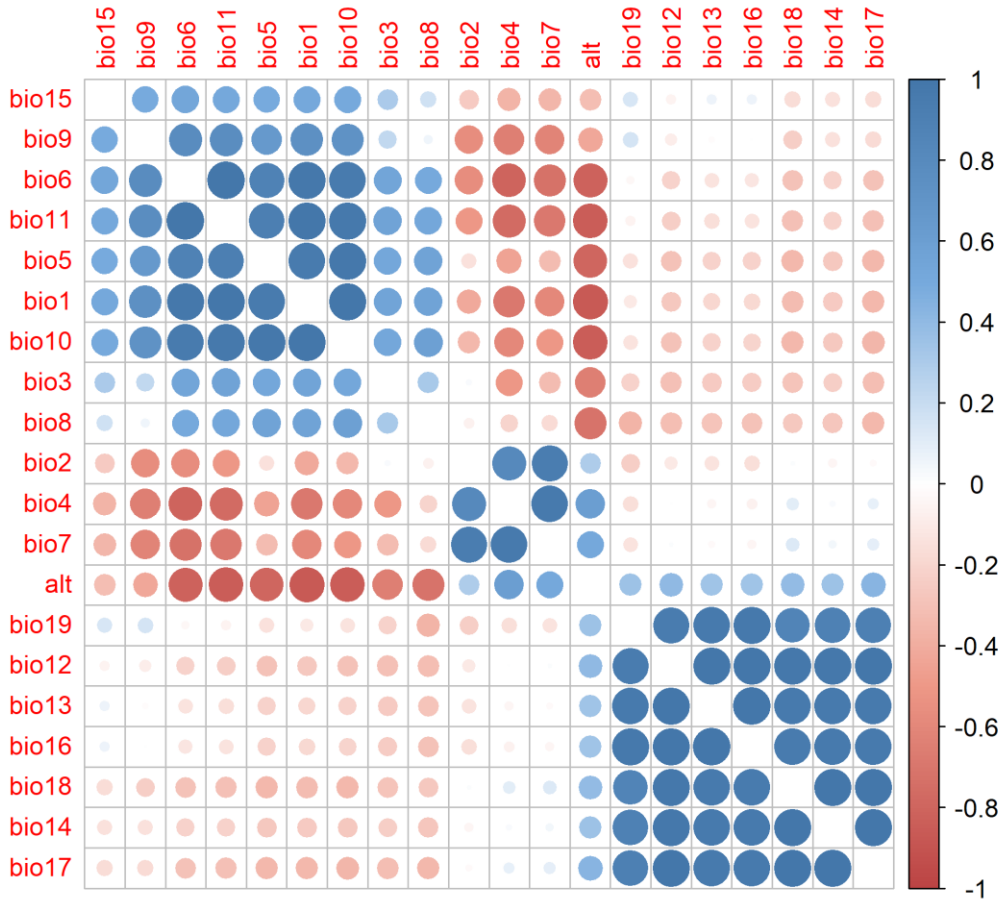


Figure S4 Correlation plot between 19 bioclimate environmental layers (bio1 to bio19) and the elevation layer (alt) that were used for ecological niche modelling of mainland New Zealand species of *Azorella*. The size of the circle indicates correlation level, with larger circles representing higher correlation levels between two environmental layers.

Supplementary Tables

Table S1 Names of species, subspecies and sections of *Azorella* individuals sampled for phylogenetic tree reconstruction (Plunkett and Nicolas, 2017). Ploidy levels were inferred from previously published chromosome numbers (Beuzenberg and Hair, 1983; Hair, 1980; Moore, 1967). Species are ordered first by section, and then by ploidy levels.

Species taxon names	Section	Ploidy level	Chromosome number	Geographical distribution
Azorella ranunculus d'Urv.	*Ranunculus*	2*x*	16	South America
Azorella burkartii (Mathias & Constance) G.M.Plunkett & A.N.Nicolas	*Ranunculus*	?*x*		South America
Azorella lycopodioides Gaud.	*Glabratae*	2*x*	16	South America
Azorella allanii (Cheeseman) G.M.Plunkett & A.N.Nicolas	*Schizeilema*	4*x*	32	North Island
Azorella exigua (Hook.f.) Drude	*Schizeilema*	4*x*	32	South Island
Azorella haastii (Hook.f.) Drude subsp. *haastii*	*Schizeilema*	4*x*	32	South Island
Azorella haastii subsp. *cyanopetala* (Domin) G.M.Plunkett & A.N.Nicolas	*Schizeilema*	4*x*	32	South Island
Azorella hydrocotyloides (Hook.f.) Kirk	*Schizeilema*	4*x*	32	South Island
Azorella roughii (Hook.f.) Kirk	*Schizeilema*	4*x*	32	South Island
Azorella pallida (Kirk) Kirk	*Schizeilema*	6*x*	48	South Island
Azorella cockaynei Diels	*Schizeilema*	6*x*	48	South Island
Azorella hookeri Drude	*Schizeilema*	6*x*	48	North & South Islands
Azorella nitens Petrie	*Schizeilema*	6*x*	48	North & South Islands
Azorella colensoi (Domin) G.M.Plunkett & A.N.Nicolas	*Schizeilema*	10*x*	80	North & South Islands
Azorella fragosea (F.Muell.) Druce	*Schizeilema*	?		Australia
Azorella polaris (Hombr. & Jacquinot ex Hook.f.) G.M.Plunkett & A.N.Nicolas	*Stilbocarpa*	6*x*	48	Subantarctic islands
Azorella lyallii (Armstr.) G.M.Plunkett & A.N.Nicolas	*Stilbocarpa*	?		Subantarctic islands

Table S2 The Illumina ID and PacBio ID of sequenced individuals of New Zealand *Azorella* for phylogenetic tree reconstruction. Voucher specimens are lodged at the following herbaria: CHR, WELT, MPN, CHR and NYBG.

Species	Illumina Tree ID	PacBio Tree ID	Herbarium voucher
Azorella allanii	Azallanii_HikB10_illumina	Azallanii_HikB7_pacbio	WELT SP111286
Azorella burkartii	Azburkartii_NYBG714_illumina	Azburkartii_NYBG714_pacbio	NYBG Ccalv714
Azorella cockaynei	Azcockaynei_Man5_illumina	Azcockaynei_Man5_pacbio	MPN 52530
Azorella colensoi	Azcolensoi_BellA10_illumina	Azcolensoi_BellA2_pacbio	WELT SP110028
Azorella exigua	Azexigua_Wet4_illumina	Azexigua_Wet4_pacbio	MPN 52528
Azorella fragosea	Azfragosea_CANB798456_illumina	Azfragosea_CANB798456_pacbio	CANB 798456
Azorella haastii subsp. *haastii*	Azhaastii_Bri1_illumina	Azhaastii_Bri1_pacbio	WELT SP111342
Azorella haastii subsp. *haastii*	Azhaastii_Pat5_illumina	Azhaastii_Pat5_pacbio	WELT SP111326
Azorella haastii subsp. *haastii*	Azhaastii_Tow10_illumina	Azhaastii_Tow10_pacbio	WELT SP107453
Azorella haastii subsp. *cyanopetala*	Azcyanopetala_Bro5_illumina	Azcyanopetala_Bro5_pacbio	MPN 52529
Azorella haastii subsp. *cyanopetala*	Azcyanopetala_Tak10_illumina	Azcyanopetala_Tak10_pacbio	WELT SP108844
Azorella hookeri	Azhookeri_Cas_illumina	Azhookeri_Cas_pacbio	CHR 674155
Azorella hookeri	NA	Azhookeri_Mon_pacbio	MPN 52533
Azorella hydrocotyloides	Azhydrocotyloides_Dob6_illumina	Azhydrocotyloides_Dob6_pacbio	WELT SP106684
Azorella lycopodioides	Azlycopodioides_NYBG2433_illumina	N/A	NYBG Plunkett 2433
Azorella nitens	Aznitens_Wai3_illumina	Aznitens_Wai3_pacbio	MPN 52525
Azorella nitens	Azpallida_Bro5_illumina	Azpallida_Bro5_pacbio	MPN 52531
Azorella pallida	NA	Aznitens_Boy_pacbio	WELT SP114493
Azorella polaris	Azpolaris_Camp2_illumina	Azpolaris_Camp2_pacbio	CHR 677157
Azorella ranunculus	Azranunculus_NYBG2447_illumina	Azranunculus_NYBG2447_pacbio	NYBG Plunkett 2447
Azorella roughii	Azroughii_Sta5_illumina	Azroughii_Sta5_pacbio	MPN 52523

Table S3 The measured stomatal guard cell length differences (μm) between different samples (silica gel-dried vs. herbarium specimen) in five tested individuals of New Zealand *Azorella* species.

Species	Herbarium voucher	Mean guard cell length (Silica gel-dried) ± SD	Mean guard cell length (Herbarium specimen) ± SD	Difference
A. allanii	MPN 52524	20.6 ± 2.2	21.39 ± 2.52	0.78 (3.78%)
A. haastii subsp. *cyanopetala*	MPN 52529	13.76 ± 2.04	14.46 ± 1.18	0.69 (5.01%)
A. nitens	MPN 52525	17.51 ± 1.52	18.21 ± 1.38	0.7 (4.00%)
A. pallida	MPN 52532	19.42 ± 1.68	20.34 ± 2.15	0.91 (4.68%)
A. roughii	MPN 52534	23.11 ± 3.74	23.39 ± 2.84	0.28 (1.21%)

Table S4 The measured guard cell length (μm) for each *Azorella* taxon. Species names, herbarium voucher, their number of replicates measured and sample source (silica gel-dried vs. herbarium specimen) are included. Species are sorted in alphabetical order, and each taxon is ordered by the measured guard cell length.

Species	Herbarium voucher	*n*	Mean guard cell length ± SD	Sample source
A. allanii	WELT SP110008	3	17.59 ± 1.68	Silica gel-dried
A. allanii	WELT SP108812	3	19.62 ± 1.67	Silica gel-dried
A. allanii	WELT SP111286	3	20.3 ± 2.32	Silica gel-dried
A. allanii	WELT SP111287	3	20.41 ± 2.74	Silica gel-dried
A. allanii	MPN 52524	3	20.6 ± 2.2	Silica gel-dried
A. burkartii	NYBG Ccalv714	1	18.57 ± 1.68	Silica gel-dried
A. burkartii	NYBG GM2434	1	18.75 ± 1.61	Silica gel-dried
A. cockaynei	WELT SP114481	1	17.18 ± 1.63	Silica gel-dried
A. cockaynei	MPN 52530	1	17.97 ± 2.26	Silica gel-dried
A. colensoi	WELT SP110011	3	18.87 ± 1.6	Silica gel-dried
A. colensoi	AK167696	1	19.34 ± 1.3	Herbarium specimen
A. colensoi	WELT SP110035	3	19.93 ± 1.55	Silica gel-dried
A. colensoi	WELT SP110028	3	20.05 ± 1.98	Silica gel-dried
A. colensoi	AK289484	1	20.5 ± 1.42	Herbarium specimen

A. exigua	WELT SP111277	1	13.5 ± 1.27	Silica gel-dried
A. exigua	MPN 52528	1	16.63 ± 2	Silica gel-dried
A. fragosea	CANB798456	1	14.08 ± 1.59	Herbarium specimen
A. fragosea	CANB797854	1	15.24 ± 1.49	Herbarium specimen
A. haastii subsp. *cyanopetala*	MPN 52529	3	13.76 ± 2.04	Silica gel-dried
A. haastii subsp. *cyanopetala*	WELT SP108840	3	15.67 ± 1.42	Silica gel-dried
A. haastii subsp. *cyanopetala*	WELT SP108844	3	16.63 ± 1.91	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP107477	3	15.25 ± 1.25	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP108837	3	15.26 ± 1.58	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP107453	3	15.33 ± 1.79	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP108832	3	15.75 ± 1.54	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP107484	3	16.3 ± 1.43	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP114488	2	18.04 ± 1.78	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP111342	3	18.44 ± 2.55	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP111326	3	18.88 ± 3.71	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP108816	3	19.29 ± 3.51	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP114479	2	19.98 ± 2.59	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP108848	3	22.1 ± 4.06	Silica gel-dried
A. haastii subsp. *haastii*	WELT SP114501	1	22.78 ± 2.29	Silica gel-dried
A. hookeri	AK288199	1	16.82 ± 1.1	Herbarium specimen
A. hookeri	MPN 52537	1	17.15 ± 1.84	Silica gel-dried
A. hookeri	CHR 674155	4	17.87 ± 2.16	Silica gel-dried
A. hookeri	AK351327	1	19.79 ± 2.03	Herbarium specimen
A. hookeri	AK209018	1	19.97 ± 1.61	Herbarium specimen
A. hydrocotyloides	WELT SP106684	3	19.9 ± 2.35	Silica gel-dried
A. hydrocotyloides	WELT SP111280	3	23.4 ± 2.73	Silica gel-dried
A. hydrocotyloides	WELT SP107474	1	23.64 ± 2.72	Silica gel-dried
A. lyallii	CHR 542359	1	17.71 ± 1.29	Herbarium specimen
A. lycopodioides	NYBG GM2433	1	13.92 ± 1.48	Silica gel-dried

A. nitens	MPN 52536	1	16.22 ± 1.68	Silica gel-dried
A. nitens	MPN 52525	2	17.51 ± 1.52	Silica gel-dried
A. nitens	WELT SP114493	3	17.79 ± 2	Silica gel-dried
A. nitens	WELT SP114485	2	18.69 ± 1.65	Silica gel-dried
A. nitens	WELT SP114484	2	18.86 ± 2.29	Silica gel-dried
A. pallida	NYBG GM2533	1	16.46 ± 1.87	Silica gel-dried
A. pallida	MPN 52531	3	16.62 ± 2.02	Silica gel-dried
A. pallida	MPN 52532	2	19.42 ± 1.68	Silica gel-dried
A. polaris	CHR 677157	3	14.8 ± 1.88	Silica gel-dried
A. polaris	CHR 677160	3	16.13 ± 1.71	Silica gel-dried
A. ranunculus	NYBG GM2447A	1	17 ± 1.77	Silica gel-dried
A. ranunculus	NYBG GM2447B	2	17.77 ± 3.33	Silica gel-dried
A. roughii	MPN 52523	3	23.1 ± 3.01	Silica gel-dried
A. roughii	MPN 52534	3	23.11 ± 3.74	Silica gel-dried
A. roughii	MPN 52535	3	23.61 ± 2.61	Silica gel-dried
A. schizeilema	CHR 572151	1	13.82 ± 1.37	Herbarium specimen

Table S5 Total number of reads and the read lengths (in bp) of PacBio sequenced individuals of New Zealand *Azorella*. Individuals are sorted in alphabetical order.

Individual ID	No. of total HiFi reads	0 - 500 bp (%)	500 - 1000 bp (%)	1 - 1.5 kbp (%)	1.5 - 2 kbp (%)	> 2 kbp (%)
Azallanii_HikB7_pacbio	89213	54.68	40.89	4.28	0.13	0.02
Azburkartii_NYBG714_pacbio	67529	28.27	51.12	19.29	1.29	0.04
Azcockaynei_Man5_pacbio	109318	28.77	56.2	14.67	0.35	0.01
Azcolensoi_BellA2_pacbio	127881	59.49	36.23	4.01	0.27	0.01
Azcyanopetala_Bro5_pacbio	300372	13.71	52.1	31.24	2.78	0.17
Azcyanopetala_Tak10_pacbio	221470	23.46	63.04	12.37	1.06	0.08
Azexigua_Wet4_pacbio	138012	27.41	61.31	10.9	0.36	0.02
Azfragosea_CANB798456_pacbio	64025	42.82	45.17	11.46	0.53	0.02
Azhaastii_Bri1_pacbio	220931	10.5	40.57	39.2	9.24	0.5
Azhaastii_Pat5_pacbio	212926	23.59	60.17	15.54	0.65	0.05
Azhaastii_Tow10_pacbio	211578	20.46	54.33	21.87	3.08	0.26
Azhookeri_Cas_pacbio	360801	11.84	40.71	34.43	11.76	1.27
Azhookeri_Mon_pacbio	167781	34.57	61.52	3.87	0.03	0
Azhydrocotyloides_Dob6_pacbio	207579	28.62	56.91	13.86	0.6	0.01
Aznitens_Boy_pacbio	126066	15.78	45.78	32.74	5.42	0.27
Aznitens_Wai3_pacbio	139533	24.63	55.9	18.49	0.96	0.02
Azpallida_Bro5_pacbio	188217	30.15	60.94	8.75	0.15	0
Azpolaris_Camp2_pacbio	149922	20.56	52.48	22.39	4.18	0.39
Azranunculus_NYBG2447_pacbio	10930	34.65	50.92	13.65	0.75	0.03
Azroughii_Sta5_pacbio	64758	25.96	63.08	9.98	0.93	0.05

Table S6 Hybpiper2 output for Illumina- and PacBio-sequenced individuals of *Azorella* (see column, Group). The percentage of on-target reads (PctOnTarget) shows the proportion of mapped reads (ReadsMapped) among the total number of reads (NumReads). Genes with exon contigs extracted (GenesWithSeq), extracted genes with flanking regions (GenesAt150pct), and two categories of paralog detection by length (ParalogWarningsLong) and depth (ParalogWarningsDepth), were calculated for each individual. The genes without (GenesWithoutStitchedContigs) or with (GenesWithStitchedContigs) stitched exons are also shown. Individuals are sorted in alphabetical order.

Individual ID	NumReads	ReadsMapped	PctOnTarget	GenesWithSeqs	GenesAt150pct	ParalogWarningsLong	ParalogWarningsDepth	GenesWithoutStitchedContigs	GenesWithStitchedContigs	Group
Azallanii_HikB10_illumina	6726752	2938154	43.7	284	252	17	30	98	244	illumina
Azallanii_HikB7_pacbio	89213	47852	53.6	250	205	6	8	207	43	pacbio
Azburkartii_NYBG714_illumina	5359878	1388609	25.9	336	306	5	18	124	218	illumina
Azburkartii_NYBG714_pacbio	67529	22740	33.7	185	172	3	3	164	21	pacbio
Azcockaynei_Man5_illumina	7732782	2756112	35.6	267	247	27	71	93	251	illumina
Azcockaynei_Man5_pacbio	109318	54078	49.5	274	258	50	63	214	60	pacbio
Azcolensoi_BellA10_illumina	13593530	5078993	37.4	236	220	17	54	85	256	illumina
Azcolensoi_BellA2_pacbio_Azcol	127881	74436	58.2	254	206	11	12	215	39	pacbio
Azcyanopetala_Bro5_illumina	7401850	3107549	42	289	263	7	24	95	250	illumina
Azcyanopetala_Bro5_pacbio	300372	144003	47.9	325	319	48	63	230	95	pacbio
Azcyanopetala_Tak10_illumina	5803678	2456261	42.3	318	292	14	31	104	240	illumina
Azcyanopetala_Tak10_pacbio	221470	104133	47	319	312	32	46	213	106	pacbio
Azexigua_Wet4_illumina	4897002	1733798	35.4	322	300	11	35	112	230	illumina
Azexigua_Wet4_pacbio	138012	59631	43.2	294	281	30	32	210	84	pacbio
Azfragosea_CANB798456_illumina	110484	16342	14.8	72	2	0	0	62	10	illumina
Azfragosea_CANB798456_pacbio	64025	34142	53.3	241	218	6	8	211	30	pacbio
Azhaastii_Bri1_illumina	6404528	2678521	41.8	329	305	11	37	91	249	illumina
Azhaastii_Bri1_pacbio	220931	107498	48.7	317	315	48	62	246	71	pacbio
Azhaastii_Pat5_illumina	6091628	2243493	36.8	300	283	14	27	99	244	illumina
Azhaastii_Pat5_pacbio	212926	97545	45.8	318	314	34	49	221	98	pacbio
Azhaastii_Tow10_illumina	6331790	2553087	40.3	303	282	13	31	97	246	illumina

Azhaastii_Tow10_pacbio	211578	103237	48.8	310	306	33	43	228	82	pacbio
Azhookeri_Cas_illumina	8862644	4261276	48.1	280	259	18	54	96	246	illumina
Azhookeri_Cas_pacbio	360801	177535	49.2	323	322	93	115	256	67	pacbio
Azhookeri_Mon_pacbio	167781	93121	55.5	290	272	38	45	208	83	pacbio
Azhydrocotyloides_Dob6_illumina	7242830	3113665	43	287	264	16	35	98	244	illumina
Azhydrocotyloides_Dob6_pacbio	207579	106041	51.1	317	311	37	52	225	92	pacbio
Azlycopodioides_NYBG2433_illumina	11483664	3396155	29.6	307	304	12	27	143	200	illumina
Aznitens_Boy_pacbio	126066	58372	46.3	285	275	70	81	234	51	pacbio
Aznitens_Wai3_illumina	6831066	2750941	40.3	286	265	30	77	100	243	illumina
Aznitens_Wai3_pacbio	139533	74239	53.2	284	271	48	55	218	66	pacbio
Azpallida_Bro5_illumina	6405592	2754676	43	306	279	9	24	101	244	illumina
Azpallida_Bro5_pacbio	188217	92394	49.1	311	303	27	37	207	104	pacbio
Azpolaris_Camp2_illumina	8852854	3099283	35	289	274	27	67	107	239	illumina
Azpolaris_Camp2_pacbio	149922	57459	38.3	274	268	58	66	226	48	pacbio
Azranunculus_NYBG2447_illumina	625324	79651	12.7	204	34	1	1	145	59	illumina
Azranunculus_NYBG2447_pacbio	10930	4486	41	14	2	0	0	12	2	pacbio
Azroughii_Sta5_illumina	6953656	2412261	34.7	317	302	9	15	117	226	illumina
Azroughii_Sta5_pacbio	64758	28791	44.5	214	179	11	12	191	23	pacbio

Table S7 Hybpiper2 output for the Illumina and PacBio merged dataset of New Zealand species of *Azorella*. See Table S6 for additional information. Individuals are sorted in alphabetical order.

Individual ID	NumReads	ReadsMapped	PctOnTarget	GenesWithSeqs	GenesAt150pct	ParalogWarningsLong	ParalogWarningsDepth	GenesWithoutStitchedContigs	GenesWithStitchedContigs	Group
Azallanii_HikB7_merged	2062405	937475	45.5	317	289	46	59	208	109	merged
Azburkartii_NYBG714_merged	1140225	324560	28.5	309	288	15	24	232	77	merged
Azcockaynei_Man5_merged	1264618	604546	47.8	331	320	140	187	205	126	merged
Azcolensoi_BellA2_merged	2379302	1329084	55.9	326	286	85	130	206	120	merged
Azcyanopetala_Bro5_merged	2123768	995296	46.9	337	335	129	165	225	112	merged
Azcyanopetala_Tak10_merged	1494154	745825	49.9	334	331	104	140	205	129	merged
Azexigua_Wet4_merged	791906	366681	46.3	327	320	73	105	211	116	merged
Azfragosea_CANB798456_merged	90564	38023	42	246	221	27	28	216	30	merged
Azhaastii_Bri1_merged	1724365	833459	48.3	331	329	112	135	248	83	merged
Azhaastii_Pat5_merged	1459762	642010	44	336	332	97	132	219	117	merged
Azhaastii_Tow10_merged	1645434	778369	47.3	337	331	99	129	230	107	merged
Azhookeri_Cas_merged	2471469	1556368	63	333	332	193	230	265	68	merged
Azhookeri_Mon_pacbio	167781	93121	55.5	290	272	89	102	208	82	pacbio
Azhydrocotyloides_Dob6_merged	1966386	950442	48.3	338	331	106	132	220	118	merged
Azlycopodioides_NYBG2433_illumina	11483664	3396155	29.6	301	298	25	42	144	201	illumina
Aznitens_Boy_pacbio	126066	58372	46.3	285	275	119	134	234	51	pacbio
Aznitens_Wai3_merged	1334864	774626	58	328	319	129	170	212	116	merged
Azpallida_Bro5_merged	1856900	877750	47.3	342	334	98	127	210	132	merged
Azpolaris_Camp2_merged	2006230	801107	39.9	330	325	138	152	238	92	merged
Azranunculus_NYBG2447_merged	162514	23359	14.4	56	15	0	0	49	7	merged
Azroughii_Sta5_merged	1136617	524657	46.2	324	306	71	94	206	118	merged

Table S8 Measured genome sizes of 13 *Azorella* taxa from 28 individuals. For information about genetic groups and ploidy level see Table S1 and Fig.1. Species within each genetic groups are in alphabetical order, and by increasing 2C values within each taxon.

Species	Individual ID	Voucher ID	Standard	Group	Ploidy	CV (%)	CV-standard (%)	2C (pg)
A. allanii	Azallanii_HikB	WELT SP111286	Pea	NZ1	4*x*	3.49	3.69	6.59
	Azallanii_Mau	WELT SP110008	Pea	NZ1	4*x*	3.3	3.13	6.68
	Azallanii_Rau	WELT SP111287	Pea	NZ1	4*x*	3.61	3.14	6.76
A. roughii	Azroughii_Rob	MPN 52535	Broadbean	NZ1	4*x*	4.62	2.58	7.95
	Azroughii_StA	MPN 52534	Broadbean	NZ1	4*x*	5.01	3.61	8.23
A. cockaynei	Azcockaynei_Cer	WELT SP114481	Broadbean	NZ1	6*x*	5.92	3.94	8.37
	Azcockaynei_Man	MPN 52530	Broadbean	NZ1	6*x*	4.37	3.02	8.81
A. hookeri	Azhookeri_Mon	MPN 52533	Broadbean	NZ1	6*x*	5.58	3.68	8.71
	Azhookeri_Tai	MPN 52537	Broadbean	NZ1	6*x*	4	3.16	8.77
	Azhookeri_Cas	CHR 674155	Broadbean	NZ1	6*x*	3.8	3.07	8.84
A. nitens	Aznitens_CasB	WELT SP114485	Broadbean	NZ1	6*x*	4.47	3.37	8.25
	Aznitens_Boy	WELT SP114493	Broadbean	NZ1	6*x*	4.76	3.79	8.29
	Aznitens_CasA	WELT SP114484	Broadbean	NZ1	6*x*	6.01	3.93	8.35
	Aznitens_Wai	MPN 52525	Broadbean	NZ1	6*x*	4.9	4.11	8.78
A. colensoi	Azcolensoi_BellA	WELT SP110028	Pea	NZ1	10*x*	2.95	3.94	14.07
	Azcolensoi_MauA	WELT SP110011	Pea	NZ1	10*x*	2.37	2.96	14.86
	Azcolensoi_MauB	WELT SP110035	Pea	NZ1	10*x*	2.52	3.1	14.86
A. hydrocotyloides	Azhydrocotyloides_Dun	WELT SP111280	Pea	NZ2	4*x*	4.52	3.26	4.32
A. haastii subsp. *cyanopetala*	Azcyanopetala_Bro	MPN 52529	Pea	NZ2	4*x*	4.71	4.66	4.53
A. pallida	Azpallida_Bro	MPN 52531	Pea	NZ2	4*x*	5.39	3.81	4.53
	Azpallida_Cra	MPN 52532	Pea	NZ2	4*x*	6.54	5.05	4.54
A. haastii subsp. *haastii*	Azhaastii_Ach	WELT SP114501	Pea	NZ2	4*x*	3.41	3.47	5.46
	Azhaastii_Cou	WELT SP114488	Pea	NZ2	4*x*	5.41	3	5.49
	Azhaastii_Cer	WELT SP114479	Pea	NZ2	4*x*	4.8	3.6	5.61
	Azhaastii_Bri	WELT SP111342	Pea	NZ2	4*x*	4.85	3.62	5.64
A. exigua	Azexigua_Cara	MPN 52528	Pea	NZ2	4*x*	4.32	3.45	6.76

A. lyallii	Azlyallii_Ste	WELT SP114496	Broadbean	Sub	?*x*	5.45	2.86	7.83
A. polaris	Azpolaris_End	CHR 677160	Broadbean	Sub	6*x*	3.78	2.55	8.09

Table S9 The important environmental predictors in ecological niche modelling using ENMTools for species of mainland New Zealand *Azorella*. Darker colours represent more important predictors.

	BIO2	BIO3	BIO5	BIO8	BIO9	BIO13	BIO15
A. allanii	0.01	0.03	0	0.21	0.02	0.03	0
A. cockaynei	0	0.08	0.15	0	0.03	0.1	0
A. colensoi	0	0.01	0.02	0.41	0.01	0.01	0.01
A. exigua	0.18	0	0.01	0.02	0	0	0.02
A. haastii subsp. *cyanopetala*	0	0	0.41	0.02	0.02	0	0
A. haastii subsp. *haastii*	0.01	0.03	0.09	0.02	0.04	0.07	0
A. hookeri	0.02	0.01	0.17	0.01	0.06	0.03	0.07
A. hydrocotyloides	0.01	0	0.34	0.05	0.04	0.03	0.01
A. nitens	0.02	0.02	0.23	0	0.02	0.09	0.02
A. pallida	0.05	0.05	0	0	0.05	0.01	0
A. roughii	0.13	0	0.24	0	0.01	0.01	0.01

Chapter 4. Thesis summary and perspective

The New Zealand polyploid-rich genus *Azorella* was shown here to be a useful model to investigate species diversification due to whole genome duplication (WGD) (Murray *et al.*, 2005; Plunkett & Nicolas, 2017). In this thesis, phylogenomic methods and bioinformatic tools for studying polyploid species relationships were first introduced in Chapter 1. In Chapter 2, two phylogenomic approaches, single copy nuclear gene capture via Hyb-Seq employing the Angiosperms353 bait set (Johnson *et al.*, 2018) and genome-skim sequencing of nrDNA and plastome DNA, were used to resolve the origins, phylogenetic relationships, and biogeographical histories of New Zealand *Azorella* species. The post-polyploidization macroevolutionary patterns of New Zealand *Azorella* were then compared in Chapter 3 using polyploidy-associated traits, including genome sizes, stomatal guard cell length, homeologous copy number variation of target-enriched genes, and environmental niche space. Finally, Chapter 4 presents a summary of the thesis and offers future perspectives.

4.1 Aims of Thesis

This thesis aimed to answer two main questions regarding the polyploid-rich genus, *Azorella,* in New Zealand: 1) What do analyses of phylogenomic data tell us about the origins and relationships of these species, and can phylogenetic inference of polyploid–rich genera be improved? and 2) Can comparison of genomic, morphological, and ecological traits among polyploids provide insight into their post-WGD diversification ?

To improve on previous phylogenetic inference of *Azorella* in New Zealand, a total of 125 individuals representing 20 *Azorella* taxa from 72 sites in New Zealand and South America were sampled. In addition to comprehensive sampling, phylogenetic inference of *Azorella* polyploids (Chapter 2 and Chapter 3) was also improved by: 1) using Hyb-Seq with the Angiosperms353 bait set and capturing genome-wide nuclear markers; 2) inferring networks using the discordance levels among target-enriched gene trees; 3) comparing the topological incongruence between single copy and high copy gene trees; 4) including additional homeologous gene copies of the target-enriched genes in some analyses; and 5) determining the intraspecific variation within non-monophyletic species (e.g., three genetic groups in *A. haastii* subsp. *haastii*) using SNPs extracted from Hyb-Seq data.

Overall, polyploidization and reticulation contributed to the origin of several lineages of New Zealand *Azorella*, which likely explains previous difficulties with phylogenetic reconstruction of

this group. For example, Chapter 2 networks identified the hybrid origins of the NZ1 group of hexaploids (*A. hookeri*, *A. nitens*, *A. cockaynei*) and the NZ2 group of tetraploids (*A. exigua*, *A. haastii* subsp. *haastii*, *A. haastii* subsp. *cyanopetala*, *A. hydrocotyloides* and *A. pallida*), and the allopolyploid origin of the sole New Zealand decaploid *A. colensoi* from the hexaploid maternal parent, *A. hookeri*, and the tetraploid paternal parent, *A. allanii*. Although the phylogenies based on homogenized nrDNA copies and maternally inherited plastomes showed limited usage or even misleading results when compared to the single copy nuclear gene tree, the Hyb-Seq networks in turn explained the different genetic groups within and between NZ *Azorella* nrDNA and plastome gene trees (Chapter 2), which further highlights the importance of combining different phylogenomic approaches for phylogenetic inference of polyploid-rich genera.

Comparison of polyploidy-associated traits in Chapter 3 not only provided additional information about species relationships, but also offered insight into the post-WGD macroevolutionary patterns of New Zealand *Azorella*. Specifically, the genomic trait of holoploid 2C value was strongly phylogenetically correlated and thus informative about species relationships. The holoploid 2C genome sizes of the hexaploids (*A. hookeri*, *A. nitens*, *A. cockaynei*, and *A. polaris*), as well as their high number of homeologs, suggest their subgenome donors might originate from more dissimilar genomes, which may include taxa that have similar 2C genome size in South American sections *Huanaca*, *Ranunculus* or *Azorella* that previously showed unresolved phylogenetic relationships with New Zealand *Azorella* sections *Stilbocarpa* and *Schizeilema* (Plunkett & Nicolas, 2017; Ptáček *et al.*, 2022). Indeed, phylogenetic incongruence of nrDNA vs. single copy nuclear gene tree in Chapter 2 also revealed the hybrid origins of *Azorella* hexaploids in NZ (*A. hookeri*, *A. nitens*, and *A. cockaynei*) that are related to South American relatives. In addition, allopolyploid, *A. colensoi* (2C = 14.60 pg) showed an additive genome size of its parental species, *A. hookeri* (2C = 8.77 pg) and *A. allanii* (2C = 6.68 pg). On the other hand, another genomic trait, monoploid 1Cx genome size, as well as environmental niche comparison, shed light on New Zealand *Azorella* post-WGD diversification patterns, including expansion or contraction of genome size, and niche shift or niche conservatism. Such patterns varied for taxa in different genetic groups, which may relate to the origins of the species (including their subgenome donors), reticulate histories, and their age.

4.2 Future Perspectives

Unexpected delays and difficulties due to the COVID-19 pandemic meant that additional research questions were unable to be fully addressed in this thesis. For future studies of New Zealand *Azorella*, two main future directions are listed below and discussed briefly.

4.2.1 Taxonomic Implications

Rapid and reticulate speciation within polyploid-rich genera can lead to similar morphological traits and overlapping ecological habitats, which can cause difficulties for taxonomic classification of species. This situation is encountered frequently in New Zealand polyploid-rich genera such as *Veronica* (Wagstaff *et al.*, 2002; Thomas *et al.*, 2021), *Myosotis* (Prebble *et al.*, 2019; Meudt, 2021), and *Plantago* (Meudt, 2011). In addition to potential new and as yet unnamed *Azorella* species (A.sp_CHR617214, A.sp_CHR617283 and A. sp_AN58) listed in Chapter 2, New Zealand *Azorella* polyploids exhibited similar morphological characteristics, which presented challenges to identify some species when using the identification keys in Allan (1961). These new species need formal description (to be undertaken by collaborator Dr Peter Heenan, Manaaki-Whenua Landcare Research) and the key to New Zealand *Azorella* needs to be updated given the new species and current phylogenetic knowledge.

Three trifoliate-leaved plants, *Azorella nitens*, *A. hookeri* and *A. colensoi*, are genetically divergent but have similar morphological characteristics (Chapter 2). The whole leaf size of *A. nitens* (< 10 mm) is often smaller compared to *A. hookeri* and *A. colensoi* (~15 mm) (Allan, 1961). Whereas *A. nitens* and *A. hookeri* leaflets can be either entire or lobed (Fig. 4.1; Fig 4.2), the whole leaf size may be variable in different populations. Furthermore, *A. hookeri* and its allopolyploid descendant *A. colensoi* are very morphologically similar to one another (Fig. 4.2; Fig. 4.3), so that the current key (Allan, 1961) does not adequately identify them. More intensive morphological studies to quantify the variables in the leaf traits (e.g., size, trichomes, leaf thickness, number of leaflets, etc.), number of flowers per inflorescence, or microscale phenotype traits (stomatal guard cell structure, pollen structure, etc.) will be required for taxonomic revision.

Figure 4.1 Leaf morphological variation within *Azorella nitens*. The images were modified from iNaturalist records. a) Westland region; **https://inaturalist.nz/observations/4257476**, iNaturalist.nz © Alex Fergus; b) Otago region; **https://inaturalist.nz/observations/69845630**. iNaturalist.nz © Dave Holland; c) Wellington region; **https://inaturalist.nz/observations/2567861**, iNaturalist.nz © Pat Enright.

Figure 4.2 Leaf morphological variation within *Azorella hookeri*. The images were modified from iNaturalist records. a) Canterbury region, https://inaturalist.nz/observations64920874, iNaturalist.nz © Alice Shanks; b) Otago region, https://inaturalist.nz/observations/7670183, iNaturalist.nz © John Barkla; d) Hawke's Bay region, https://inaturalist.nz/observations/10027159, iNaturalist.nz © Alex Fergus; d) Nelson region, https://inaturalist.nz/observations/109791185, iNaturalist.nz © Chris Ecroyd.

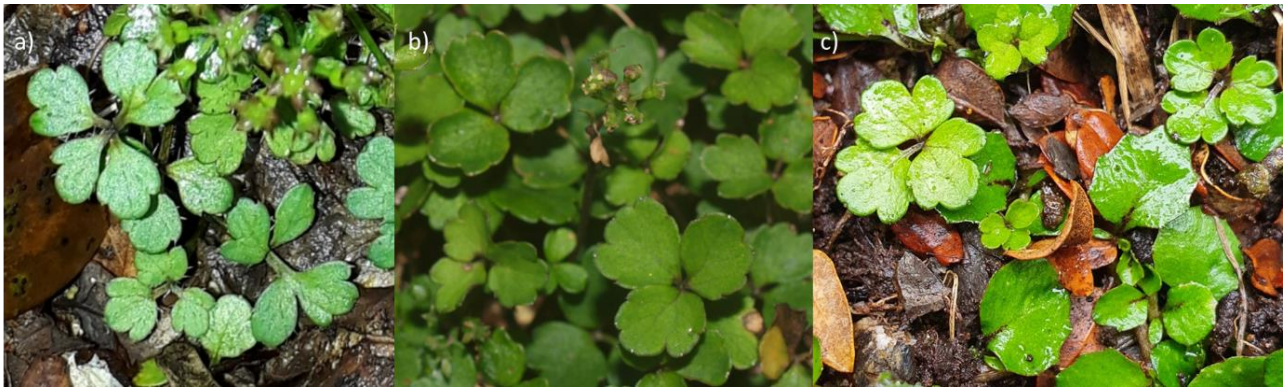


Figure 4.3 Leaf morphological variation within *Azorella colensoi*. The images were modified from iNaturalist records. a) Manawatu-Wanganui region, **https://inaturalist.nz/observations/9442149**, iNaturalist.nz © Leon Perrie; b) Hawke's Bay region, **https://inaturalist.nz/observations/2619947**, iNaturalist.nz © Mike Lusk; c) Manawatu-Wanganui region, **https://inaturalist.nz/observations/80523717**, iNaturalist.nz © Oscar Grant.

Additionally, one specimen of *A. nitens* [WELT SP114484; (https://collections.tepapa.govt.nz/object/2039735)] exhibited an intermediate leaf morphology between partially lobed and simple leaf, similar to *A. cockaynei*, that has not been described before. The sister species pair *A. nitens* and *A. cockaynei* sometimes are non-monophyletic to each other in nuclear phylogenetic trees (nrDNA, Hyb-Seq single copy gene tree; Chapter 2 and Chapter3). In the future studies, increasing the sampling of the widely distributed *A. nitens* will be required to determine the variation within this species and the levels of genetic divergence to its sister species *A. cockaynei*.

Two South Island endemic subspecies of *A. haastii*, i.e. *A. haastii* subsp. *haastii* and *A. haastii* subsp. *cyanopetala*, are very similar in their morphological characteristics, and both are common on the western side of the Southern Alps. However, the two subspecies are not sister to one another in any phylogeny (Chapter 2). Among *A. haastii* subsp. *haastii* sampled individuals, there were three genetic groups represented (Chapter 2). In addition, *A. haastii* subsp. *cyanopetala* also has a smaller 2C genome size (on average, 4.53 pg) compared to *A. haastii* subsp. *haastii* (5.55 pg) (Chapter 3). Therefore, *A. haastii* subsp. *cyanopetala* may be recognized at the species rank, instead of as a subspecies of *A. haastii*. However, making this taxonomic change is outside the scope of this thesis and a future taxonomic revision is required to define the boundaries within *A. haastii* subsp. *haastii*, and between *A. haastii* subsp. *haastii* and *A. haastii* subsp. *cyanopetala*, to determine the most appropriate rank.

4.2.2 Chromosome Counting

Chromosome counting is essential to determine the ploidy level of a polyploid species in an effort to understand post-WGD diversification (Otto & Whitton, 2000; Moraes *et al.*, 2022). This is especially true for the New Zealand flora, as many of its species have gone through multiple rounds of WGD (Murray *et al.*, 2005; Meudt *et al.*, 2021). In this thesis, the chromosomes of New Zealand *Azorella* were attempted to be recounted using harvested root tips for a few individuals with the assistance of Prashant Joshi (Massey University, Palmerston North). The root tips were cut in 0.5 cm long pieces and pre-treated with 0.02mol/L 8-hydroxyquinoline for 3 hours, then fixed in 3:1 ethanol:acetic acid fixative overnight (Viruel *et al.*, 2019). The fixed root tips were cleaned with distilled water and placed on glass slides. The chromosomes of root tip cells were stained with propidium iodide (PI, Sigma) (1 mg/mL), and then root tips were squashed by adding a drop of 45% acetic acid to help to separate the individual cells. However, no actively dividing cells with clear chromosomes were able to be counted under a dissecting microscope. Unfortunately, due to

COVID-19 restrictions, many greenhouse grown plants died before additional counts could be attempted. On the other hand, flow cytometry is often used to infer ploidy level of plants when compared to a known diploid (Dolezel, 1997). However, the variable genome sizes within tetraploids, i.e., the smallest in *A. hydrocotyloides* 2C = 4.32 pg to the largest *A. roughii* 2C = 8.09 pg, indicates that flow cytometry may not be able to measure ploidy level variation accurately. Especially the tetraploid *A. roughii* showed similar 2C genome size as the hexaploids (2C around 8 to 8.5 pg; *A. polaris*, *A. nitens*, *A. cockaynei* and *A. hookeri*) in New Zealand. In the future, it will be necessary to confirm the chromosome number of New Zealand *Azorella* to fully understand the 1Cx genome sizes changes.

4.2.3 Phylogenomic Implications

Combining Hyb-Seq and genome-skimming was a useful approach for phylogenetic study of the New Zealand polyploid-rich genus *Azorella*. The Angiosperm353 bait set targeted 353 single copy nuclear genes (Johnson *et al.*, 2018), which were highly conserved at the genus level in *Azorella* (at least 340 loci were assembled with contigs, Chapter 2), and even to subfamily level of Azorelloideae [over *c.* 300 loci (Clarkson *et al.*, 2021)]. Therefore, this bait set may be particularly useful to resolve South American *Azorella* species relationships further (Plunkett & Nicolas, 2017), especially the lineages involved in WGD (Ptáček *et al.*, 2022). However, extracting the homeologs of targeted loci remains challenging, and requires improvements of capturing the whole targeted gene sequence (exons and introns) (Chapter 3), as well as removing the homologous gene copies. In the future, Hyb-Seq and genome-skimming sequenced reads from this study may assist in the development of genus-level or family-level Hyb-Seq bait set design to include not only exons but also introns from more genetic markers.

4.3 Reference cited

Allan HH 1961. *Flora of New Zealand. Vol. I.*: R.E. Owen Government Printer, Wellington, New Zealand.

Clarkson JJ, Zuntini AR, Maurin O, Downie SR, Plunkett GM, Nicolas AN, Smith JF, Feist MAE, Gutierrez K, Malakasi P, et al. 2021. A higher-level nuclear phylogenomic study of the carrot family (Apiaceae). *American Journal of Botany* 108(7): 1252-1269.

Dolezel J. 1997. Application of flow cytometry for the study of plant genomes. *Journal of Applied Genetics* 38(3): 285-302.

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, et al. 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68(4): 594-606.

Meudt HM. 2011. Amplified fragment length polymorphism data reveal a history of auto- and allopolyploidy in New Zealand endemic species of *Plantago* (Plantaginaceae): new perspectives on a taxonomically challenging group. *International Journal of Plant Sciences* 172(2): 220-237.

Meudt HM. 2021. Taxonomic revision of five species groups of ebracteate-erect *Myosotis* (Boraginaceae) endemic to New Zealand, based on morphology, and description of new subspecies. *Australian systematic botany* 34(3): 252-304.

Meudt HM, Albach DC, Tanentzap AJ, Igea J, Newmarch SC, Brandt AJ, Lee WG, Tate JA. 2021. Polyploidy on islands: its emergence and importance for diversification. *Frontiers in Plant Science* 12(336).

Moraes AP, Engel TBJ, Forni-Martins ER, de Barros F, Felix LP, Cabral JS. 2022. Are chromosome number and genome size associated with habit and environmental niche variables? Insights from the Neotropical orchids. *Annals of Botany* 130(1): 11-25.

Murray BG, De Lange PJ, Ferguson AR. 2005. Nuclear DNA variation, chromosome numbers and polyploidy in the endemic and indigenous grass flora of New Zealand. *Annals of Botany* 96(7): 1293-1305.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34(1): 401-437.

Plunkett GM, Nicolas AN. 2017. Assessing *Azorella* (Apiaceae) and its allies: Phylogenetics and a new classification. *Brittonia* 69(1): 31-61.

Prebble JM, Meudt HM, Tate JA, Symonds VV. 2019. Comparing and co-analysing microsatellite and morphological data for species delimitation in the New Zealand native *Myosotis* pygmaea species group (Boraginaceae). *Taxon* 68(4): 731-750.

Ptáček J, Sklenář P, Pinc J, Urfusová R, Calviño CI, Urfus T. 2022. A pentaploid endosperm and a Penaea-type embryo sac are likely synapomorphies of *Azorella* (Apiaceae, Azorelloideae). *Plant Systematics and Evolution* 308(6): 40.

Thomas AE, Igea J, Meudt HM, Albach DC, Lee WG, Tanentzap AJ. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *American Journal of Botany* 108(7): 1289-1306.

Viruel J, Conejero M, Hidalgo O, Pokorny L, Powell RF, Forest F, Kantar MB, Soto Gomez M, Graham SW, Gravendeel B, et al. 2019. A Target Capture-Based Method to Estimate Ploidy From Herbarium Specimens. *Frontiers in Plant Science* 10.

Wagstaff SJ, Bayly MJ, Garnock-Jones PJ, Albach DC. 2002. Classification, origin, and diversification of the New Zealand hebes (Scrophulariaceae). *Annals of the Missouri botanical garden* 89(1): 38-63.