Genome-wide patterns of genetic diversity, population structure and demographic history in mānuka (Leptospermum scoparium) growing on indigenous Māori land

Emily Koot, Elise Arnst, Melissa Taane, Kelsey Goldsmith, Amali Thrimawithana, Kiri Reihana, Santiago C. Gonzalez-Martinez, Victor Goldsmith, Gary Houliston, David Chagné

Horticulture Research

Horticulture Research, 2022, **9**: uhab012
https://doi.org/10.1093/hr/uhab012

Article

Genome-wide patterns of genetic diversity, population structure and demographic history in mānuka (*Leptospermum scoparium*) growing on indigenous Māori land

Emily Koot[1], Elise Arnst[2], Melissa Taane[1], Kelsey Goldsmith[3], Amali Thrimawithana[4], Kiri Reihana[2], Santiago C. González-Martínez[5], Victor Goldsmith[3], Gary Houliston[2] and David Chagné[1,*]

[1]The New Zealand Institute for Plant and Food Research Limited (Plant & Food Research), Batchelar Rd, Palmerston North 4410, New Zealand
[2]Manaaki Whenua Landcare Research, 54 Gerald St, Lincoln 7608, New Zealand
[3]Kaiaka Consulting Limited, Wellington 5016, New Zealand
[4]Plant & Food Research, 120 Mt Albert Rd, Auckland 1025, New Zealand
[5]Université de Bordeaux, INRAE, UMR BIOGECO, 33610 Cestas, France
*Corresponding author: E-mail: David.Chagne@plantandfood.co.nz

Abstract

Leptospermum scoparium J. R. Forst et G. Forst, known as mānuka by Māori, the indigenous people of Aotearoa (New Zealand), is a culturally and economically significant shrub species, native to New Zealand and Australia. Chemical, morphological and phylogenetic studies have indicated geographical variation of mānuka across its range in New Zealand, and genetic differentiation between New Zealand and Australia. We used pooled whole genome re-sequencing of 76 *L. scoparium* and outgroup populations from New Zealand and Australia to compile a dataset totalling ~2.5 million SNPs. We explored the genetic structure and relatedness of *L. scoparium* across New Zealand, and between populations in New Zealand and Australia, as well as the complex demographic history of this species. Our population genomic investigation suggests there are five geographically distinct mānuka gene pools within New Zealand, with evidence of gene flow occurring between these pools. Demographic modelling suggests three of these gene pools have undergone expansion events, whilst the evolutionary histories of the remaining two have been subjected to contractions. Furthermore, mānuka populations in New Zealand are genetically distinct from populations in Australia, with coalescent modelling suggesting these two clades diverged ~9–12 million years ago. We discuss the evolutionary history of this species and the benefits of using pool-seq for such studies. Our research will support the management and conservation of mānuka by landowners, particularly Māori, and the development of a provenance story for the branding of mānuka based products.

Introduction

According to Māori lore, all trees in Aotearoa are the children of Tāne Mahuta, God of the forest [1, 2]. Māori have many uses for the taonga (culturally significant/treasure) tree species mānuka (*L. scoparium* J. R. Forst et G. Forst) including for medicine [3, 4], food [5, 6], hunting [7], fishing [6, 8], weaponry [6, 9, 10] and as a building material [6]. More recently, mānuka has been used to produce high value honey, including by Māori-owned agribusinesses, pushing the export value of the New Zealand honey industry to $348 million (NZD) in 2018 [11, 12]. Indigenous communities and businesses across New Zealand are seeking knowledge about the genetic variation and evolutionary history of mānuka as it is a culturally, ecologically and economically important species. This is so that an authentic and higher value honey industry based on indigenous plants of known New Zealand provenance can be established, and so that this genetic resource can be managed more sustainably for the future.

The genus *Leptospermum* J. R. Forst et G. Forst of the Myrtaceae family comprises approximately 87 species, predominantly distributed throughout south-east Australia (Victoria, New South Wales, Tasmania); species are also present in South East Asia, New Guinea, Rarotonga and New Zealand [13–18]. Specifically, one native species resides in New Zealand – *L. scoparium*. *L. scoparium*, which is also native to south-east Australia, is a woody tree species, distributed from coastal to sub-alpine environments (1800 m above sea level) [11]. In New Zealand, *L. scoparium* is indigenously known as mānuka or kahikātoa, and is found across all of New

Received: June 22, 2021; **Accepted:** September 2, 2021; **Published:** 20 January 2022

Zealand's major islands, including the Chatham Islands [19]. Mānuka is able to establish in adverse environments (e.g. low fertility soils, peat swamps, volcanic soils, geothermal areas, exposed coastal, sub alpine) and is a primary coloniser of recently disturbed habitats (e.g. post-fire and deforestation) [19, 20].

Current understanding of the population structure within New Zealand mānuka is predominantly based on studies of its antibacterial metabolic compounds and morphological variation [20–23]. Perry, *et al.* [22] were able to establish via common garden experiments that regional differences in mānuka's triketone chemotypes are not determined by abiotic factors, demonstrating there is a genetic contribution to the regional differentiation of essential oil compounds in mānuka. Additionally, Perry, *et al.* [22] identified distinct chemotypic variation between *L. scoparium* samples from Australia and New Zealand, and recommended a taxonomic revision of this species. Expanding on this earlier work, Douglas, *et al.* [21] carried out extensive chemotypic sampling of mānuka across New Zealand, elucidating 11 chemotypes that displayed degrees of geographic association. Unique compounds are also found in the nectar of mānuka (dihydroxyacetone) and its honey (methylglyoxal) [24–27], leading to the development of several medical wound care products in recent years, as well as demand as a health food. Studies by Williams, *et al.* [28] and Stephens [29] indicate regional variation in the dihydroxyacetone content of mānuka's nectar and the non-peroxide antibacterial activity of mānuka honey, respectively, across New Zealand. Additionally, morphological and phenological variation of mānuka across New Zealand was found to correlate with geographic and environmental factors, and common garden experiments indicated that differences were likely linked to genetic variation [20].

Recently, a molecular study of mānuka was published [30], revealing the phylogenetic grouping of New Zealand mānuka into three clades – a Northland clade, a central and southern North Island clade and a South Island clade. However, as an inadequate sample size prevented Buys *et al.* [30] from carrying out population genetic analyses, the population genetic structure of mānuka across New Zealand remains unexplored. Similarly, gaps remain in our knowledge regarding the demographic and evolutionary history of *L. scoparium* in New Zealand and the relationships between New Zealand and Australian populations of the species. It is hypothesised that *L. scoparium* arrived in New Zealand from Australia (either from the mainland or from Tasmania) during the Miocene (23.03–5.33 Mya) [13, 32]. During this time, New Zealand has been subjected to geological events (mountain building, volcanism) [33, 34] and glacial cycling (12000–5 Mya) [35, 36] that may have shaped the diversity and distribution of mānuka following its establishment in New Zealand.

The use of pooled sequencing (pool-seq) to obtain genome-wide variants provides a relatively economical means to study genomic variation at a population scale [37–42]. Equal quantities of high-quality DNA from multiple individuals (preferably n= >25 [43],) are pooled for each population and sequenced using next-generation sequencing (NGS). Mapping to a high-quality reference genome reveals genetic variants in the NGS data – commonly single nucleotide polymorphisms (SNPs). Provided that read coverage across sites is high (>100x) to ensure reliability [43, 44], various analyses can then be carried out, including: population genomic analyses [37, 45, 46]; genome wide association studies [47, 48]; gene environment association studies [38, 49]; and the modelling of evolutionary histories [50, 51]. Pool-seq is an increasingly popular method, and provided the correct quality control steps are taken, validation studies have proven it to be a valuable and informative tool [39, 44, 52, 53]. Alternatives to pool-seq include genotype-by-sequencing (GBS) methods, however such approaches only capture a fraction of the genome variants, are error-prone and produce missing data [54].

A complete assembly of the mānuka genome was recently developed, scaffolded into the expected 11 pseudo-chromosomes, which are syntenic to the related Myrtaceae model species *Eucalyptus grandis* [55]. In the present study we applied pooled whole genome re-sequencing of mānuka and kānuka (*Kunzea robusta* de Lange et Toelken, used as outgroup) collected from naturally occuring stands on indigenous Māori owned land across New Zealand and in native stands on public land in Australia to identify genome-wide DNA variants. With this data, we differentiate local provenances of *L. scoparium* within New Zealand, and between New Zealand and Australian populations, as well as provide insights into the complex demographic and evolutionary history of this species.

Results

Whole genome re-sequencing

The DNA of approximately 30 trees was pooled for each New Zealand site (n = 68) and 15 to 30 trees from each Australian site (n = 6) (Supplementary material 1; 2). For some populations we were unable to extract adequate DNA from all 30 samples, thus pools for some populations comprised fewer than 30 individuals. From the Novaseq 6000 sequencing runs a total of 5811.53×10^9 nucleotidic bases ($(1 \times 10^9 = \text{Gb})$ from 19.27×10^9 sequencing reads were generated, with an average 253.61 billion sequencing reads and 76.46 Gb per collection site. The quality of the sequencing reads was high, with most sequences having a quality score greater than Q30 across the 150 bp reads (data not shown). Read mapping to the reference genome was consistent among populations and ranged between 50 and 64% of the *L. scoparium* reads mapping to the reference genome of *L. scoparium* "Crimson Glory" and 26% of the kānuka reads mapping to the mānuka reference genome (Supplementary material 3). The remaining sequences

Table 1. Whole genome variant detection in *Leptospermum scoparium* and *Kunzea robusta* using pool sequencing. Seven different datasets are shown, based on including *Kunzea robusta* (kānuka) or not, including Australian populations or not and with minor allele frequencies (MAF) of 2% and 5% applied. Variant filtering was performed with VCFtools. Only single nucleotide polymorphisms (SNPs) with no missing data and a coverage greater than 100X were included (filtered SNPs)

Number of populations	MAF	Filtered SNPs	Filtered SNP density
76 populations (All)	0%	5 503 881	45.21
76 populations (All)	2%	2 513 694	98.99
76 populations (All)	5%	1 498 405	166.07
74 populations (No Kānuka)	2%	4 049 649	61.45
74 populations (No Kānuka)	5%	2 580 451	96.43
68 populations (NZ only)	2%	3 270 864	76.08
68 populations (NZ only)	5%	2 526 589	98.49

(on average 41% of reads) were from fungal and bacterial associated microbiome and not used for further analysis in this study. The equivalent coverage obtained from reads mapped had a mean depth of 169X across the genome (Supplementary material 4). Base pair coverage for the complete 76 population dataset ranged from 150 419 718 bp (for kānuka population P011, equivalent to 60% of the reference genome assembly) to 247 091 534 bp (population P003, equivalent to 99.5% of the reference genome assembly) (Supplementary material 5). When this was filtered to include only nucleotide bases with a read depth greater than 100X, the range in base pair coverage reduced to 65 445 232 (for kānuka population P008, equivalent to 26.2% of the reference genome assembly) to 196 036 101 bp (for population P060, equivalent to 79% of the reference genome assembly). Kānuka populations (P008 and P011) had the lowest coverage, as expected for a relative in a different genus.

Variant detection

A total of 10.14 million candidate variants (SNPs and indels) were detected across the 76 DNA pools and throughout the genome prior to filtering. A stringent filtering protocol was applied to the raw set of variants to remove sequencing errors, rare variants and variants occurring in low depth regions, only keeping SNPs with an average read depth greater than 100X and no missing data, resulting in 5 503 881 SNPs being detected across the reference assembly of "Crimson Glory". When further filtering was applied using minor allele frequencies (MAF) of 5% and 2%, 1 498 405 and 2 513 694 SNPs were detected across all 76 collection sites (including both kānuka and Australian *L. scoparium* samples), respectively (Table 1). When both kānuka populations were excluded (i.e. only *L. scoparium* populations taken into account), a total of 2 580 451 and 4 049 649 SNPs were detected using MAF of 5% and 2%, respectively. When only the 68 New Zealand mānuka populations were included (i.e. Australian *L. scoparium* and kānuka populations excluded), a total of 2 526 589 and 3 270 864 SNPs were detected using MAF of 5% and 2%, respectively.

Genetic differentiation and population structure

Allele frequencies generated from 1 498 405 SNPs for 76 populations (including Australian *L. scoparium* and kānuka) revealed the dataset consists of seven clusters (Fig. 1A) as determined by the optimized K-means clustering algorithm find.clusters in Adegenet v2.1.1 [59]. The function was used to run 30 successive K-means analyses, with an optimal K of seven selected based on Bayesian Information Criterion (BIC) (Fig. 1A). A Discriminant Analysis of Principal Components (DAPC) [59] using this optimal K value supported the segregation of the 76 populations into seven genetically and geographically distinct clusters (Fig. 1 C,D). Linear dimensions (LD) one, two, three and four (LD1, LD2, LD3 and LD4) of the DAPC (K = 7) explained 69.55%, 13.55%, 7.54% and 4.56% of the data variation, respectively (Fig. 1 C, D). New Zealand populations from the north-eastern South Island (NESI), south-western South Island (SWSI), northern North Island (NNI), central and southern North Island (CSNI) and the East Cape of the North Island (ECNI) each formed separate clusters, suggesting strong geographic structuring of mānuka populations into five distinct gene pools across New Zealand (Fig. 1B). There was also evidence of strong partitioning between *L. scoparium* populations in New Zealand and those sampled from Australia (including Tasmania), with the six populations of *L. scoparium* from Australia forming their own cluster.

The exploration of DAPC posterior-membership probabilities at five different K values (K = 5 to K = 9) further demonstrated the strong geographic structuring of the New Zealand mānuka populations (Fig. 2). When K = 5 three New Zealand clusters were identified – the NNI cluster, a central, eastern and southern North Island cluster and one cluster containing all South Island populations. When K = 6 the central, eastern and southern North Island cluster split into two – into the CSNI cluster and the ECNI cluster. When K = 7 (the optimal K value), the South Island cluster divided into the NESI and SWSI clusters. The single Tasmanian population was identified as an additional cluster when K = 8, however at K = 9 this cluster was lost, with the NNI and SWSI clusters both dividing in two. The NNI populations that formed an additional cluster at K = 9 (P037, P040 and P041) are distributed at the northern tip of the North Island, and the four SWSI populations that formed an additional cluster at K = 9 are all distributed on the West Coast of the South Island, indicating additional geographical structuring. P050 is the only population to demonstrate mixture between clusters (NNI and CSNI), predominantly being clustered with CSNI at K = 5, K = 6, K = 7 and K = 8, and is entirely clustered with NNI at K = 9. This population occurs at the northern edge of the CSNI distribution, and is in close proximity to an NNI population (P066).

Pairwise *Fst* genetic distances were calculated in PoolFstat v1.0.0 [60] using mpileup and sync input files (Fig. 3A). Overall average *Fst* distances for all New Zealand populations was 0.128 (SD = 0.057) (Table 2).

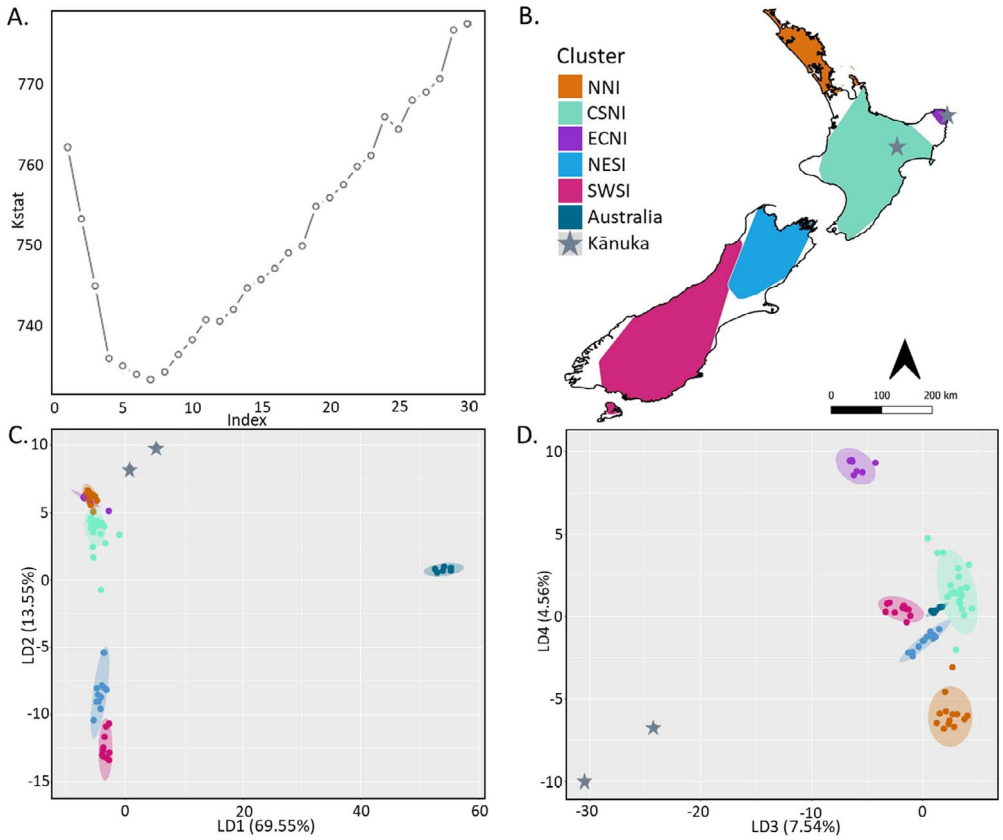

Figure 1. Population structure of *Leptospermum scoparium* in New Zealand and Australia based on whole genome pooled sequencing. A. Broken elbow plot of Bayesian Information Criterion (BIC) scores from Kmeans clustering analysis used to determine the optimal number of clusters (K) within the dataset. The optimal number of clusters is seven, including the kānuka (outgroup), Australia and five New Zealand clusters. C. Geographical distribution of DAPC clusters in New Zealand coloured by cluster. Colour codes and symbols are the same for B, C, and D. NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north-eastern South island, SWSI = south-western South Island. C. & D. Discriminant Analysis of Principal components (DAPC) (K = 7) of *L. scoparium* (dots) and *Kunzea robusta* (stars; kānuka). B. displays the first two dimensions (linear discriminants) of the DAPC analysis (LD1 and LD2), accounting for 69.55% and 13.55% of the variation, respectively. D. displays LD3 and LD4, accounting for 7.54% and 4.56% of the data, respectively.

Between the New Zealand gene pools, the lowest mean *Fst* values were observed between populations from CSNI and ECNI (average *Fst* = 0.079; SD = 0.023), whilst highest mean *Fst* values were between populations from ECNI and NESI populations (average *Fst* = 0.196, SD = 0.019). Within the New Zealand gene pools, average *Fst* values varied between 0.024 (SD = 0.016) (EC NI) and 0.073 (SD = 0.024) (NESI). Pairwise *Fst* distances between all populations can be found in supplementary material (Supplementary material 6). Average *Fst* distances between New Zealand and Australian populations, New Zealand and Tasmanian populations, and Australian and Tasmanian populations were 0.357 (SD = 0.02), 0.383 (SD = 0.011) and 0.308 (SD = 0.006) respectively. Similarly, an unrooted, neighbour-joining network generated in SplitsTree4 v4.14.8 [61] using the pairwise *Fst* distance matrix revealed strong structuring between the aforementioned gene pools; however, there was also an apparent conflicting signal observed within the dataset, visible in the boxes formed within the SplitsTree network (Fig. 3B). An Isolation by Distance (IBD) analysis revealed a significant relationship between genetic distances and geographic distances when Australian populations were included (p-value = 0.022) but no significant relationship within New Zealand populations when analysed alone (p-value = 0.383).

Genetic diversity and demographic modelling

Summary statistics generated by NPStats v1 [46] were calculated in non-overlapping 10 kb windows and results are presented as the weighted medians of these windows averaged per gene pool (Table 3). Weighted medians by population can be found in supplementary material (Supplemental material 7). The average weighted median for segregating sites (S) varied from 28 (SD = 11) (for kānuka) to 118 (SD = 92.8) (for Australian populations, including Tasmania) across all 76 populations. Within the New Zealand gene pools, S varied from 39.83 (SD = 4.02) (ECNI) to 46.03 (SD = 9.51) (CSNI). Watterson's θ varied from 0.0072 (SD = 0.0003) (SWSI) to 0.0123 (SD = 0.002) (Australia) across all populations, and from 0.0072 (SD = 0.0003) (SWSI) to 0.0087 (SD = 0.0014) (CSNI) within the New Zealand mānuka populations. Nucleotide diversity (π) ranged from 0.0066 (SD = 0.0002) (ECNI) to 0.0108 (SD = 0.0012) (Australia), and from 0.0066

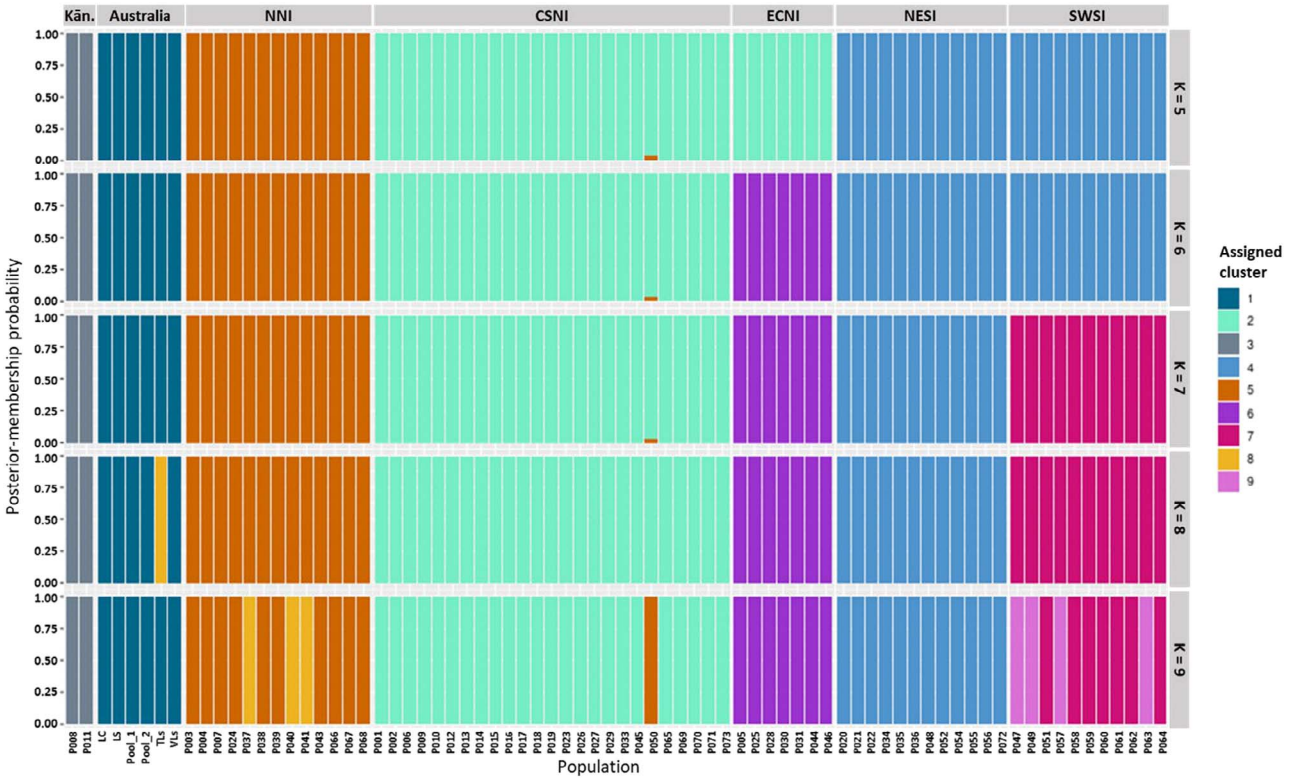

Figure 2. Posterior-membership probability. Stacked bar plots of posterior-membership probability for each population for five different K values. Kā. = Kānuka, NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north-eastern South island, SWSI = south-western South Island. Populations are colour coded according to clustering in Figure 1, with the addition of yellow for Assigned cluster eight and light pink for Assigned cluster nine.

Table 2. Average pairwise Fst for each gene pool with standard deviations in brackets. Matrix of pairwise Fst values for all populations can be found in Supplementary material 6. NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north eastern South Island, SWSI = south western South Island, NZ = New Zealand

	Australia	Tasmania	NNI	CSNI	ECNI	NESI	SWSI	NZ
Australia	0.187 (0.093)							
Tasmania	0.302 (0.006)	-						
NNI	0.363 (0.021)	0.388 (0.011)	0.049 (0.034)					
CSNI	0.356 (0.018)	0.382 (0.007)	0.143 (0.040)	0. 048 (0.024)				
ECNI	0.377 (0.017)	0.399 (0.006)	0.172 (0.036)	0.079 (0.023)	0.024 (0.016)			
NESI	0.357 (0.019)	0.384 (0.009)	0.184 (0.028)	0.165 (0.021)	0.196 (0.019)	0.073 (0.024)		
SWSI	0.346 (0.019)	0.373 (0.010)	0.178 (0.030)	0.146 (0.026)	0.175 (0.027)	0.106 (0.024)	0.069 (0.030)	
NZ	0.357 (0.020)	0.383 (0.011)	-	-	-	-	-	0.128 (0.057)

(SD = 0.0002) (ECNI) to 0.0074 (SD = 0.0009) (CSNI) within the New Zealand mānuka populations. Average Tajima's D weighted medians ranged from −0.387 (SD = 0.077) (SWSI) to −0.785 (SD = 0.0.136) (NNI) across populations, and) within the New Zealand mānuka populations, indicating an excess of low-frequency alleles.

An allele count dataset consisting of 74 populations (MAF of 0.05, and excluding kānuka) totalling 2 580 451 SNPs was used for the TreeMix [62] analysis. Based on the explained variance of eleven different migration events, it was determined that five to six migration events best explained the observed data used in the TreeMix model (Fig. 4A, B). Phylogenetic trees produced by this analysis support the clustering of New Zealand mānuka into five gene pools (NNI, CSNI, ECNI, and two South Island clades), with clear division also seen between New Zealand mānuka and Australian *L. scoparium* (Fig. 4C). The five migration event analysis suggests two major migration events have occurred between New Zealand and external gene pools: one between mainland Australia and Northland (NNI), and another from an ancestral population in Australia or Tasmania (not sampled in our study) into the NNI gene pool. Within New Zealand, there have been major migration events between the SWSI gene pools and NNI, and between the CSNI gene pool and NESI. An additional migration event is seen within the CSNI gene pool. When a sixth migration event was applied alongside the aforementioned events, the model determined an additional event from NNI to the SWSI gene pool (Fig. 4C). Migration events with the greatest weight are seen between the CSNI and NESI gene pools, and between NNI

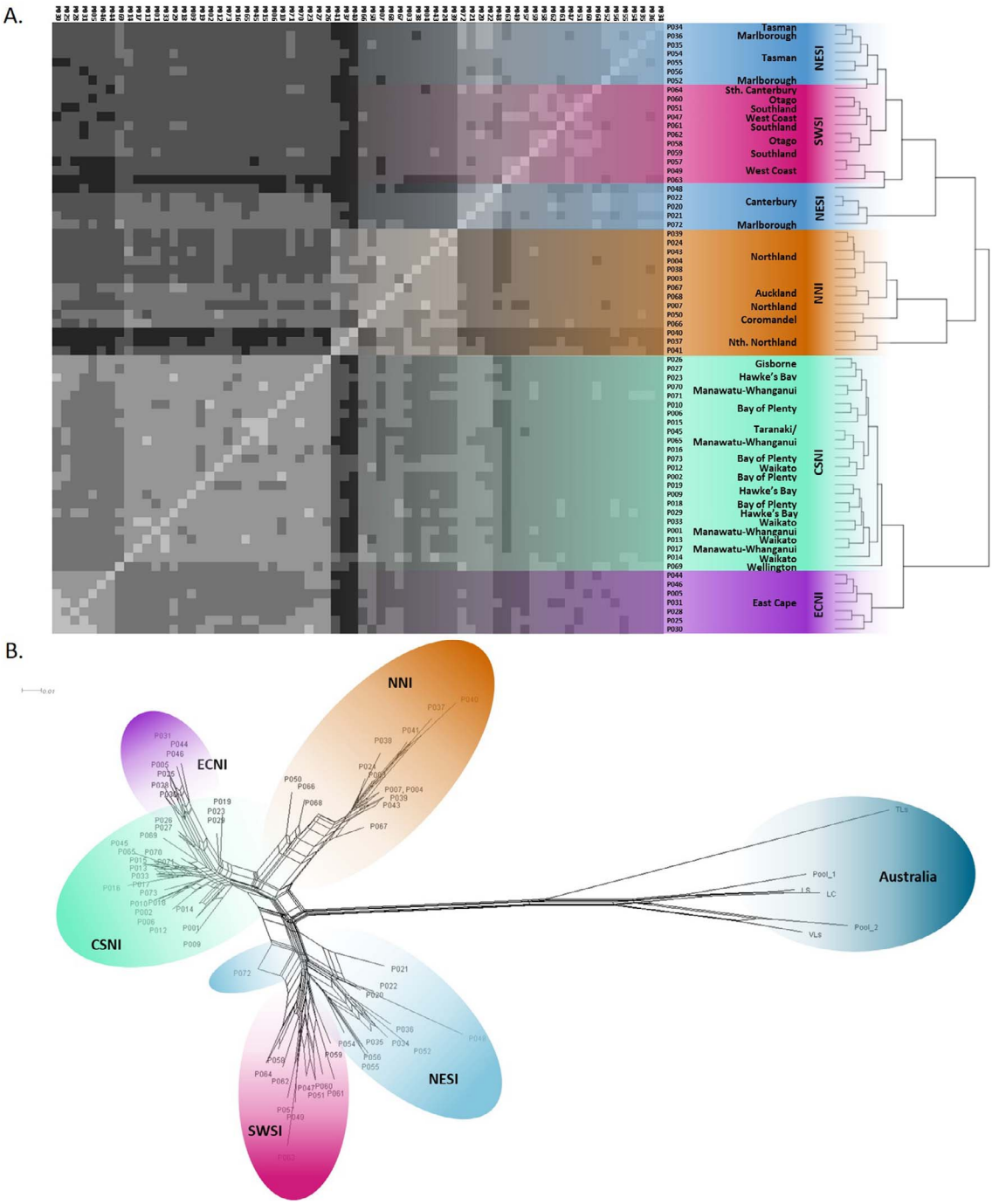

Figure 3. Pairwise Fst plot and neighbour-joining network. A. Pairwise Fst distance heat map of new Zealand populations with dendrogram based on hierarchical clustering analysis of Fst distances, darker greys indicating higher Fst distances, and lighter greys indicating lower Fst distances. B. Phylogenetic neighbour-joining network of *Leptospermum scoparium* in New Zealand and Australia indicating population differentiation using Fst. The tree was obtained using SplitsTree. NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north-eastern South island, SWSI = south-western South Island. Populations are colour coded according to Figure 1 (with the exception of P050 which is clustering with NNI rather than CSNI).

and SWSI gene pools. The results suggest no substantial migration events have occurred from New Zealand back to Australia.

SFS-based demographic models that explored four hypotheses (population stability, bottleneck, expansion and contraction models) suggest that NNI, CSNI and SWSI gene pools have all undergone expansion events, whilst the ECNI, NESI and Australian gene pools have all undergone contractions (Fig. 5A; Table 4). Parameter estimates and associated summary statistics inferred from 100 parametric bootstraps of the most suitable model for each gene pool, as well as output of the best overall model based on AIC can be found in Supplementary material 8. In the case of the three gene

Table 3. Summary statistics averaged by gene pool of population genetic parameters calculated for using NPStats. S = segregating sites, π = nucleotide diversity, SD = standard deviation, NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north eastern South Island, SWSI = south western South Island. Individual summary statistics for each populations can be found in Supplementary material 7

Region	S (SD)	Watterson's θ (SD)	π (SD)	Tajima D (SD)
Australia	118.83 (92.80)	0.0123 (0.0022)	0.0108 (0.0012)	−0.514 (0.363)
NNI	44.5 (9.33)	0.0084 (0.0011)	0.0069 (0.0005)	−0.785 (0.136)
CSNI	46.03 (9.51)	0.0087 (0.0014)	0.0074 (0.0009)	−0.691 (0.147)
ECNI	39.83 (4.02)	0.0077 (0.0007)	0.0066 (0.0002)	−0.632 (0.160)
NESI	43.16 (5.7)	0.0079 (0.00038)	0.0072 (0.0003)	−0.520 (0.158)
SWSI	43.36 (4.01)	0.0072 (0.0003)	0.0069 (0.0002)	−0.387 (0.077)
Kānuka	28.00 (11.00)	0.0099 (0.0044)	0.0081 (0.0035)	−0.723 (0.065)

pools that have undergone population size contractions, results for bottleneck and contraction models were very similar and in some instances the bottleneck model had better support. However, results of the bottleneck models all indicated that the effective population size of the gene pools does not recover following the bottleneck, and is less than the effective population size during the bottleneck – effectively indicating that the population sizes have continued to contract (hence the contraction model being selected as the better model). Here we report the results of the models with the lowest AIC from 100 runs of the best model for each gene pool. NNI is estimated to have begun expanding ∼243 602 generations ago, increasing effective population size (N_e) by ∼2739.8%; the CSNI gene pool is estimated to have begun expanding ∼23 721 generations ago, increasing N_e by ∼941.35%; and the SWSI gene pool is estimated to have begun expanding ∼362 230 generations ago, increasing N_e by ∼458.03%. Whilst the ECNI gene pool is estimated to have begun contracting ∼964 generations ago, with N_e decreasing by ∼97.02%, the NESI gene pool is estimated to have begun contracting ∼550 generations ago, decreasing N_e by ∼95.62%, and the Australian gene pool is estimated to have begun contracting ∼1155 generations ago, decreasing N_e by ∼98.98%.

The model that best explains the evolutionary history of Australian and New Zealand populations (represented by the NNI gene pool) and the arrival of mānuka to New Zealand was the D-CGF-CE model (Fig. 5B; Fig. 6; Table 5). The parameter estimates for this model suggest there has been continuous, asymmetrical gene flow between the two countries, with more gene flow occurring from Australia to New Zealand than vice versa (Australia to NNI = 1.03E-05, NNI to Australia = 2.67E-06). Results of this model also suggest that the split between New Zealand and Australian populations occurred ∼581 569 generations ago. The parameter estimates for the Australian gene pool are similar under this model to the parameter estimates determined in the above individual models, with contraction of the Australian gene pool estimated to have begun ∼1235 generations ago, with a decrease in N_e by ∼99.0%. However, parameter estimates for the Northland gene pool are comparably different from the individual model estimates, and this gene pool is estimated to have begun expanding ∼17 507 generations ago (cf. ∼243 602 generations ago), with N_e increasing only by ∼605.86% (c.f. the individual NNI expansion model of ∼2739.8%).

Discussion

A genomics study supporting indigenous branding

This first population genomic study of *L. scoparium* indicates clear geographic structuring of mānuka across New Zealand and a strong genetic differentiation between New Zealand and Australian populations. This research was co-developed in response to demand for this information by indigenous agribusinesses, with an interest in developing strong regional branding associated with their cultural connections to the land. Such evidence will assist Māori by addressing economic, cultural and ecological aspects. In terms of economic benefits for Māori, our research will underpin the veracity of novel mānuka-based products and assist Māori to achieve the potential of this natural resource, leveraging off global demand for provenance information – particularly for food. Determining the provenance of the plants will provide indigenous innovators producing honey with marketing opportunities unique to Māori. This branding will identify these products as being premium and authentic, and create a strong differentiation from other types of honeys, including from Australia. Finally, understanding mānuka genetic diversity will contribute to better management of this resource, improving the resilience of the honey industry.

Pool-seq

Pooled whole genome re-sequencing (pool-seq) was applied in this study as an alternative to individual genotyping. We chose this strategy as opposed to reduced representation methods such as genotyping by sequencing (GBS) as it gives access to the full set of variants across the genome. In our experience, GBS can yield variable read depth across genomic regions, with many errors and missing data [54]. Whilst pool-seq does have its limitations [63] (no true sample genotypes, inability to calculate linkage disequilibrium,

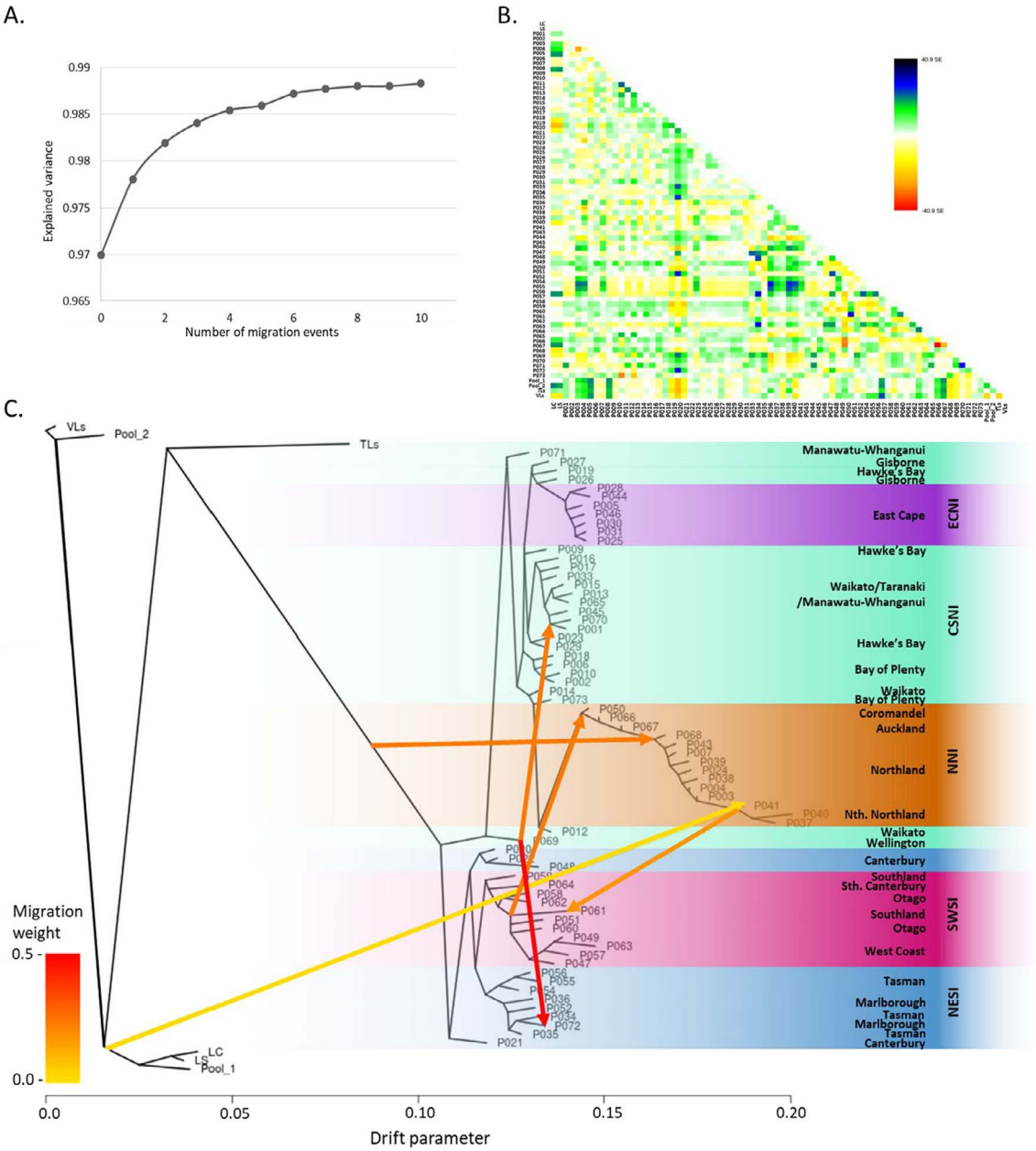

Figure 4. TreeMix analysis of past migration events among New Zealand and Australia *Leptospermum scoparium*. A. Estimation of the number of migration events that best explains the data. The explained variance reaches a plateau around 98.6% at five to six migration events. B. Heat map of residuals from TreeMix analysis with six migration events. C. Phylogenetic representation of TreeMix analysis with six migration events. Colour of arrows indicates the significance of each migration event (yellow = less significant, red = more significant). NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north-eastern South island, SWSI = south-western South Island. Populations are colour coded according to Figure 1 (with the exception of P050 which is clustering with NNI rather than CSNI).

limited downstream applications), we believe our dataset is superior to a GBS dataset as we obtained a large proportion of the reference genome with a consistent read depth. Additionally, pool-seq is an economical choice compared with barcoding the complete sampling set, as only a limited number of costly sequencing indices are required. This allowed us to sequence a larger set of individual trees across more regions (2325 trees across 76 sites). One hundred times coverage and > 25 individuals per pool were recommended to attain reliable results [43, 44], and we were able to achieve this for the majority of populations. This enabled us to estimate allelic frequencies within populations. Our results are consistent with morphological [20], phenological [20] and chemotypic [21, 22] studies that have suggested there is geographical structuring of mānuka across New Zealand, and builds on previous genetic work by Buys *et al.* (2019).

Population genetic structure

Genetics are commonly mentioned as a factor influencing the observed variation in the chemotypic compounds,

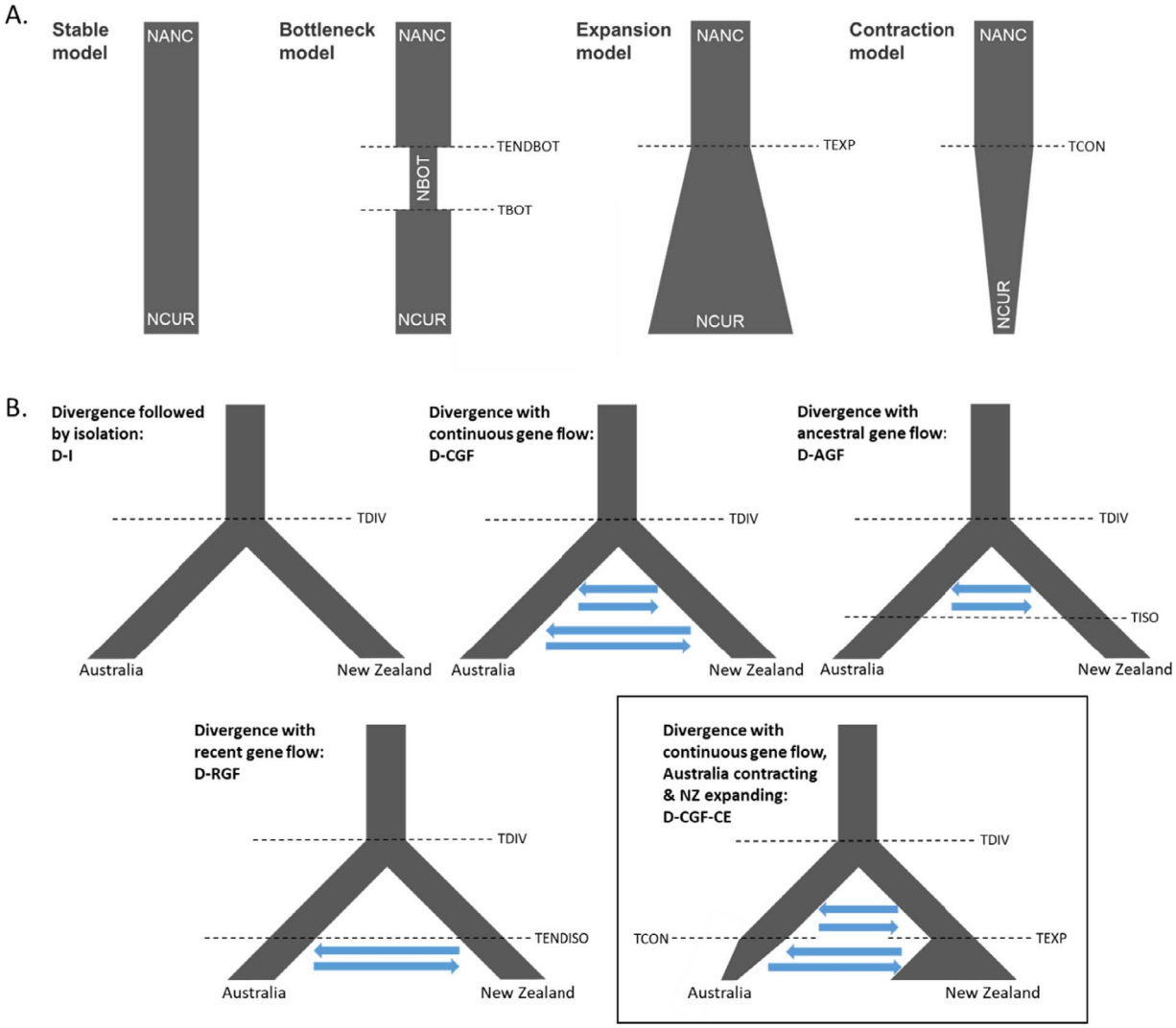

Figure 5. Diagram of the demographic models. A. models tested on individual gene pools and B. the demographic models used to investigate the divergence of Australian and New Zealand *Leptospermum scoparium* gene pools. NANC = Effective population size of ancestral population, NCUR = Effective population size of current population, NBOT = Effective population size of population during bottleneck, TBOT = time of bottleneck, TENDBOT = time of end of bottleneck, TEXP = time of expansion, TCON = time of contraction, TDIV = time of division, TISO = time of isolation, TENDISO = time of end of isolation. Dotted lines indicate when historic events occurred. Blue arrows indicate gene flow and direction of gene flow. Best model (D-CGF-CE) is highlighted inside box.

morphology and phenology across mānuka's range. However, no previous studies have investigated the population genetic structure of *L. scoparium* across New Zealand, or between New Zealand and Australia, or provided insight into the demographic history of this species. Our population genetic investigation indicates that there are five genetically and geographically distinct gene pools of *L. scoparium* in New Zealand, and that these populations are genetically distinct from Australian *L. scoparium* populations. DAPC, *F*st, SplitsTree and IBD analyses all support the genetic differentiation of these gene pools. However, the TreeMix analyses suggest that despite this genetic differentiation, gene flow has occurred between these gene pools – including between New Zealand and Australian populations. Genetic differentiation (*F*st) indicates that New Zealand populations have been separated from Australian populations for a long period of time; however again, the TreeMix analyses suggest that multiple migration events have occurred from Australia to New Zealand (most likely to NNI populations). There is no evidence to support significant migration occurring from New Zealand back to Australia.

Buys *et al.* (2019) described three genetically and geographically distinct clades of mānuka (northern North Island, central and southern North Island, South Island), with some weakly supported sub-structuring within these clades. With more sampling sites and a larger genetic dataset, we were able to elucidate two additional clades – dividing the Buys *et al.* (2019) South Island clade into two (NESI and SWSI), and establishing East Cape populations as a separate gene pool (ECNI). Buys *et al.* (2019) also discussed the potential reason for poor resolution within the branches of the three New Zealand mānuka clades they described, suggesting it

Table 4. Maximum likelihood and AIC statistics for the individual demographic models tested. The model considered to be the best fit is in bold. These models were then run 100x (See Supplementary material 8). DeltaL = the difference between the maximum estimated likelihood and the maximum observed likelihood. AIC = Akaike's Information Criterion. Note: in some instances the model with the lowest AIC was not always selected as the best model, as the parameter estimates were not sensible (see Results section). NNI = northern North Island, CSNI = central and southern North Island, ECNI = East Cape North Island, NESI = north eastern South Island, SWSI = south western South Island

Gene pool(s)	Model	No. of parameters	Maximum estimated likelihood	deltaL	AIC
NNI	Stable	2	−4 624 575	316 784	21 296 959
	Bottleneck	6	−4 498 474	68 461	20 716 252
	Expansion	**4**	**−4 498 275**	**190 484**	**20 715 329**
	Contraction	4	−4 620 187	312 395	21 276 754
CSNI	Stable	2	−5 222 475	492 068	24 050 390
	Bottleneck	6	−6 627 286	458 419	30 519 790
	Expansion	**4**	**−5 122 746**	**392 339**	**23 591 125**
	Contraction	4	−5 177 958	447 551	23 845 384
ECNI	Stable	2	−4 880 280	350 141	22 474 523
	Bottleneck	**6**	**−4 817 860**	**287 721**	**22 187 079**
	Expansion	4	−4 845 841	315 702	22 315 933
	Contraction	**4**	**−4 823 543**	**293 404**	**22 213 246**
NESI	Stable	2	−3 497 510	70 215	16 106 634
	Bottleneck	**6**	**−3 483 627**	**56 332**	**16 042 707**
	Expansion	4	−3 490 089	62 794	16 072 461
	Contraction	**4**	**−3 484 327**	**57 032**	**16 045 928**
SWSI	Stable	2	−4 602 161	172 147	21 193 737
	Bottleneck	6	−6 068 501	117 142	27 946 491
	Expansion	**4**	**−4 497 117**	**67 103**	**20 709 997**
	Contraction	4	−4 598 972	168 958	21 179 058
Australia	Stable	2	−5 501 975	427 386	25 337 533
	Bottleneck	**6**	**−5 305 720**	**231 132**	**24 433 757**
	Expansion	4	−5 503 242	428 654	25 343 375
	Contraction	**4**	**−5 308 952**	**234 363**	**24 448 635**

was the result of conflicting signal within their dataset. We can now confirm that the conflicts in signal they encountered are likely the result of gene flow occurring between populations and gene pools within New Zealand and between Australian and New Zealand populations.

Between the New Zealand populations, pairwise *Fst* was relatively low (cf. difference between New Zealand and Australian populations), particularly between neighbouring populations where gene flow is more readily facilitated. This is demonstrated by ECNI and CSNI gene pools having the lowest average pairwise *Fst*, and NESI and SWSI also exhibiting relatively low average pairwise *Fst*. Interestingly, the highest pairwise *Fst* were not between the most geographically distant gene pools, however, but between ECNI and NESI. The seeds and pollen of mānuka are dispersed via wind [64], and it is possible that wind may more easily carry seeds and pollen from the South Island to the northern North Island, than across the North Island to the East Cape.

Evolutionary history

The genus *Leptospermum* is thought to have evolved in Australia during the Miocene (5 – 23Mya), with the split between *L. scoparium* and its sister species *Leptospermum trinervium* estimated to have occurred 15.9 Mya (95% confidence interval = 6.7–26.8 Myr) in Australia; arrival of an ancestral *L. scoparium* lineage to New Zealand would

Table 5. Maximum likelihood and AIC statistics for the demographic models used to investigate the divergence of Australian and New Zealand *Leptospermum scoparium* gene pools. The model considered to be the best fit is in bold. These models were then run 100x (See Supplementary material 8). DeltaL = the difference between the maximum estimated likelihood and the maximum observed likelihood. AIC = Akaike's Information Criterion

Model description	Model ID	No. of parameters	Maximum estimated likelihood	deltaL	AIC
Divergence followed by isolation	D-I	5	−12 083 216	923 326	55 645 277
Divergence with continuous gene flow	D-CGF	7	−11 874 074	714 183	54 682 144
Divergence with ancestral gene flow	D-AGF-I	8	−11 878 674	718 783	54 703 330
Divergence with recent gene flow	D-RGF	8	−11 859 600	699 709	54 615 490
Divergence with continuous gene flow, Australia contracting& New Zealand expanding	**D-CGF-CE**	**11**	**−11 622 817**	**462 926**	**53 525 072**

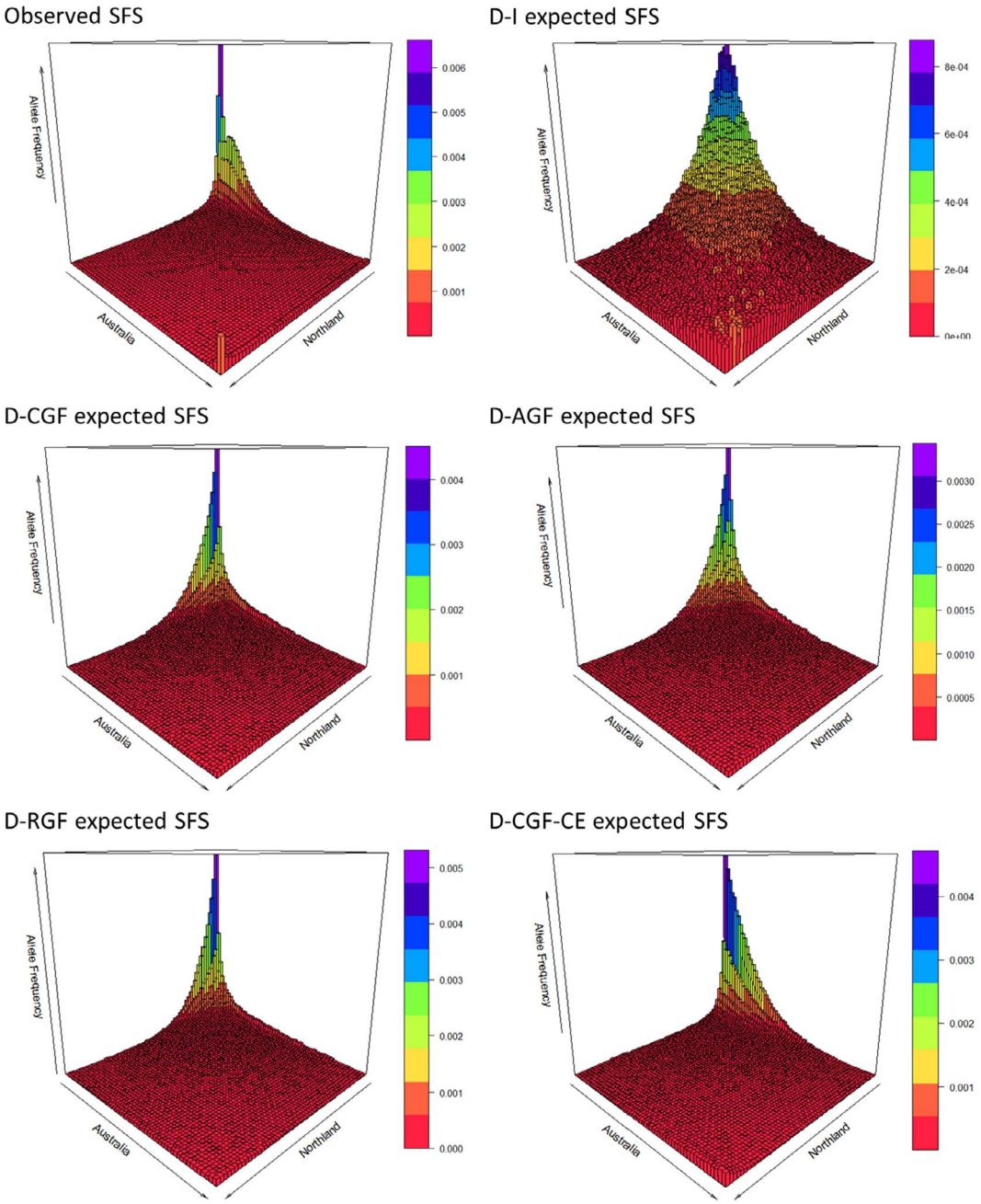

Figure 6. 3D heat maps displaying 2D SFS of Australian and New Zealand *Leptospermum scoparium* gene pools. Top left plot is the observed 2D SFS of the Australian and New Zealand (represented by NNI) gene pools used to run the demographic models, and against which the expected SFS output of all models was compared.

have occurred subsequently [13, 32]. Results of our demographic models are congruent with this. If we assume a generation time of 15–20 years for mānuka presuming that larger and more mature *L. scoparium* trees contribute more to reproduction, our results suggest that the New Zealand lineage of *L. scoparium* split from the Australian lineage ~9–12 Mya. This timing also fits with the arrival and divergence of other woody angiosperm species in New Zealand, following the end of the Miocene Thermal Optima ~15 Mya [65, 66]. However, mānuka can flower from three to five years after germination, which would correspond to ~2–5 Mya. Therefore, the population structure we have observed in New Zealand mānuka (into five gene pools) likely reflects phylogeographic patterns associated with climatic and glacial cycling of the Plio-Pleistocene (12 000 Ka – 5 Mya) [35, 36], and/or New Zealand's turbulent geological history of mountain building and volcanic activity [34, 67].

For two gene pools (NNI, SWSI), the individual demographic models indicate that N_e began increasing ~4–7 Mya, suggesting these gene pools may have benefitted from the cooler climate of the Plio-Pleistocene, expanding their ranges in the north of the North Island and in the south western South Island, respectively.

These two regions are considered to have been relatively geologically and climatically stable over the last ~5 Myr [68]. In particular, the southern South Island during this time would have consisted predominantly of tussock grasslands and shrub-lands, which may have provided suitable open habitat for mānuka to establish and then to expand its range [34, 69, 70]. If we take the estimated date of expansion of NNI from the Australian and New Zealand divergence model, it suggests that NNI did not begin expanding until more recently (~260 000–350 140 years ago). This timing is similar to that estimated for the CSNI gene pool, with expansion estimated to have begun ~350 000–475 000 years ago. Fires occurred regularly in acid mires during the Plio-Pleistocene in the central and northern North Island, whilst volcanic activity in the central North Island would have also caused fires and significant disturbance to landscapes due to ash fall [34, 71–73]. *L. scoparium* is a seral species, known to establish in disturbed landscapes, particularly after fire, an ancestral adaptation that evolved to enable it to survive forest fires in Australia [73]. Serotinous *L. scoparium* populations are more common in the North Island (correlating with the history of natural fires in New Zealand) and this may be one reason for the success and expansion of NNI and CSNI gene pools in these two regions over the last ~350 000 years [73].

For the two remaining New Zealand gene pools (ECNI, NESI), demographic modelling suggests N_e has been decreasing over the last ~8250–19 000 years. This timing coincides with the end of the Last Glacial Maximum (LGM) (when global ice sheets were at their maximum integrated volume) [74] and warming of the climate, suggesting that as glaciers, tussock grassland and alpine shrub-land receded, so too did the distribution of mānuka in the north of the South Island and in the East Cape in the North Island. For the northern South Island in particular, there would have been considerable habitat loss at this time as sea levels rose, and the land bridge connecting the North and South Islands was inundated [34, 69], potentially explaining the contraction of the NESI gene pool. Results of the demographic modelling also suggest that N_e of the Australian gene pool has decreased significantly (~98.98% decrease) over the last ~17 000–23 000 years, however, our sampling of this gene pool is very limited (considering the distribution of *L. scoparium* in Australia). Sampling across the complete distribution of *L. scoparium* in Australia is required to draw any further conclusions about the evolutionary history of this gene pool.

It has been hypothesised, based on morphological data, that *L. scoparium* originated in Tasmania and simultaneously dispersed to Australia and New Zealand [13]. Although we only have one population from Tasmania, our results provide no evidence to support the hypothesis that *L. scoparium* originated in Tasmania and dispersed to Australia and New Zealand simultaneously. Based on our findings, it is likely that *L. scoparium* originated on mainland Australia, and subsequently dispersed to Tasmania and then to New Zealand, with migration events also occurring from mainland Australia to New Zealand. Wider sampling within Australia and Tasmania would help to confirm this hypothesis, however similar conclusions were reached by Buys *et al*. (2019) and Stephens (2006) in their phylogenetic and chemotype studies, respectively. Similarly to those of Stephens (2006), our findings suggest that there have been multiple introductions of *L. scoparium* from Australian populations to New Zealand. Eastward, long-distance dispersal of flora [75–79] and fauna [80–82] from Australia to New Zealand is not uncommon, and is often attributed to the Tasman Current and the Antarctic Circumpolar Current phenomenon of the Southern Hemisphere (formed following the separation of Antarctica from South America and then Australia in the mid Oligocene ~28-35Mya) [77, 83, 84].

Conclusion

The findings from this study will support the commercialisation of mānuka honey based on its indigenous and geographic origins. Additionally, data produced in this study can contribute to the conservation management of the mānuka germplasm, ensuring genetic diversity is maintained in an industry where selective planting is occurring. Furthermore, the genetic structure of New Zealand mānuka as five different clusters, all distinct from one another and from Australian populations, could form the basis for a future reclassification and taxonomic definition of this species. Finally, knowledge of genome variants across its natural range can be used in future research to understand key traits such as tolerance to the tree pathogen myrtle rust (*Austropuccinia psidii*) (Smith *et al*. 2020) and adaptation to climatic and environmental conditions.

Materials and methods

Indigenous considerations and data sovereignty

Māori, the indigenous people of New Zealand, currently have a claim under the Waitangi Tribunal (the mechanism for redress via Te Tiriti o Waitangi (The Treaty of Waitangi) – the founding document of New Zealand), relating to intellectual property pertaining to native flora and fauna [56]. This claim has not been settled to date, but current convention in New Zealand research is to recognise Māori connection and guardianship over native flora and fauna, acknowledging that this claim is still before the tribunal. Therefore, before samples for genetic analysis were collected from natural stands of mānuka grown on Māori freehold land, mana whenua (Māori who have historical and territorial rights over the land) were individually approached and were briefed about the objectives of the research study. The privilege to access taonga material was granted for the purpose of this research and further research on these samples would require consent, including access to more samples, and

further analysis or disclosure of the exact identity of the trees. For this reason, our sampling sites are named according to their region of origin and the exact location of the samples and identity of the people who contributed has purposefully been occulted to address the sensitivity of this indigenous intellectual property. For further comments on the involvement and interaction of Māori and science see Morgan *et al.* [57] and Hudson, *et al.* [58].

Sampling

In total, samples from 2325 individual mānuka (n = 2265) and kānuka (*Kunzea robusta*) (n = 60) trees were collected from 70 sites (mānuka = 68 sites, kānuka = two sites) around New Zealand and six sites in Australia (Supplementary material. 1; 2). All trees were collected from naturally established stands, and not from commercially grown plantations. For each tree, fifteen young expanding leaves were sampled in duplicate, with at least 50 m between trees to ensure they were unlikely siblings, and metadata (e.g. latitude, longitude) were recorded. Samples were kept on silica beads in 2 mL screw cap tubes before being stored at −80°C. For all New Zealand samples DNA was extracted using a modified CTAB protocol [85] and quantified using the Quant-iT™ PicoGreen™ dsDNA Assay kit (Invitrogen, Burlington, ON, Canada) and a SpectraMax® Gemini EM Microplate Reader (Molecular Devices, San Jose, California, USA). Australian samples were extracted using the Qiagen DNeasy Plant Mini Kit on the QIAGEN QIAcube, the ISOLATE II Plant DNA Kit (Bioline) using buffer PA1 to manufacturer's protocol and extending the lysis incubation to 40 minutes, or by using the Machery-Nagel Nucleospin Plant II Kit with the PL2/PL3 buffer system. Results were examined using SoftMax® Pro Software (Molecular Devices, San Jose, California, USA). DNA of each sample was normalised to the same concentration and then pooled for each collection site to obtain two micrograms of total DNA per site (Supplementary material 2).

Pooled sequencing and variant calling

All 76 DNA pools were indexed using Illumina (Illumina, Inc., San Diego, USA) commercial indices, following the manufacturer's protocol. DNA pools were sequenced using the Illumina NovaSeq 6000 System utilising S2 flow cell technology and seven lanes of 2x150 base pair (bp) output at the Australian Genome Research Facility (AGRF, Melbourne, Australia). The libraries were approximately 600 bp. Raw fastq sequencing files from the 76 pools were separated by indices and quality control was carried out using md5 checksum v3.3 [86] and FastQC v0.11.8 [87]. Read mapping was performed against the reference genome of mānuka "Crimson Glory" [55] using Bowtie2 v3.4.3 [88], applying the —end-to-end parameter for each site independently. SAMtools v1.7 [89] was utilised to convert output SAM files into BAM files, which were sorted and indexed against "Crimson Glory". Coverage statistics for each BAM file were acquired by applying the SAMtools flagstat utility. BCFtools v1.9 [90] was used to detect variants for each pseudo-chromosome of the reference genome, implementing the –m and –v options of BCFtools. Filtering of raw variants into high confidence SNPs was carried out using VCFtools v0.1.14 [91], applying the following parameter settings: a minor allelic frequency (MAF) of 5%, 2% and 0% (—maf 0.05, —maf 0.02 and –maf 0.00); no indels (—remove-indels); a minimum mean read depth of 100X for each site (—min-meanDP 100); and no missing data (—max-missing 1) (Table 1).

Population genetics and demographic modelling

Allele frequencies were calculated following SNP calling in the R v3.5.0 [92] package vcfR v1.8.0 [93]. Population structure was examined via k-means clustering (find.clusters) and a Discriminant Analysis of Principal Components (DAPC) functions from the R package Adegenet v2.1.1 [94]. The find.clusters analysis was performed using max.n = 30 and n.pca = 100, whilst an initial DAPC analysis was performed using n.pca = 100, and n.da = 100. An optimum number of Principal Components (PC) was discerned from this initial run using the optim.a.score function (also from the Adegenet package), before a final DAPC analysis was run using both optimised K and PC values. Population pairwise Fixation Indices (Fst) were calculated utilising PoolFstat v1.0.0 [60], using mpileup and sync input files. Neighbour-net networks and IBD analyses were implemented in SplitsTree4 v4.14.8 [61] and ade4 v1.7–15 [95], respectively, both using the aforementioned Fst genetic distance matrix. Diversity statistics and neutrality tests (segregating sites (S), nucleotide diversity (π), Watterson's θ and Tajima's D) were carried out using NPStats v1 [46], applying -l 10 000 -mincov 25 -maxcov 500. An allele count dataset consisting of 74 populations (filtered by MAF of 0.05, and excluding kānuka) was utilised for the TreeMix v1.13 analyses [62], where eleven migration events (0–10) were explored in 1000 SNP blocks, and the trees were rooted by the Australian population VLs. Covariance data were extracted from the TreeMix output files and an equation provided by Pickrell and Pritchard [62] (fraction *f*, page 5) was used to calculate the explained variance of each migration run.

To explore the demographic history within and between New Zealand and Australian gene pools (established in our DAPC, Fst and TreeMix analyses), we applied a coalescent simulation-based method in fastsimcoal v2.6.0.3 [96]. Rare variants were called in VCFtools using the settings: MAF 0.00, missing = 0, depth = 100, resulting in a dataset of 5 503 881 SNPs. The minor allele frequencies of these SNPs were averaged across the populations within each gene pool, to fully encapsulate the genetic diversity and structure of each. Unfolded Site Frequency Spectra (SFS) and two-dimensional SFS (2D-SFS) (using P011 (kānuka) as an outgroup) were calculated for each gene pool using SweepFinder2 [97] and a custom R script (Supplementary material 9), respectively. For every

demographic model tested, 100 000 coalescent simulations were applied in fastsimcoal2, with maximum likelihood estimates calculated based on differences between the input observed SFS and the output expected SFS. Models were repeated 100 times, and a global maximum likelihood estimate was obtained from these independent runs and Akaike's Information Criterion (AIC) calculated for model comparison and selection.

Individual demographic models were implemented for each gene pool in fastsimcoal2, with neutral, bottleneck, expansion and contraction models explored. Multi-gene pool models were then carried out. Additionally, a 2D unfolded SFS matrix between the Australian gene pool and a representative New Zealand gene pool (Northland) was calculated to explore the demographic history of divergence and gene flow between New Zealand and Australian *L. scoparium*. Five models were tested: divergence followed by isolation (D-I); divergence followed by continuous gene flow (D-CGF); divergence with ancestral gene flow followed by isolation (D-AGF-I); divergence with only recent gene flow (D-RGF); and divergence with continuous gene flow as well as contraction for Australia and expansion for New Zealand (D-CGF-CE). The structure of the final model was based on findings of the initial individual demographic models.

Acknowledgements

We thank the Māori landowners who gave us the privilege to access plant materials for this study; Treigh Akuhata-Christy (Ngāti Porou), Trey Thomson (Ngāti Porou), Morgan Coleman, Paul Peterson, and Hamish Maule, Chris Morse, Rowan Buxton, Alex Fergus and Jessie Prebble from Manaaki Whenua Landcare Research for their help with sample collection. We thank the Honey Landscape Māori project governance group, Aaron Taikato (Prime Holding Ltd), Shar Amner (Atihau-Whanganui), Huti Watson (Ngāti Porou Miere Ltd), Brenda Tahi (Tūhoe Tuawhenua Trust), Victoria Henstock (Ngāi Tahu) and Twyla McDonald (Tai Tokerau Honey Ltd) for their advice; Emily Buck, Cecilia Deng, John McCallum, Susan Thomson, Samantha Baldwin, Linley Jesson, Chris Kirk, Andrew Granger, Claire Hall, Jenny Green, Nicolas Bordes and Isabel Moller from The New Zealand Institute for Plant & Food Research Ltd. for their advice and assistance during the project design, sampling, data analysis and Māori engagement; Matthew Tinning and the Australian Genome Research Facility (AGRF) for the next-generation sequencing; Dr Phillip Wilcox (University of Otago), Holden Hohaia (Manaaki Whenua Landcare Research) and Maui Hudson (University of Waikato) for discussions about Māori engagement and data sovereignty.

Author Contributions

Emily Koot: performed research, analysed data, wrote the paper. Elise Arnst: collected samples, wrote the paper. Melissa Taane: extracted DNA. Kelsey Goldsmith: collected samples, extracted DNA. Amali Thrimawithana: analysed data. Kiri Reihana: collected samples. Santiago C. González-Martínez: analysed data, wrote the paper. Victor Goldsmith: coordination of the access to plant material with Maori landowners. Gary Houliston: designed research, wrote the paper. David Chagné: designed research, analysed data, wrote the paper.

Data Availability

Permission from representatives of the Indigenous Peoples (Māori) was obtained for using the plant material used for this study. Further studies using this material, raw sequencing data and final genome assembly will require consent from the Māori iwi (tribe) and landowners who exercises guardianship for this material according to Aotearoa New Zealand's Treaty of Waitangi and the international Nagoya protocol on the rights of indigenous peoples. Access to raw and analyzed data will require permission from representatives of the iwi.

Conflicts of interests

The authors declare no competing interests.

Supplementary data

Supplementary data is available at *Horticulture Research Journal* online.

References

1. Gudgeon WE. Maori religion. *J Polyn Soc*. 1905;**14**:107–30.
2. Grey G. *Maori Lore: The Traditions of the Maori People, with the more Important of their Legends*. By authority, J. Mackay. South Africa: Government Printer, 1904.
3. Riley M. Viking Sevenseas NZ, 1994.
4. Best E. Maori medical lore. Notes on sickness and disease among the Maori people of New Zealand, and their treatment of the sick; together with some account of various beliefs, superstitions and rites pertaining to sickness, and the treatment thereof, as collected from the Tuhoe tribe. Part I. *J Polyn Soc*. 1904;**13**:213–37.
5. Crowe A. *A field guide to the native edible plants of New Zealand, 3rd edn*. Auckland: Godwit Publishing Limited, 2001.
6. Colenso W. On the geographical and economic botany of the North Island of New Zealand. *Trans N Z Inst*. 1868;**1**:1–58.
7. Beattie H. *Nature-Lore of the Southern Maori*. New Zealand: New Zealand Institute; 1920.
8. Best E. *Fishing Methods and Devices of the Maori*. New Zealand: Ams Pr Inc; 1929.
9. Hiroa TR. *The Coming of the Maori. Maori Purposes Fund Board*. Whitcombe & Tombs: Christchurch, New Zealand; 1949.
10. Palmer J. The Maori Kotaha. *J Polyn Soc*. 1957;**66**:175–91.
11. Stephens J, Molan PC, Clarkson BD. A review of Leptospermum scoparium (Myrtaceae) in New Zealand. *N Z J Bot*. 2005;**43**: 431–49.
12. Ministry for Primary Industries. *Apiculture: Ministry for Primary Industries 2018 Apiculture Monitoring Programme*. New Zealand: Wellington; 2018.

13. Thompson J. Revision of the genus leptospermum (Myrtaceae). *Telopea*. 1989;**3**:301–449.
14. Bean A. The genus leptospermum Forst. Et Forst. F.(Myrtaceae) in northern Australia and Malesia. *Austrobaileya*. 1992;**3**:643–59.
15. Bean A. Three new species of leptospermum (Myrtaceae) from Queensland and northern New South Wales. *Telopea*. 2004;**10**: 831–8.
16. Lyne A. Leptospermum namadgiensis (Myrtaceae), a new species from the Australian Capital Territory–New South Wales border area. *Telopea*. 1993;**5**:319–24.
17. Lyne A, Crisp M. Leptospermum jingera (Myrtaceae–Leptospermoideae): a new species from north-eastern Victoria. *Aust Syst Bot*. 1996;**9**:301–6.
18. Cheeseman TF. VI. The flora of Rarotonga, the chief island of the cook group. In: *Transactions of the Linnean Society of London. 2nd series. Botany*. Vol. **6**. 1925,261–313.
19. Allan HH. Zealand. *Flora of New Zealand. Volume I*. New Zealand: Manaaki Whenua Press, 1961.
20. Ronghua Y, Mark A, Wilson J. Aspects of the ecology of the indigenous shrub Leptospermum scoparium (Myrtaceae) in New Zealand. *N Z J Bot*. 1984;**22**:483–507.
21. Douglas MH, van Klink JW, Smallfield BM. *et al.* Essential oils from New Zealand manuka: triketone and other chemotypes of Leptospermum scoparium. *Phytochemistry*. 2004;**65**:1255–64.
22. Perry NB, Brennan NJ, van Klink JW. *et al.* Essential oils from New Zealand manuka and kanuka: chemotaxonomy of leptospermum. *Phytochemistry*. 1997;**44**:1485–94.
23. Porter NG, Wilkins AL. Chemical, physical and antimicrobial properties of essential oils of Leptospermum scoparium and Kunzea ericoides. *Phytochemistry*. 1999;**50**:407–15.
24. Adams CJ, Boult CH, Deadman BJ. *et al.* Isolation by HPLC and characterisation of the bioactive fraction of New Zealand manuka (Leptospermum scoparium) honey. *Carbohydr Res*. 2008;**343**:651–9.
25. Adams CJ, Manley-Harris M, Molan PC. The origin of methylglyoxal in New Zealand manuka (Leptospermum scoparium) honey. *Carbohydr Res*. 2009;**344**:1050–3.
26. Mavric E, Wittmann S, Barth G. *et al.* Identification and quantification of methylglyoxal as the dominant antibacterial constituent of Manuka (Leptospermum scoparium) honeys from New Zealand. *Mol Nutr Food Res*. 2008;**52**:483–9.
27. Atrott J, Haberlau S, Henle T. Studies on the formation of methylglyoxal from dihydroxyacetone in Manuka (Leptospermum scoparium) honey. *Carbohydr Res*. 2012;**361**:7–11.
28. Williams S, King J, Revell M. *et al.* Regional, annual, and individual variations in the dihydroxyacetone content of the nectar of manuka (Leptospermum scoparium) in New Zealand. *J Agric Food Chem*. 2014;**62**:10332–40.
29. Stephens JMC. *The Factors Responsible for the Varying Levels of UMF® in mānuka (Leptospermum scoparium) Honey*. New Zealand: The University of Waikato; 2006.
30. Buys MH, Winkworth RC, de Lange PJ. *et al.* The phylogenomics of diversification on an island: applying anchored hybrid enrichment to New Zealand Leptospermum scoparium (Myrtaceae). *Bot J Linn Soc*. 2019;**191**:1–17.
31. Lamb S, Mortimer N, Smith E. *et al.* Focusing of relative plate motion at a continental transform fault: Cenozoic dextral displacement >700 km on New Zealand's Alpine Fault, reversing >225 km of Late cretaceous sinistral motion. *Geochem Geophys Geosyst*. 2016;**17**:1197–213.
32. Thornhill AH, Ho SY, Külheim C. *et al.* Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. *Mol Phylogenet Evol*. 2015;**93**:29–43.
33. Kamp P, Late J. Cretaceous-Cenozoic tectonic development of the Southwest Pacific region. *Tectonophysics*. 1986;**121**:225–51.
34. Trewick S, Bland K. Fire and slice: palaeogeography for biogeography at New Zealand's North Island/south island juncture. *J R Soc N Z*. 2012;**42**:153–83.
35. McCulloch GA, Wallis GP, Waters JM. Onset of glaciation drove simultaneous vicariant isolation of alpine insects in New Zealand. *Evolution*. 2010;**64**:2033–43.
36. Patterson MO, McKay R, Naish T. *et al.* Orbital forcing of the East Antarctic ice sheet during the Pliocene and Early Pleistocene. *Nat Geosci*. 2014;**7**:841–7.
37. Mimee B, Duceppe M-O, Véronneau P-Y. *et al.* A new method for studying population genetics of cyst nematodes based on Pool-Seq and genomewide allele frequency analysis. *Mol Ecol Resour*. 2015;**15**:1356–65.
38. Fischer MC, Rellstab C, Tedder A. *et al.* Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Mol Ecol*. 2013;**22**:5594–607.
39. Schlötterer C, Tobler R, Kofler R. *et al.* Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nat Rev Genet*. 2014;**15**:749.
40. Dennenmoser S, Vamosi SM, Nolte AW. *et al.* Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (Cottus asper) revealed by Pool-Seq. *Mol Ecol*. 2017;**26**:25–42.
41. Roda F, Ambrose L, Walter GM. *et al.* Genomic evidence for the parallel evolution of coastal forms in the Senecio lautus complex. *Mol Ecol*. 2013;**22**:2941–52.
42. Futschik A, Schlötterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*. 2010;**186**:207–18.
43. Fracassetti M, Griffin PC, Willi Y. Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata*. *PLoS One*. 2015;**10**:e0140462.
44. Lynch M, Bost D, Wilson S. *et al.* Population-genetic inference from pooled-sequencing data. *Genome Biol Evol*. 2014;**6**: 1210–8.
45. Corander J, Majander KK, Cheng L. *et al.* High degree of cryptic population differentiation in the B altic S Ea herring C lupea harengus. *Mol Ecol*. 2013;**22**:2931–40.
46. Ferretti L, Ramos-Onsins SE, Pérez-Enciso M. Population genomics from pool sequencing. *Mol Ecol*. 2013;**22**:5561–76.
47. Endler L, Betancourt AJ, Nolte V. *et al.* Reconciling differences in pool-GWAS between populations: a case study of female abdominal pigmentation in Drosophila melanogaster. *Genetics*. 2016;**202**:843–55.
48. Bertelsen HP, Gregersen VR, Poulsen N. *et al.* Detection of genetic variation affecting milk coagulation properties in Danish Holstein dairy cattle by analyses of pooled whole-genome sequences from phenotypically extreme samples (pool-seq). *J Anim Sci*. 2016;**94**:1365–76.
49. Turner TL, Bourne EC, Von Wettberg EJ. *et al.* Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat Genet*. 2010;**42**:260.
50. Christe C, Stölting KN, Paris M. *et al.* Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol Ecol*. 2017;**26**:59–76.
51. Leblois R, Gautier M, Rohfritsch A. *et al.* Deciphering the demographic history of allochronic differentiation in the pine processionary moth Thaumetopoea pityocampa. *Mol Ecol*. 2018;**27**: 264–78.

52. Rellstab C, Zoller S, Tedder. A *et al.* Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One*. 2013;**8**:e80422.
53. Anderson EC, Skaug HJ, Barshis DJ. Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Mol Ecol*. 2014;**23**:502–12.
54. Bhatia D, Wing R, Singh K. Genotyping by sequencing, its implications and benefits. *Crop Improv*. 2013;**40**:101–11.
55. Thrimawithana AH, Jones D, Hilario E. *et al.* A whole genome assembly of *Leptospermum scoparium* (Myrtaceae) for mānuka research. *N Z J Crop Hortic Sci*. 2019;**47**:1–28.
56. Tribunal W. *Ko Aotearoa tēnei: A Report into Claims Concerning New Zealand Law and Policy Affecting Māori Culture and Identity: Te Taumata Tuatahi (Wai 262)*. Wellington, New Zealand: Waitangi Tribunal Report; 2011.
57. Morgan ER, Perry NB, Chagné D. Science at the intersection of cultures–Māori, Pākehā and mānuka. *N Z J Crop Hortic Sci*. 2019;**47**:225–32.
58. Hudson M, Garrison NA, Sterling R. *et al.* Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat Rev Genet*. 2020;**21**:377–84.
59. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 2010;**11**:94.
60. Hivert V, Leblois R, Petit EJ. *et al.* Measuring genetic differentiation from Pool-seq data. *Genetics*. 2018;**210**:315–30.
61. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2005;**23**:254–67.
62. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 2012;**8**:1–17, e1002967.
63. Cutler DJ, Jensen JD. To pool, or not to pool? *Genetics*. 2010;**186**:41–3.
64. Esler AE, Astridge SJ. Tea tree (leptospermum) communities of the Waitakere range, Auckland New Zealand. *N Z J Bot*. 1974;**12**:485–501.
65. Heenan PB, McGlone MS. 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*. 2019;**57**:204–26.
66. Prebble JG, Reichgelt T, Mildenhall DC. *et al.* Terrestrial climate evolution in the Southwest Pacific over the past 30 million years. *Earth Planet Sci Lett*. 2017;**459**:136–44.
67. Kamp PJ, Green PF, White SH. Fission track analysis reveals character of collisional tectonics in New Zealand. *Tectonics*. 1989;**8**:169–95.
68. McGlone M. Plant biogeography and the late Cenozoic history of New Zealand. *N Z J Bot*. 1985;**23**:723–49.
69. Alloway BV, Lowe DJ, Barrell DJA. *et al.* Towards a climate event stratigraphy for New Zealand over the past 30 000 years (NZ-INTIMATE project). *J Quat Sci*. 2007;**22**:9–35.
70. McGlone MS, Newnham RM, Moar NT. The vegetation cover of New Zealand during the last glacial maximum: do pollen records under-represent woody vegetation. *Terra Australis*. 2010;**32**:49–68.
71. Kitzberger T, Perry GLW, Paritsis J. *et al.* Fire–vegetation feedbacks and alternative states: common mechanisms of temperate forest vulnerability to fire in southern South America and New Zealand. *N Z J Bot*. 2016;**54**:247–72.
72. Wyse SV, Wilmshurst JM, Burns BR. *et al.* New Zealand forest dynamics: a review of past and present vegetation responses to disturbance, and development of conceptual forest models. *N Z J Ecol*. 2018;**42**:87–106.
73. Battersby PF, Wilmshurst JM, Curran TJ. *et al.* Exploring fire adaptation in a land with little fire: serotiny in Leptospermum scoparium (Myrtaceae). *J Biogeogr*. 2017;**44**:1306–18.
74. Clark PU, Dyke AS, Shakun JD. *et al.* The last glacial maximum. *Science*. 2009;**325**:710–4.
75. McGlone M, Duncan R, Heenan P. Endemism, species selection and the origin and distribution of the vascular plant flora of New Zealand. *J Biogeogr*. 2001;**28**:199–216.
76. Lockhart PJ, McLenachan PA, Havell D. *et al.* Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. *Ann Missouri Bot Gard*. 2001;**88**:458–77.
77. Winkworth RC, Wagstaff SJ, Glenny D *et al.* Plant dispersal news from New Zealand. *Trends Ecol Evol*. 2002;**17**:514–20.
78. Knapp M, Stöckler K, Havell D. *et al.* Relaxed molecular clock provides evidence for long-distance dispersal of Nothofagus (southern beech). *PLoS Biol*. 2005;**3**:e14.
79. Breitwieser I, Glenny DS, Thorne A. *et al.* Phylogenetic relationships in Australasian Gnaphalieae (Compositae) inferred from ITS sequences. *N Z J Bot*. 1999;**37**:399–412.
80. Goldberg J, Morgan-Richards M, Trewick SA. Intercontinental island hopping: colonization and speciation of the grasshopper genus Phaulacridium (Orthoptera: Acrididae) in Australasia. *Zool Anz*. 2015;**255**:71–9.
81. Goldberg J, Trewick SA, Paterson AM. Evolution of New Zealand's terrestrial fauna: a review of molecular evidence. *Philos Trans R Soc B: Biol Sci*. 2008;**363**:3319–34.
82. Koot EM, Morgan-Richards M, Trewick SA. An alpine grasshopper radiation older than the mountains, on Kā Tiritiri o te Moana (Southern Alps) of Aotearoa (New Zealand). *Mol Phylogenet Evol*. 2020;**147**:1–19, 106783.
83. Sanmartín I, Wanntorp L, Winkworth RC. West wind drift revisited: testing for directional dispersal in the southern hemisphere using event-based tree fitting. *J Biogeogr*. 2007;**34**:398–416.
84. Sanmartíin I, Ronquist F. Southern hemisphere biogeography inferred by event-based models: plant versus animal patterns. *Sys Biol*. 2004;**53**:216–43.
85. Gardiner S, Bassett HC, Noiton DA. *et al.* A detailed linkage map around an apple scab resistance gene demonstrates that two disease resistance classes both carry the V f gene. *Theor Appl Genet*. 1996;**93**:485–93.
86. Tung P-Y, Blischak JD, Hsiao CJ. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci Rep*. 2017;**7**:39921.
87. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. 2016. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
88. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;**9**:357.
89. Li H, Handsaker B, Wysoker A. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**:2078–9.
90. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;**27**:2987–93.
91. Danecek P, Auton A, Abecasis G *et al.* The variant call format and VCFtools. *Bioinformatics*. 2011;**27**:2156–8.
92. R Core Team. R: A language and environment for statistical computing. 2013.
93. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour*. 2017;**17**:44–53.
94. Jombart T. Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;**24**:1403–5.

95. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw*. 2007;**22**:1–20.
96. Excoffier L, Dupanloup I, Huerta-Sánchez E *et al*. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013;**9**:1–17, e1003905.
97. DeGiorgio M, Huber CD, Hubisz MJ *et al*. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*. 2016;**32**: 1895–7.