Ngā Uri o Karaka:

A genetic study of the karaka/kōpi tree in Aotearoa/New Zealand

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Genetics

at Massey University/Te Kunenga ki Pūrehuroa,
Palmerston North/Te Papaioea, New Zealand/Aotearoa.

Robin Amber Atherton

July 2014

Preface

Mihi

Ko Cymru te whenua
Ko Eryri te maunga
Ko Banwy te awa
Ko Vyrnwy te moana
Ko Robin Amber Atherton tōku ingoa

I was born in England to my father, whose father was a Yorkshireman, and whose mother was a Welshwoman from Anglesey, and my mother, whose parents were both from Yorkshire. I was raised partly in South Africa, but mostly in Cymru (Wales) in a small village called Y Foel nestled in the hills in the mid-central part of the principality. At primary school I learnt Welsh in full-immersion and delved into the Welsh world feet first, learning to recite Welsh poetry, sing Welsh songs and participating in cultural competitions, known as *eisteddfod*. My roots are firmly planted in the alluvial soils of the Banwy region, it is where I feel empowered and connected; it is my foundation, my home, my *tūrangawaewae*.

My love for *Papa-tū-a-nuku* (Mother Earth), the world around us, and my interest in languages and travel, brought me to Aotearoa to continue my studies. It felt comfortable here, like a second home, and I started to learn Te Reo Māori. Being the mother of a Māori child, my world and Te Ao Māori (the Māori world) become closer with each passing day.

My PhD research has taken me all over this beautiful land, collecting leaf samples and measuring karaka/kōpi tree trunks. I am fortunate to have seen hidden coves and inlets, cliffs and coastal banks, isolated hilltops and bluffs, that few others have. Through my study of the karaka tree, my roots have sunk deep into Papa-tū-a-nuku, and Aotearoa is now my home.

Mā te rongo, ka mōhio;

Mā te mōhio, ka mārama;

Mā te mārama, ka mātau;

Mā te mātau, ka ora.

Through resonance comes cognisance;

through cognisance comes understanding;

through understanding comes knowledge;

through knowledge comes life and well-being.

ABSTRACT

Polynesians translocated a number of plant species around the Pacific region. Many of these tropical crops were probably introduced to New Zealand, however, only a few survived owing to the cooler climate. Compensating for the loss of introduced crops, Māori cultivated endemic species they discovered in New Zealand. This project focuses on cultural and evolutionary aspects of the cultivation of one of these, karaka (*Corynocarpus laevigatus* Forst. & Forst.), which was cultivated for its highly nutritious kernel. Originally it is thought to have been restricted to the northern North Island. Its occurrence in the southern North Island, the South Island, Chatham and Kermadec Islands is strongly associated with Māori and Moriori archaeological sites and considered to have resulted from translocations as part of its cultivation. For this project, hypotheses were formulated based on existing written accounts of oral histories, published studies on karaka and informal observations and recollections. Oral histories exist regarding the origins of some translocated populations and have the potential to play an important role in tracing the history of karaka.

The relationships among the five *Corynocarpus* species were investigated by analyzing DNA sequences amplified using universal nuclear and chloroplast markers to test hypotheses of the inter- and intraspecific relationships of the genus. Nuclear markers suggest a closer relationship between *C. laevigatus* and *C. dissimilis* whereas the interpretation from chloroplast markers is less clear. This is indicated by the *rbcL* and *trnL-trnF* networks, which both show a reticulation suggesting support for both *C. laevigatus* and *C. similis* being more closely related to each other and *C. laevigatus* and *C. dissimilis* being more closely related. Nevertheless, in all cases, all markers suggest a close relationship between *C. laevigatus* and *Corynocarpus* species to the north of New Zealand (*C. dissimilis* in New Caledonia and *C. similis* in Vanuatu).

Using universal primers, intraspecific variation within karaka was found to be too low for studying translocation histories within New Zealand and extensive marker development was necessary. The first step in the development of chloroplast markers was characterisation of the chloroplast genome as a reference for different strategies in

molecular marker identification. A protocol was developed for the isolation of chloroplasts and the sequencing of the chloroplast genome using the Illumina Genome Analyser II. This protocol was also shown to be effective in the characterisation of chloroplast genomes in other elements of the New Zealand flora.

The sequence variability of the karaka chloroplast genome was investigated as a potential source for seed dispersal markers. A set of seven chloroplast molecular markers was developed and evaluated in terms of their potential for elucidating the history of karaka translocation during Māori settlement of New Zealand. Long-range polymerase chain reaction products were amplified from the chloroplast genome sequenced using Illumina Genome Analyser II, which enabled the identification of 48 putative chloroplast single nucleotide polymorphisms (SNPs). Sanger sequencing validated 16 of these detected SNPs. High resolution melting (HRM) was evaluated as an accurate, sensitive and fast PCR-based method to screen SNP variations in the chloroplast genome of karaka. Sufficient resolution in the data enabled an evaluation of the phylogeographic distribution of karaka to provide insight into the extent of human-mediated dispersal of the tree in New Zealand.

The results of the analysis of species-specific markers show the potential of the chloroplast genome to study recent events in plant history, and the use of HRM to assay several hundred accessions for a suite of chloroplast SNPs. They show an interesting relationship between Kermadec Island karaka and mainland karaka, and between Rekohu/Chatham Islands karaka and mainland karaka. To be able to pinpoint the location of the source for Rekohu/Chatham Islands karaka, more genetic work is required. However, these results are promising in their ability to trace the translocation of one of New Zealand's most important ethnobotanical species. By developing a more detailed picture of the genetic variation of karaka, this work has the potential to be the foundation for a deeper study into the translocation of the species. This has implications for further understanding the level of domestication in karaka, which at present cannot be ascertained.

Acknowledgments

Ehara taku toa, he taki tahi, he toa taki tini

My success should not be bestowed onto me alone,
as it was not individual success, but success of a collective

It would not have been possible to write this doctoral thesis without the help and support of some wonderful and kind people around me. I thank everyone who has offered help and advice over the last five years.

Above all, I would like to thank my principal supervisor Dr Lara Shepherd for her vision, hard work, personal support and great patience at all times. My supervisor Professor Peter Lockhart who has given me his unequivocal support throughout, as always, for which this expression of thanks does not suffice. This thesis would not have been possible without the help, support and patience of my third supervisor Dr Nick Roskruge (Te Ātiawa), not to mention his advice and unsurpassed knowledge of Māori customs, culture and protocol. Trish McLenachan, no words can express my gratitude for all your tireless efforts to help me in the lab and for all the sequencing you did for me. Thanks too for all your emotional support and comments on each chapter.

I am most grateful to Peter de Lange for providing me hundreds of karaka leaf and herbarium accessions from some far flung places on these two islands, the collection of which has been valuable, time-saving and at times, for Peter, both exciting and dangerous. Peter, you are like a mountain goat, I am sincerely thankful for all your efforts. Thank you too to the following people who have also provided karaka samples or assisted with sample collection: Lara Shepherd, Leon Perrie, Jeremy Rolfe, Patrick Brownsey, Barry Sneddon, Jean-Claude Stahl, Bill Wallace, Simon Cox, David Havell, Janene Collings, Geoff Wall, Kay Kitchener, Craig McGill, Anna McNaughton, Patricia Aspinal, Kevin Matthews, Z Stevenson, Joseph Potangaroa, Eleanor Burton, Mike Shepherd, Stephen King and Jill Rapson. Thank you Bex Smith and Josie Monaghan who both helped me when I was heavily pregnant and still out collecting samples!

The good advice, support (especially through those last tough weeks) and friendship of my dear friend Dr Andrew Clarke, has been invaluable on both an academic and a personal level, for which I am extremely grateful. Andrew, you have been a pleasure to come to know, and, besides me, you are the funniest person I know! Big thanks to Chris Stowe whose Masters thesis was a fantastic source of information, references and location data. I was fortunate enough to accidentally meet Chris at a beach house on the Bay of Plenty coast (my brother is friends with Chris's brother) so we got to swap notes on karaka.

This project would not have been able to get off the ground without close contact with those involved in Te Ao Māori and Te Ao Moriori (the Māori World and the Moriori World). I have learnt so much about tikanga Māori and have a deep respect for Māori and Moriori culture and traditions. I would like to sincerely thank Maui Solomon (Hokotehi Moriori Trust) and Susan Thorpe, tēnā kōrua, your acceptance, love, help, guidance and support has been second to none, thank you for accepting both Kōpi and me into your whānau, and for supporting me naming Kōpi after the kōpi groves. Many thanks to Tom Lanauze, a fount of knowledge; Horipo (Dane) Rimene and Joseph Potangaroa (Rangitāne ō Wairarapa); Haami Te Whaiti (Ngāti Hinewaka); Henare Manaene (Ngāti Kahungunu ki Wairarapa); Clive Stone (Ngāti Wai); Alex Nathan (Te Rōroa Whatu Ora Trust); Barney Haami (Te Rūnanga o Tamaupoko); Utiku Potaka (Ngāti Hauiti); Jon Proctor (Rangitāne o Manawatu); Marty Davis (Te Kahui ō Rauru); Luana Pirihi (Patuharakeke Trust Board), Frances and Greg White (Ngāti Tama), Michelle Wi (Te Aupōuri); Victor Holloway and Alan Hetaraka (Ngāti Kahu); Rongo Bentson and Ani Walker (Te Rarawa); Jonda Subritsky (Ngāti Kuri); Rachel Puentener (Ngai Tahu); Mark Te One (Te Ātiawa, Taranaki Whanui); Paula Wilson (Pātaka Komiti); Bobby Morehu (Ngāti Tūwharetoa); Ngāti Rārua, Te Ātiawa, Ngāti Tama, Wakatu Inc and Ngāti Rārua Ātiawa Iwi Trust (Tiakina te Taiao Ltd.).

Sincere thanks to the following landowners, who allowed me access to their land to collect karaka samples: Hugh Wilson, Ōtanerito Arboretum Hinewai Reserve; Bryan and Helen Hocken (QE2 covenant), Tarata; Murray and Pahau Thacker, Okains Bay; Stan and Mary, Marakopa; John and Mary McGuiness, Flatpoint Station, Wairarapa; Jo Tuanui, Waihi, Rekohu/Chatham Islands; Arthur Bowen, Mahanga. Thank you to the wonderful people who picked me up hitchhiking around Banks Peninsula when I forgot

my driving license and couldn't hire a car. Thank you Bill Wallace (QE2 Trust), Tony Silbery, Rod Wallace, John Ogden, Mere Roberts, Helen Leach and Michael Taylor for your support and advice.

I would like to acknowledge the financial, academic and technical support of the Allan Wilson Centre and Massey University, Palmerston North and its staff, particularly in the award of an Allan Wilson Centre scholarship, a Massey Doctoral Research Scholarship and an Institute of Fundamental Sciences departmental scholarship that provided the necessary stipend support for this research. A Royal Society of New Zealand Marsden grant (MAU 07-MAU-090) financed the lab consumables and fieldtrip costs. A JP Skipworth Scholarship for Plant Biology, a Heseltine Bursary and two travel awards from the Institute of Fundamental Sciences funded three separate trips to Rekohu/Chatham Islands. Many thanks to Ann Truter, Cynthia Cresswell and Joy Wood for secretarial support and Katrina Ross for advice and guidance throughout this time. I owe a debt of gratitude to the ladies at the International Student Office at Massey University: Natalia Benquet, Olive Pimentel, Diane Reilley and Sylvia Hooker.

For technical support, Dr Lesley Collins, Senior Research Fellow, Massey University, thank you for suggesting edits to all chapters and for your emotional support, patience and comments. Thank you Dr Patrick Biggs, Senior Lecturer in Computational Biology, mEpilab and Infectious Disease Research Centre, Massey University, for all your help and for producing the Circos plot in Figure 4.6, Chapter 4. Huge thanks to Dr Andrew Clarke, Research Fellow at University of Warwick, for producing Figure 1.9 in Chapter 1, to Matt Irwin, GIS/Remote Support Officer, Institute of Agriculture & Environment at Massey University, for producing Figure 4.6 in Chapter 4 and to Rachael Ouwejan for producing Figure 4.7 in Chapter 4.

Special thanks also to all my graduate friends, especially PLEB, Farside and Phoenix lab members: Lizzie Daly, Dr Barbara Schönfeld, Dr Bennet McComish, Juan Carlos Garcia-Ramirez, Dr Simon Hills, Dr Gillian Gibb, Dr Nick Albert, Simon Cox, Tariq Mahmood, Josie Monaghan, Ibrar Ahmed and Dr Jian Han for sharing the literature and invaluable assistance. Thank you to Nick, Andrew, Barbara and Gillian for comments on some of the chapters. I really appreciated comments, advice and corrections from all

three examiners: Dr Phil Wilcox (Scion), Professor David Penny (Massey University) and Dr Michael Knapp (Bangor University).

Without the support of friends, I am not sure I would have made it to publishing this thesis. Jade, my soul-sister, thank you for all the love and support you gave to me, your friendship is pure gold. Peter Horsley, thank you from the bottom of my heart, you showed me such kindness and compassion, I truly appreciate the refuge of your lovely big house, where Kōpi was born, where I spent several months sitting in the garden listening to the birds. Thanks to Amy and Pete for letting us live with them for six weeks; to Jess and Richard for providing a refuge for us at Westoe for the last five months of thesis writing; thanks to Tabitha who looked after Kōpi one day a week for months on end so I could write; to Debbie, Kristal, Nirbha, Rachel H, Rachel A, Michelle, Bianca, Jules, Sophie, Annie, Susan, Heather, Rebekah, Lucy and Lynlee. Huge thanks to my sister Melanie, for standing by my side through all the important and life-changing parts. It really does take a village to raise a child (and support a mother writing a thesis at the same time). Mum and dad, I know you have no idea if I actually DO anything important, but I know you believed in my ability to nail this beast, and that's all I needed.

I dedicate this doctoral thesis to my son Kōpi, named for the kōpi trees on Rekohu/Chatham Islands. Having a small child AND a thesis to write is like having twins: both demand your constant attention and keep you up all night worrying. Kōpi, you have been such a patient and happy child, and although it has been anything but plain sailing, it has been a joy to share this journey with you. You have been a constant reminder of why I needed to complete this thesis. I love you with all my heart.

Table of Contents

List of Figures

List of Tables

1

General Introduction

1.1 Chapter overview

The aim of this project was to use a molecular (DNA) approach to reconstruct the translocation history, and dispersal in New Zealand, of the evergreen tree karaka, *Corynocarpus laevigatus* (Forst & Forst), and to use the inferences of the patterns in the genetic data as a proxy for human mobility. The use of molecular markers to determine the natural and translocated range of karaka can also be utilised to determine the nature of domestication events; when a plant begins its journey towards full domestication, is there an initial loss of genetic variation, or does this loss occur over time?

Karaka was one of the most important staple food crops for the ancestors of modern Māori. It replaced some of the Polynesian crops that were introduced to New Zealand but failed to thrive due to their tropical nature. This chapter serves as an introduction to the species; as well as its taxonomy, biology, dispersal, distribution and uses. Domestication is defined and described and the term applied to the translocation and possible cultivation of karaka.

1.2 INTRODUCTION

Karaka (*Corynocarpus laevigatus*, Figure 1.1) is a broadleaved evergreen lowland tree endemic to New Zealand and its outer islands. The nutritious kernels were a staple part of the Māori diet. Karaka is a feature of Māori oral history, and these accounts tell of its arrival in New Zealand, while others talk of its uses other than as a food source. There is no doubt karaka had begun its journey along the domestication continuum, but to what extent that has occurred, is not currently known.

1.2.1 CORYNOCARPACEAE

1.2.1.1 TAXONOMY

Corynocarpus was circumscribed by J. R. and G. Forster[1] in 1775 and described from specimens collected in New Zealand during James Cook's second voyage (1772-1775) (Hemsley, 1903). The species of *Corynocarpus* have been clearly defined, however, the genus has proved difficult to place within the natural phylogenetic system (Carlquist & Miller, 2001). It had previously been placed in the Myrsinaceae, Theophrastaceae, Terebinthaceae and Anacardiaceae, amongst others (Hemsley, 1903). Engler (Engler, 1897) redescribed and figured *C. laevigatus* as the type of a new family, Corynocarpaceae. In 2000, analysis of sequences from the chloroplast-encoded gene *rbcL* firmly placed *Corynocarpus* in its own distinctive family, Corynocarpaceae, next to Anisophylleaceae, Begoniaceae, Coriariaceae, Cucurbitaceae and Datiscaceae which comprise the order Cucurbitales (Wagstaff & Dawson, 2000). Carlquist and Miller (2001) confirmed this placing after analysis of the wood of *Corynocarpus* within Cucurbitales, three superfamilial clades are supported by floral structure: Tetramelaceae/Datiscaceae, Tetramelaceae/Datisaceae/ Begoniaceae and Corynocarpaceae/Coriariaceae (Matthews & Endress, 2004).

[1] The Forsters were visiting botanists on Cook's Second Voyage

FIGURE 1.1: Karaka (*Corynocarpus laevigatus*) in fruit. Photo courtesy of Lana (http://www.flickr.com/photos/53936550@N03/6173892427/)

1.2.1.2 DISTRIBUTION

The family Corynocarpacaea consists of five species found in tropical to warm temperate areas in the southwest Pacific (Figure 1.3). *Corynocarpus similis* is found in Vanuatu, the Solomon Islands, New Britain, New Ireland, and the Bismarck Archipelago; *C. cribbianus* is found on the island of New Guinea (French, 2006) and northeastern Queensland (van Steenis, 1951). *Corynocarpus rupestris* occurs in isolated locations in Australia and has two subspecies: (i) *C. rupestris* subsp. *rupestris*, also known as Glenugie Karaka, occurs in the Clarence Valley near Coffs Harbour, near Grafton and in the Tenterfield area of New South Wales and is listed as vulnerable (Briggs & Leigh, 1996); (ii) *Corynocarpus rupestris* subsp. *arborescens* occurs in southeast Queensland (Guymer, 1984 cited in Wagstaff & Dawson, 2000). *Corynocarpus*

dissimilis is endemic to New Caledonia (Hemsley, 1903) and *C. laevigatus* is confined to mainland New Zealand (Aotearoa) and its offshore islands, Rekohu/Chatham Islands and the Kermadec Islands (Molloy, 1990). Figure 1.3 shows the morphological distinctiveness of each of the species. Wagstaff and Dawson (2000) suggest a palaeotropical center of origin for Corynocarpaceae, followed by two independent radiations into cooler climates. The first radiation comprised *C. cribbianus* and *C. rupestris* extending through New Guinea to central Australia, and the second comprised *C. similis*, *C. dissimilis* and *C. laevigatus*, with *C. laevigatus* reaching New Zealand, through New Caledonia several million years ago (Figure 1.2). Species are morphologically distinct from one another (Figure 1.4).

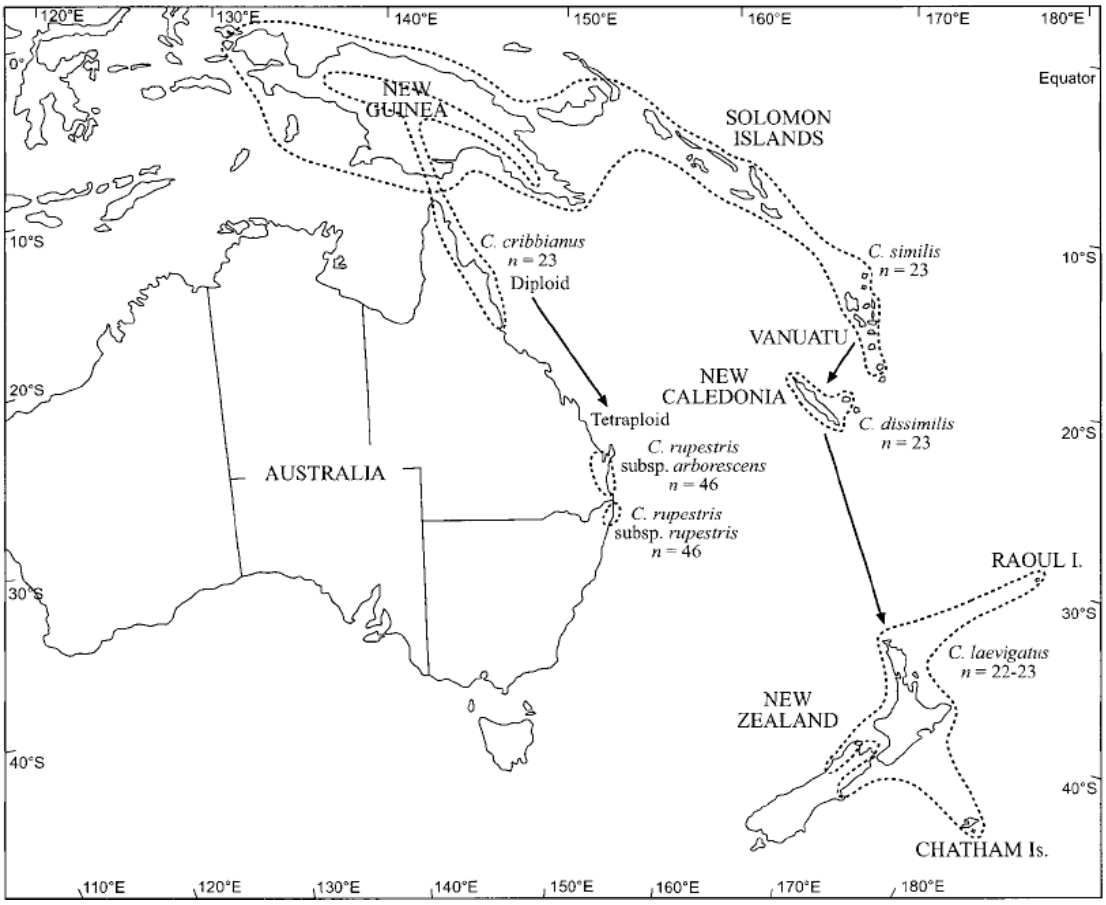

FIGURE 1.2: Map showing the present distributions and chromosome numbers of *Corynocarpus* species. Their hypothesised origin and radiation are illustrated with arrows. Figure reproduced from Wagstaff & Dawson (2000).

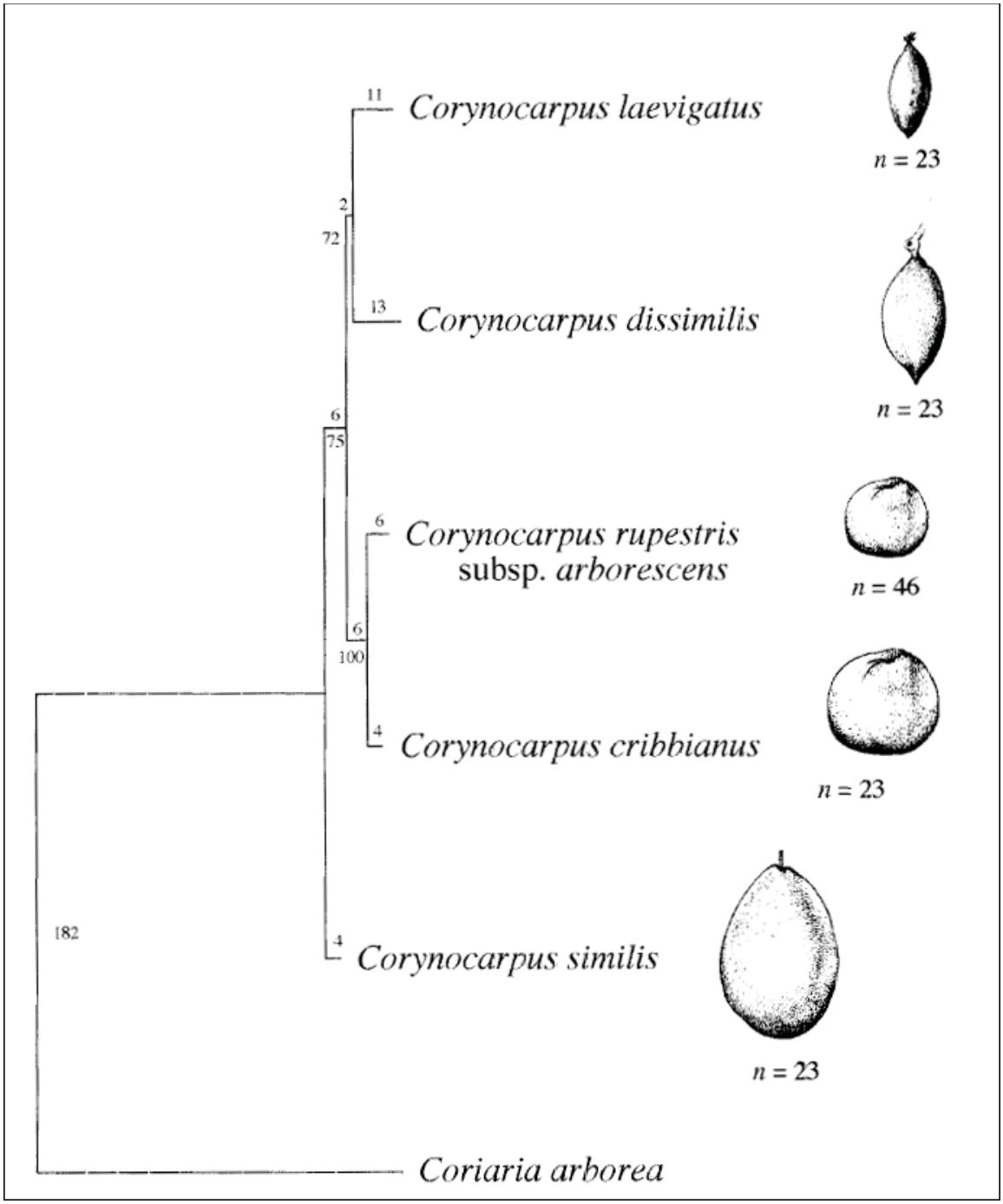

FIGURE 1.3: Relationships within *Corynocarpus* inferred from combined analysis of *rbcL* and ITS sequences. Illustrations of fruits from Molloy (1990) and reported chromosome numbers from Dawson (1997) are provided alongside each taxon. The number of changes is given above each branch and bootstrap values are given below. Figure reproduced from Wagstaff & Dawson (2000)

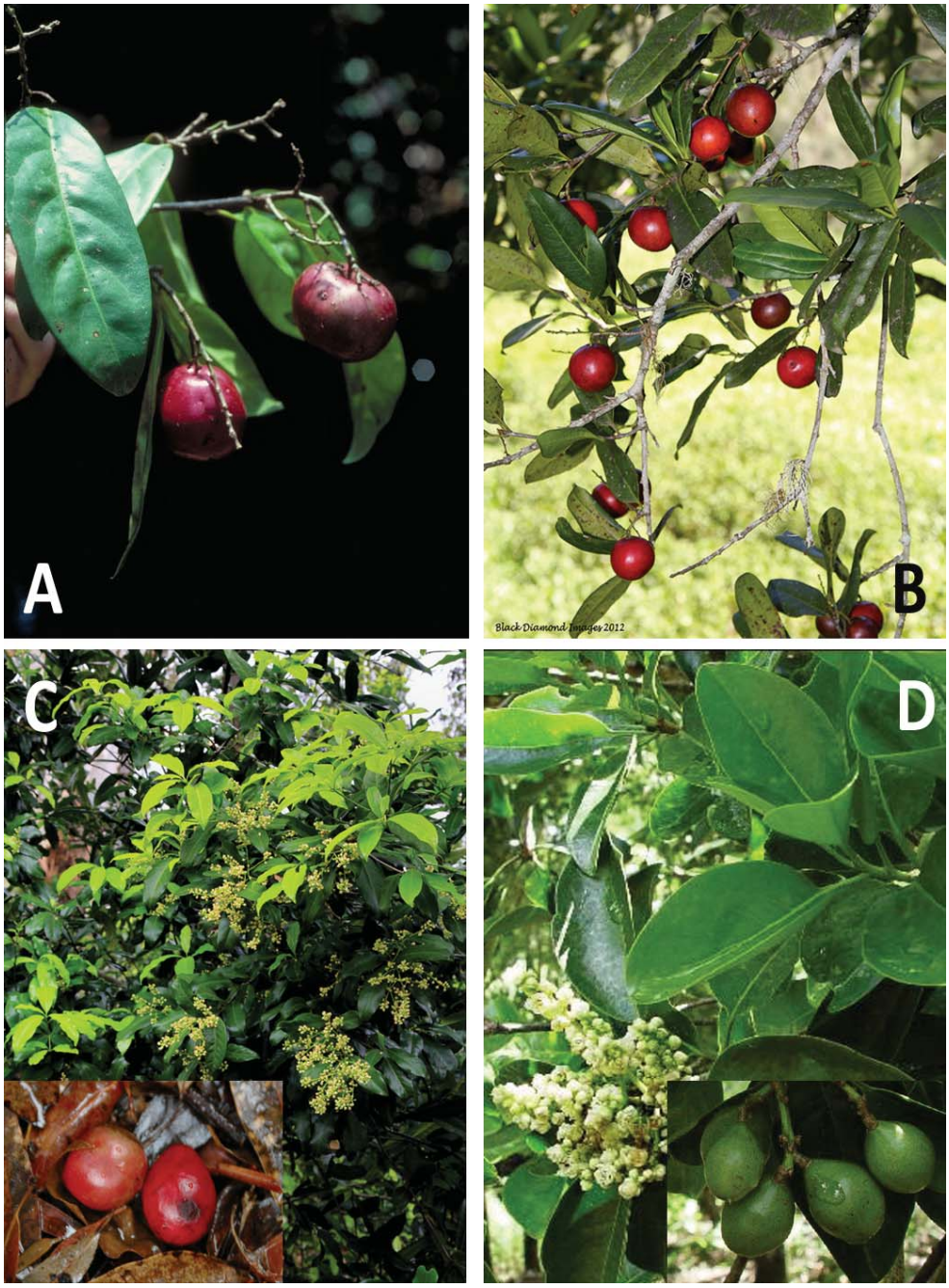

FIGURE 1.4: *Corynocarpus* species endemic to regions outside New Zealand: A) *C. cribbianus*, photo courtesy of and copyright of CSIRO Australia, source: http://www.cpbr.gov.au/cpbr/cd-keys/rfk/; B) *C. rupestris* ssp *arboreus.* photos - courtesy Black Diamond Images, source http://www.flickr.com/photos/blackdiamondimages/6986211567/; C) *C. rupestris* ssp *rupestris* (inset, fruit), photos - courtesy Black Diamond Images, sources http://www.flickr.com/photos/blackdiamondimages/6424048959/ and http://www.flickr.com/photos/97974874@N00/3036455128/; and D) *C. dissimilis* (inset, fruit) photo courtesy of and copyright of Bernard Suprin, source: http://www.endemia.nc/flore/fiche792.html

1.2.1.3 *CORYNOCARPUS* IN NEW ZEALAND

Uplifting of the southern end of the Norfolk Ridge during the Oligocene extended the New Caledonia landmass to 32˚S and land connections via island chains (Herzer *et al.*, 1997) could have facilitated dispersal of karaka into New Zealand from New Caledonia (Stowe, 2003). Fossilised kernels of karaka were discovered at Landslip Hill in Southland, New Zealand dating back to the early Miocene (~24 mya[2]) (Campbell, 2002) confirming the arrival in New Zealand during the mid-tertiary. Macrofossil remains of the other species (*Avicennia*, *Pomaderris*, and *Pouteria*) found at Landslip Hill indicate a deltaic-coastal ecosystem similar in nature to the vegetation of modern northern New Zealand and New Caledonia (Campbell, 2002).

In the late Oligocene-early Miocene the area around Gore, Southland, would have been at a latitude of more than 50˚S (Cook *et al.*, 1999). It is unlikely that these plants would survive a similar modern day latitude suggesting global temperatures in the mid-Cenozoic were warmer (Campbell, 2002). The mid-Pliocene saw a gradual reduction in the number of taxa of tropical and subtropical affinities in the northern South Island, and by the Pleistocene most of these taxa had disappeared from the flora (McGlone, 1985).

Tectonic and glacial events have determined the distribution of species in New Zealand; climate also plays a significant role. The southern limit of many of the species restricted to the northern North Island is approximately 38°S; this boundary is where the warmer climate of the northern region meets the cooler climate of the southern (Garnier, 1958). The northern North Island contains a high number of endemic plants. In the ecological zone above 39°S latitude the total number of endemics is 125, with endemics making up 5.7% of the total flora of that region; above 38°S the figures are 95 and 11% respectively (McGlone, 1985). Of these endemics, a large proportion is woody plants and tall trees. Northland has been a tectonically stable region of New Zealand, retaining a diverse flora and has acted as a refuge for some components of the flora during the Pleistocene. McGlone (1985) believes major refugia in this region occurred north of latitude 38-39°S. During glacial periods, Northland is thought to have the only large continuous tract of

[2] Mya – million years ago

forest in New Zealand (McGlone, 1985).

Karaka's association with other plants in the Miocene (Campbell, 2002), which are now confined to Northland, has been purported by some to suggest that its range prior to human arrival in New Zealand was probably restricted to Northland (Stowe, 2003). Extensive work on the extant distribution of the species classified karaka populations as either cultural or unknown (Stowe, 2003). Cultural populations were those that were found growing within 500m of a registered archaeological site (pa, storage pits, terraces, gardens, stone walls, middens or cultivation areas) and unknown populations are those that had no association with the above site types. In most regions of New Zealand karaka classified as cultural far outnumbers those classified as unknown. However, in Northland, they occur in equal numbers. This adds weight to the suggestion that Northland could be the natural range for karaka, although Stowe (2003) suggests that range could be as far south as Taranaki and Wanganui in the western North Island, and as far east as the Coromandel Peninsula, due to the number of karaka classified as unknown occurring across this region.

Karaka is a climax broadleaf forest species naturally found growing with puriri (Vitex lucens), taraire (Beilschmiedia tarairi) and kohekohe (Dysoxylum spectabile). Platt (2003) considers it reasonable to assume that where these four trees co-exist, karaka trees are natural components of the surrounding flora. Today, karaka grows mainly in coastal regions from Cape Reinga to Banks Peninsula, although populations do occur inland, particularly in the North Island (Figure 1.5). Translocated populations have subsequently naturalised in unmanaged vegetation (Burrows, 1996) to the point where it has been considered a weedy invader in forest remnants in the Wellington region, (Costall et al., 2006) where regeneration of karaka in existing plant communities has been described as aggressive (Sawyer et al., 2003).

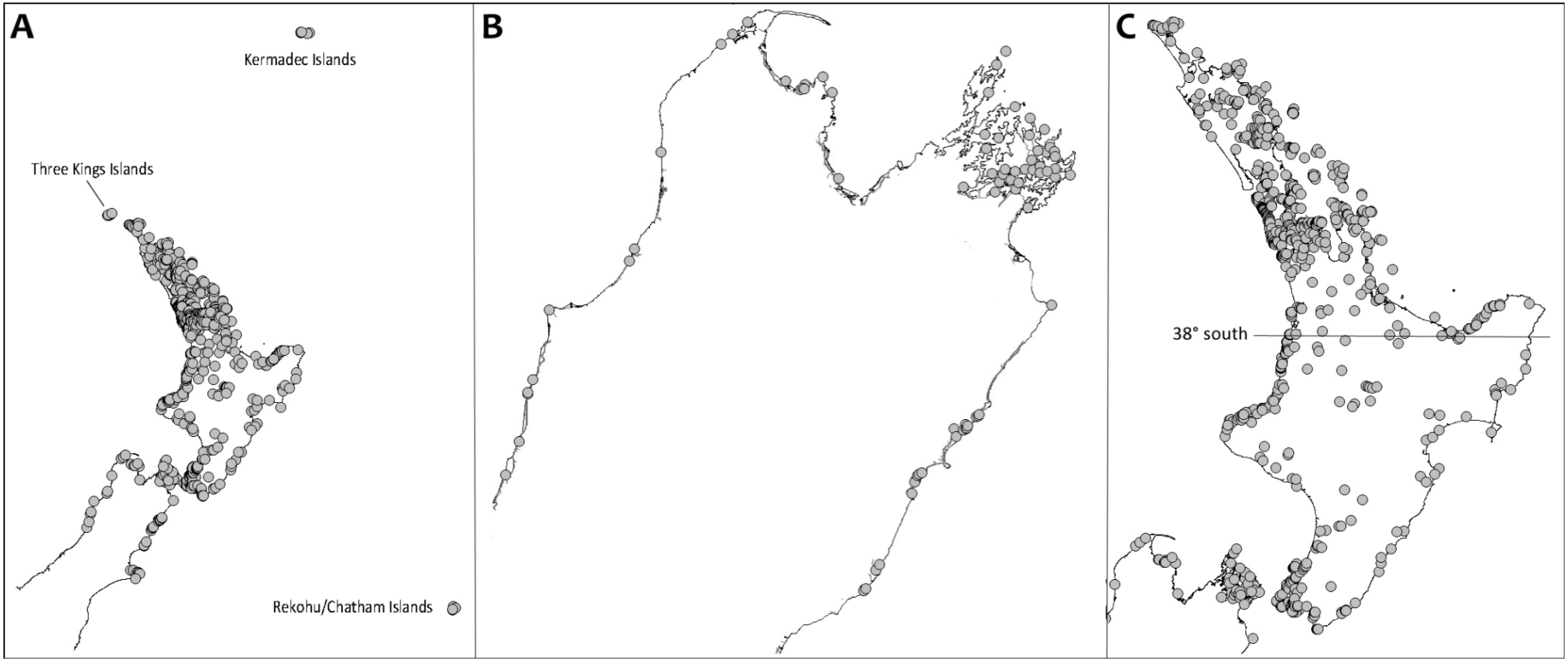

Figure 1.5: Corynocarpus laevigatus distribution showing (A) the distribution of karaka across New Zealand including the Kermadec Islands to the north and the Chatham Islands to the west; (B) the distribution of karaka in the South Island of New Zealand; and (C) the distribution of karaka in the North Island of New Zealand showing locations of inland populations based on records in the AK, CHR, NZFRI, and WELT herbaria (abbreviations follow Holmgren et al. 1990).

1.3 THE BIOLOGY OF KARAKA

1.3.1 PHENOLOGY

Karaka is a tall, spreading evergreen tree growing to a height of approximately 15 metres, found mainly in coastal regions throughout New Zealand (Clarke, 2007). Trees flower and fruit from 10 years old, sometimes younger (Molloy, 1990), and fruit ripening times range from January to April depending upon latitude (Allan, 1961 cited in Stowe, 2003). The fruit are small drupes, up to 5cm in length, with smooth skin that turns orange when ripe. The flesh of the drupe covers a tough fibrous endocarp, inside which is the highly prized seed. Seeds contain a bitter, toxic compound called karakin which, in its untreated state, is poisonous to humans (Skey, 1871). Karakin interferes with ATP synthesis resulting in weakness, hind leg paralysis, and convulsions (Parton et al., 2001; p. 345). Karakin is known to be toxic to brown kiwi causing anorexia, lethargy, and the inability to walk (Shaw & Billing, 2006) and to honeybees, causing an inability to fly (Palmer-Jones & Line, 1962). Kererū are not affected by karakin but have been described as appearing drunk after gorging on karaka fruit (Shaw & Billing, 2006). Cattle and sheep often eat the fruit whole and remain unharmed by the toxin (Molloy, 1990).

Karaka seeds show evidence of recalcitrance, like many trees of tropical and subtropical affinities. Recalcitrance is a broad term relating to the susceptibility of a plant to post-harvest desiccation and intolerance to freezing temperatures. Recalcitrance can impair germination and seeds are usually shed when the water content is high and when they are more sensitive to desiccation (Bannister et al., 1996). Despite this, karaka seed is capable of germinating within days of falling from the tree (Burrows, 1996; Dijkgraaf, 2002). Although germination can occur soon after falling from the tree or soon after sowing, the peak is usually May-July (Burrows, 1996).

1.3.2 POLLINATION BIOLOGY

Garnock-Jones et al. (2007) describe karaka as exhibiting gender dimorphism. Trees are called male or female even though many male trees set fruit. When assessed, flowers on male karaka trees were found to produce a large amount of pollen and each flower had a well-formed ovule. On female trees flowers had fully formed anthers but these contained no pollen (Garnock-Jones et al., 2007). Female trees typically set large numbers of fruit on every inflorescence but males trees varied in their fruit set, generally producing fewer fruit than female trees. The low fruit

production on male trees might be explained by early acting genetic load owing to self-pollination (Garnock-Jones et al., 2007). Figure 1.6, shows a visual comparison.

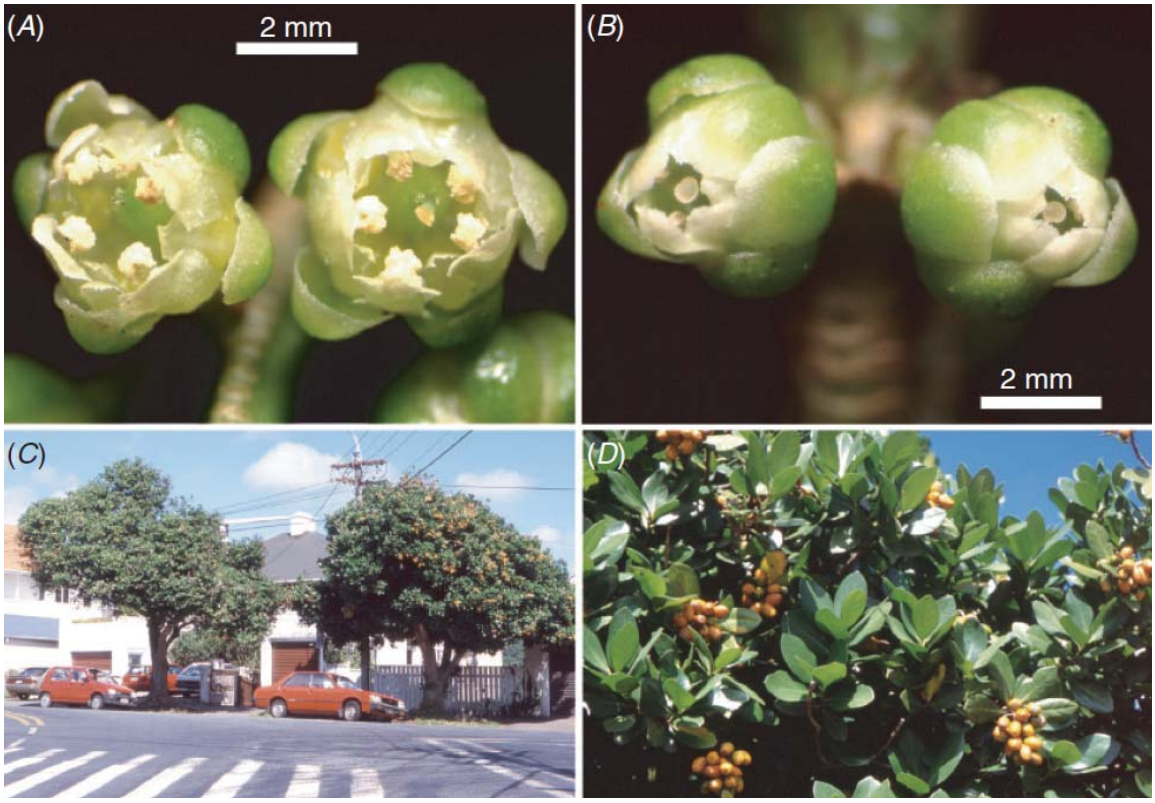

FIGURE 1.6: Corynocarpus laevigatus, showing (A) flowers from a male tree; (B) flowers from a female tree; (C) male tree (left) and female tree (right), and (D) fruiting branches on a female tree. Figure reproduced from Garnock-Jones et al. (2007)

1.3.3 LIFE-CYCLE STRATEGY

The life-cycle strategy of an organism can indicate the likelihood of it becoming invasive. MacArthur and Wilson (1967) proposed two life-cycle types, which describe opposite life-cycle strategies: *r*-selected species are those with short life-spans, reaching sexual maturity quickly and shedding numerous well-dispersed seed, whereas *K*-selected species are long-lived, produce large seed that fall and germinate under the parent plant and have shade-tolerant seedlings. In New Zealand, examples of *r*-species include *Leptospermum scoparium* (manuka) and *Schelfflera digitata* (pate), both of which are rapid invaders of disturbed areas and tree-fall gaps and K-species examples include *Beilschmeidia tawa* (tawa) and *Prumnopitys ferruginea* (miro) (Ogden, 1989). Stowe (2003) describes karaka as exhibiting the features of a K-selected species. At the extreme,

K-species should be self-perpetuating *in-situ* with a J-shaped frequency distribution for population size and/or age (Stowe, 2003) Karaka tends to form groves, and has the potential to remain in situ indefinitely (Stowe, 2003) .

The characteristics of the K-selected life-cycle of karaka would suggest it is not a coloniser of disturbed sites. However, Costall *et al.* (2006) suggest some of these life-cycle traits are what have resulted in karaka being described as a 'weedy' invasive in fourteen sites in the southern North Island, including Taranaki and Wellington regions. From their investigations into the invasive nature of karaka, Costall *et al.* (2006) recommend management of karaka invasiveness in the form of elimination or control, depending on local cultural values.

In the South Island, karaka do not appear exhibit this invasive tendency, existing in patchy and isolated dense groves (Molloy, 1990). Stowe (2003) suggests the ability of karaka to spread rapidly and become invasive varies with region, according to the presence or absence of dispersal agents, climate and predation by mammals

1.3.4 DISPERSAL OF KARAKA

Endozoochory, the ingestion and dispersal of seed by animals and birds, is the dispersal mechanism for most tree species. Karaka relies largely on frugivores, both native and non-native, for dispersal of its large fruits.

KERERŪ

It is widely believed that kererū (*Hemiphaga novaeseelandiae*) (Figure 1.7A), the New Zealand native woodpigeon, is currently the sole native dispersal agent of karaka berries (Sawyer *et al.*, 2003), (Wotton & Ladley, 2008). Kererū are the largest extant volant bird native to New Zealand (Lord *et al.*, 2002) and are capable of flying long distances. The gape size is 14 mm although it is capable of distending to enable it to swallow fruits up to 25 mm diameter (Gibb, 1970 cited in Clout & Hay, 1989). Karaka form quite a large part of the kererū diet [at certain times of year] and karaka now relies almost exclusively on kererū for dispersal (Dijkgraaf, 2002). In a study of the diet of kererū, Dijkgraaf (2002) found that karaka comprised just 4% of feeding observations, suggesting kererū do not favour karaka fruit. However, this study was conducted over a four year period in

which karaka did not fruit heavily in all years. Added to this is the short fruiting period of karaka compared to puriri (*Vitex lucens*) and taraire (*Beilschmiedia tarairi*), the preferred fruits of kererū.

FIGURE 1.7: (A) Kererū eating karaka fruit (photo credit: Monica Awasthy from (KDP, 2008)); (B) Ripe fruit of tawapou (*Pouteria costata*) a native of the Northland region of New Zealand. Photo credit Nga Manu images.

In the Chatham Islands, Pearson and Climo (1991 cited in Campbell, 2006) found that parea (*Hemiphaga novaeseelandiae chathamensis*), the Chatham Island woodpigeon, only used kōpi/karaka for loafing and preening and not for eating. In a later study (Powlesland *et al.*, 1997) feeding observations on kōpi/karaka accounted for 2.2% of total observations for the month of April confirming parea do eat them as a major part of their diet. However, when compared to matipo (*Myrsine chathamica*) and mahoe (*Melicytus chathamicus*), which, respectively, make up 24.3% and 36.5% of the observations in the same month, it is clear that karaka is not the preferred food of the parea, just as it is not on the mainland for kererū.

Kererū have long seed retention times for larger seeds such as tawa, miro, taraire and pūriri ranging from 90-180 minutes although seed passage time in kererū increases as seed size increases (Wotton *et al.*, 2008). Pigeons can fly several kilometres in one flight and this, coupled with long seed retention times, can lead to greater dispersal distances making kererū an important seed disperser for large-seeded trees (Wotton *et al.*, 2008). Campbell (2006) believed karaka seeds would be dispersed by kererū in areas of existing forest rather than in scrub or regenerating scrub where gorse is the nurse species.

The kererū is also an important dispersal agent for tawapou (*Pouteria costata*) (Dijkgraaf, 2002) a species that was found growing with karaka in Miocene deposits in Southland (Campbell, 2002) (see Section 1.2.1.3) and which looks strikingly similar to karaka (Figure 1.7B). The taxonomy of *Pouteria* is yet to be resolved; some taxonomists place it in the genus *Planchonella* others split its species into two separate genera, *Pouteria* and *Planchonella*. Many species of *Planchonella* are locally known as karaka or kalaka (or as cognates of these words) in the Pacific region (see Section 1.5).

Other birds capable of dispersing karaka and with a gape size of similar size or larger are summarised in Table 1. Of these, the species of extinct moa, being the largest of all the New Zealand birds, would seem to be an obvious disperser of large fruit. Moa had a gape size of up to 5cm (Clout & Hay, 1989) and would have been capable of swallowing karaka fruit and dispersing them over long distances.

MOA SPECIES

Clout & Hay (1989) suggest moa may have had a role similar to that of the cassowaries (*Casuarius casuarius*) in North Queensland and Papua New Guinea, consuming great quantities of fallen fruit and depositing the seeds several kilometres away (Figure 1.8). However, moa consumed a varied diet consisting mainly herbs and sub-shrubs. Wood *et al.* (2008). Upland moa (*Megalapteryx didinus*) fed on both woody and herbaceous plants and were seed dispersers for a range of plants, including *Fuchsia excorticata* (Wood *et al.*, 2012b). Wood *et al.* (2012b) found that the diet of this species of moa contained 67 different plant species and, for the first time, found nectar-rich flowers of *Fuschia* and *Phormium* made up part of the diet.

Although considered potentially important seed dispersers, larger and denser seeds (e.g. *Prumnopitys*, *Elaeocarpus*) may have been retained in moa gizzards longer, and not pass into the droppings intact, as smaller seeds do (Wood *et al.*, 2008). Another South Island moa, the little bush moa (*Anomalopteryx didiformis*), had a diet consisting of fibrous material from the forest understory (Wood *et al.*, 2012a) and based on this evidence they were probably not important dispersers of seeds. As yet, there is little information of the diets of moa living in the North Island where the three species with the largest

fruit (*Elaeocarpus* spp., *Corynocarpus* and *Dysoxylum spectabile*) grow more abundantly (Lord *et al.*, 2002). These three species do grow in the northern South Island but not in areas where moa gizzards have been studied (Lord *et al.*, 2002). Whilst it is known that moa did not selectively consume large seeds nor were they specialist frugivores (Lord *et al.*, 2002) there is no evidence that they ignored them and therefore could have been a dispersal agent for fallen karaka fruits in the North Island before humans arrived.

Table 1.1: Seed dispersing birds in New Zealand forests.

Weight (g)	Scientific name	Common name	Gape (mm)**
> 5000	*Dinornithidae* (12 spp.)	*Moa	>50
500-5000	*Apteryx australis, A. rowi, A. mantelli*	(Brown kiwi)	24
	Apteryx owenii	[Little spotted kiwi]	21
	Heteralocha acutirostris	*Huia	15
	Hemiphaga novaeseelandiae	**Kererū**	14
	Gallirallus australis	(Weka)	13
	Callaeas cinerea	(Kokako)	13

Key: *Extinct in New Zealand; Square Bracketed – species extinct on mainland New Zealand; Round Bracketed – minor frugivore and/or a species with restricted distribution; Bold type – major frugivore widely distributed. Modified from Clout & Hay (1989).

HUIA

Heteralocha acutirostris, or huia, were a species of New Zealand wattlebird that went extinct in the early 20th century. Buller (1888 in Clout & Hay, 1989) records that huia ate the fruits of pigeon wood *(Hedycarya arborea)* 6-10mm diameter, hinau (*Elaeocarpus dentatus*) 8-10mm diameter, and *Coprosma* sp. 3.5-12mm diameter. Huia had a gape of 15mm (Table 1), bigger than that of kererū (14mm), but it did not have a distensible gape. Clout & Hay (1989) believed that had the diet of huia been better recorded, the list of fruit-producing species they ate would have increased. However, the curved shape of the huia bill would have been more suited to insect foraging rather than handling and eating fruits (Dijkgraaf, 2002) and huia was probably an unlikely disperser of karaka kernels.

KŌKAKO

Kōkako are a species of forest-dwelling New Zealand wattlebird. Before kōkako were restricted to their current highly reduced range in the northern North Island, they would have been important dispersers of seeds across New Zealand. However, it is unlikely they would have rivalled the distances covered by kererū as they are weak fliers and have a permanent range not exceeding 11 hectares (Clout & Hay, 1989). Kōkako are listed as frugivores of up to 35 species including *Prumnopitys ferruginea*, *Dysoxylum spectabile*, *Litsea calicaris*, *Elaeocarpus dentatus*, *Ripogonum scandens*, *Hedycarya arborea*, *Nestegis cunninghamii*, *Rhopalostylis sapida*, *Alectryon excelsus*, *Prumnopitys taxifolia* with fruit larger than 10mm diameter being stripped of its pericarp rather than swallowed whole (J.R. Hay, unpubl. in Clout & Hay, 1989).

FIGURE 1.8: A fresh pile of cassowary dung. The large seeds are from the fruit of *Elaeocarpus bancroftii*. A whole fruit (diameter 4.5 cm) of this species is shown to one side for comparison. Figure reproduced from Stocker & Irvine (1983).

BRUSHTAIL POSSUM

The introduced common brushtail possum (*Trichosurus vulpecula*) eat the fruits of native species and seed passing through their gut is capable of germination, although results differ widely depending upon the plant species (Williams *et al.*, 2000). Williams

observed possum eating karaka berries but only the ripe fruit was eaten and not the kernel.

1.4 THE CULTURAL SIGNIFICANCE OF KARAKA

Colenso (1880) lists three wild uncultivated plants as providing staple foods to Māori: hinau (*Elaeocarpus dentatus*), karaka and tawa (*Beilschmiedia tawa*). He describes karaka fruits as 'scarcely edible' and their processing as 'incredibly labour intensive'. However, after preparation, karaka kernels could be kept for a long time, up to two or three years (Colenso, 1868). Colenso (1880) describes karaka as "of inestimable value to the Māori as a common and useful article of vegetable food, second only to their prized kūmara tuber." Whole communities would go to the karaka woods to collect fruit from the ground and trees and bring them back in baskets to prepare them (Colenso, 1880). The karaka groves did not bear fruit consistently from year to year, and seasons of sparsity were disastrous for tribes because of the importance of the kernels in the Māori diet, (Colenso, cited in Skey, 1871).

1.5 *CORYNOCARPUS* IN THE PACIFIC REGION

Corynocarpus species are also used as a food resource in the Pacific. *Corynocarpus similis* is the most widely distributed species (Wagstaff & Dawson, 2000). Cabalion & Poisson (1987) report that the kernels of *C. similis* are poisonous, containing up to 1% karakin. One of Cabalion and Poisson's co-researchers in Vanuatu recorded an oral history from the Lowo Peter family living in Happyland village south of Erromango, which recounted that the fruits were unsafe to eat and that even livestock refused to eat whole seeds, but instead removed the fruit and ate that (Cabalion & Poisson, 1987). The fruits are the largest of all *Corynocarpus* at about 10cm x 6cm (according to the diagrams drawn by Pat Molloy (1990). In coastal regions on the islands of Aneityum and Tanna in Vanuatu, *C. similis* is grown in smallholder plantations and in gardens and fallow areas amongst other important fruit crops such as coconut, breadfruit (*Artocarpus altilis*) and Tahitian chestnut (*Inocarpus fagifer*) (Clarke & Thaman, 1993). In Vanuatu, an anonymous author (Anon., 1992) wrote that fruits of *C. similis* require "…a very careful preparation in order to eliminate the toxic substances they contain." He goes on to say the fruits are "…only used in the event of a natural disaster when

famine threatens." They have a high nutritional value but only a few fruit are harvested from each tree at a time (Anon., 1992)

Corynocarpus cribbianus is one of 22 species in the Solomon Islands traditionally eaten to supply dietary carbohydrates and one of 11 that has traditional uses as a seasonal or minor food or when other food resources are scarce (Plant Genetic Resource Center, 1996b, pp11-12). The exocarp of *C. cribbianus* is known locally on the Solomon Islands as *'ibo kwao'* and *'ibo bala'*, and is used, once cooked, as a food source. The fruits are quite large (approximately 6cm x 6cm in drawings by Molloy (1990)) and are pounded until soft to make them edible (Plant Genetic Resource Center,1996b). In the south-eastern Solomon Islands, *C. cribbianus* is a locally important tree species found planted in gardens or protected in groves and is found growing around former inland settlement sites (Clarke & Thaman, 1993). *Corynocarpus cribbianus* is a native fruit tree of Manus Island, the largest of the Admiralty Islands in northern Papua New Guinea, it also grows on some small islands near Madang, on the northern coast of mainland Papua New Guinea (Plant Genetic Resource Center, 1996a). The tree is common and widespread and produces edible fruit all year round, which can be eaten raw or cooked. The fruit of *C. cribbianus* is known as 'mundroi' in Tok Pisin[3] (French, 2006). There is no mention of the kernel being eaten, nor that it may be poisonous.

Both *C. cribbianus* and *C. similis* are listed as a foraged fruit tree species in a table of Oceanic fruit trees in (Lebot, 2008). French (1994) describes *C. cribbianus* as a very fibrous fruit and probably not suitable for export. *Corynocarpus* species were at one time probably more intensively exploited, and even though they are found growing wild throughout Melanesian lowland forests, they are rarely cultivated these days (Blench, 2004).

1.5.1 The name karaka and its cognates in the Pacific region

The name karaka and its cognates are used in the Pacific region for species other than *Corynocarpus*. Polynesian settlers to New Zealand transferred their word for one species in their homeland to one with similar morphological characteristics (Leach, 1984) or uses in their new found home.

[3] **Tok Pisin** is a creole spoken throughout Papua New Guinea.

With reference to Niūe, Smith (Smith, 1903) page 181 wrote:

> *"In the names of the trees and plants there is often an identity of name with those of the Maori, though sometimes the plants themselves differ widely. Thus Kalāka (Karaka), Maile (Maire), Pilīta, (Pirita), Tara (Tawa), Kafīka (Kahika), Mohūku (Mouku)."*

A search for trees in the Pacific region with a name similar to karaka reveals many cognates for the word. In Samoa, *Planchonella linggensis* is known as ʻAlaʼa. The name is known from all the islands of Samoa but on Aunuʼu and Apilina, which lack *P. linggensis*, the name applies to *P. grayana* (a rare tree elsewhere in Samoa) (Whistler, 1984).

Kalaka is used for *Planchonella grayana* in Tonga and Niue and probably on Rapa (karaka) and Atiu (Cook Islands) in eastern Polynesia (Whistler, 1984). According to Tupou *et al.* (2001), kalaka refers to *Planchonella costata* in Tonga and Drake (1996) in the flora of 'Eua Island, Tonga, lists kalaka as the common name for *P. garberi* and *P. grayana.* In Tonga, karaka is also an inland forest tree, *Elaeocarpus tonganus*, with tough white timber that does not warp (Buse & Taringa, 1995). In Fiji qualaka (properly written nggalaka) is the name of a tree (Christian, 1925). Best (1977) records karaka as "…… a tree-name in Mangaia island, as kalaka is at Niue, but neither seems to be allied to our New Zealand tree."

Kalaka, qualaka, nggalaka and ʻAlaʼa are all cognates of karaka and whilst these names refer to a different genus and several of its species it is likely that karaka in New Zealand was so named due to its morphological similarity to *Planchonella* species in tropical Polynesia. In fact, S. Percy Smith (Smith, 1893) writes "The Kalāka is so like the New Zealand Karaka in its habit that the one might be taken for the other at a short distance, but they are different species." In Mangaia, the tree called kalaka is not used as a food, as noted by Christian (1925):

> *"The natives do not make use of the berries, either prepared for food, or crushed in order to poison fish. As in the case of the kalaka of Niue the tree is evidently called karaka from its nuts."*

In Māori, *horehore* is a term applied to the covering of the kernel of karaka; *karaka horehore* are kernels with the mealy fruit still attached. The prepared kernels used as a food supply are called *kōpia*, while *kōpi* is another name used for the tree [on Rekohu/Chatham Islands] (Best, 1977).

The Forsters record no vernacular name for karaka, and Banks and Solander (a Swedish botanist) write it *chalacha.* This would probably have been Solander's way of writing it as an Englishman would have used k's instead of ch's for the hard sound (Hemsley, 1903).

1.6 THE 'INTRODUCTION' OF KARAKA TO NEW ZEALAND

The first settlers of New Zealand relied on local vegetation as food-plants, and did not bring any crop species with them from their homelands (Wilson, JA, 1906 as cited in Buck, 1949). Wilson stated that these early settlers did not have karaka as a food source. However, (Buck, 1949) suggests this statement is not correct "…..for karaka is indigenous New Zealand." and that Wilson had inferred this based on knowledge he had of Turi[4], captain of the Aotea waka, introducing karaka from Rangitahua, in the later Fleet period (Buck, 1949).

Smith (1891) describes how Māori came "…...fully prepared to occupy a new country bringing wives, families and several plants….. and, as some traditions say, certain birds and plants which are known to be natives of the country." In the same proceedings he supposes the island where Turi stayed, Rangi-tuhia, whilst journeying to New Zealand in the Aotea waka (voyaging canoe), could, in fact, be Sunday Island, now more commonly known as Raoul Island, part of the Kermadec Group. Because the Aotea tradition lays claim to introducing karaka to New Zealand, Percy Smith suggests karaka was collected on the island and brought to New Zealand in this way (Smith, 1893). Seven years later, Heteraka Tautahi dictated the traditions of the Aotea canoe to Percy Smith (Smith, 1900). In this account, whilst no direct reference is made to collecting karaka on Rangi-tahua intentionally, it does talk of stopping at an island called Rangi-

[4] Turi was the captain of the Aotea waka whose occupants became the ancestors of the Taranaki, Ngāti Ruanui, Ngā Rauru, and Wanganui tribes of the West Coast of New Zealand.

tahua and the Aotea waka bringing karaka with them to New Zealand. The Aotea brought karaka to Taranaki and Turi planted a karaka grove in Patea and called the place Pou-o-Turi. Buck (1949) suggests Turi was attracted to the ripe berries on karaka trees on Rangitahua which means the Aotea would have been there around February or March. Buck (1949) adds that while Turi may have brought karaka from the Kermadec Islands, he certainly did not bring them with him from Hawaiki, and his introduction merely added to karaka already growing in New Zealand.

However, in another tradition it was Kupe[5], who planted a variety of karaka called *oturu* at Patea (Matorohanga, 1995). Percy Smith's notes during the translation of Te Matorohanga's account of Kupe briefly describe the karaka called *oturu*:

> *"The karaka-oturu is described to me as like the ordinary karaka (*Corynocarpus laevigatus*), but with smaller leaves and berries and fewer of them, with a low growth. There are some trees of the same species growing at Nuhaka, Hawkes Bay, the seed of which is said to have been brought here by the Kura-haupo canoe, under Whatonga. If this karaka at Patea bore a few fruit on the west side of the tree it denoted a lean year-if on the east, or inland side, it meant a prolific year for all cultivated foods. The Rev. T. G. Hammond, who knows Patea and its history better than any man, does not recognize this tree. It is also related of Turi, who commanded the Aotea canoe, and who settled down at Patea, that he brought the karaka tree with him."*

In another account written by John Houston (1965 pp. 27), Turi made the final part of his journey from Aotea [harbour] to Patea by foot. He sent Pungarehu ahead and instructed him to plant karaka seeds [brought on the Aotea canoe] all along the route to provide a plentiful supply of food. In the same account, Turi established a grove of karaka trees at Papawhero, on the north bank of the Patea River (Houston, 1965 pp. 32).

Another oral history from the East Coast (Gisborne region) mentions an iwi called Te Whakatane, whose ancestor Tama-tea-nuku-roa was the captain of the Nukutere waka (voyaging canoe). His son, Roau, was credited with the introduction of ti (cabbage tree, *Cordyline terminalis*), taro (*Colocasia esculenta*) and karaka (Best, 1972). According to Kai tahu oral history their ancestors "brought this tree [karaka] from the North Island to Kaikoura (South Island) [where] it flourishes but very few trees are further south" Beattie

[5] Kupe – the first voyager to make contact with New Zealand from Hawaiki, the traditional Māori place of origin. He appears in many Māori oral histories.

(1994, cited in (Leach & Stowe, 2005)).

It is important to note that there are many oral histories pertaining to this period of settlement in New Zealand. However, for many iwi (tribes), for example, Ngā Rauru, there are traditions of voyaging between the Pacific Islands and New Zealand that pre-date that era. This contact with the Pacific span generations, each iwi laying claim to introducing elements of the flora and fauna important for identity and survival (Nick R. Roskruge, personal communication, July 25th 2013).

1.7 THE CULTIVATION OF KARAKA

> *"According to tradition, karaka were brought by people on voyaging canoes, distributed by people living in coastal areas, and planted on tracks as a food resource, or to identify tapu places, burial grounds or caves."* (Haami, 2004)

Māori deforested large areas of New Zealand to encourage growth of aruhe (*Pteridium esculentum*, bracken) and to provide clearings for gardens, housing areas and for planting karaka (Wilmshurst *et al.*, 2004). Karaka pollen was found in pollen cores from two sites in the Mimi and Waitoetoe catchments in north Taranaki and its sudden appearance in the sections of the cores corresponded to the deforestation period and early Māori settlement period. This suggests karaka did not grow historically in Taranaki and was probably brought to the region by Māori and planted in recently deforested clearings (Wilmshurst *et al.*, 2004). In the Mimi and Waitoetoe catchments karaka are still present in small groves today (Wilmshurst *et al.*, 2004). Platt (2003) states that karaka are found at many pa[6] sites in Taranaki and that many are large-fruited compared with natural stands in Auckland, suggesting selection by Māori for increased fruit size. Karaka pollen was also found in pollen cores from the Coromandel Peninsula by Byrami (2002), with its first appearance corresponding to the same deforestation period.

In order to authenticate rights to tribal lands for native land court proceedings, Hākaraia Maumau kept a notebook of pepeha for his local iwi as a written record

[6] Pa – naturally defensible habitation sites fortified with earthworks and/or stockades.

(Haami, 2004). In this notebook he makes reference to *taupahī* (seasonal camping grounds) which were located near food resources such as kūmara, karaka groves, aka (*Metrosideros fulgens*) vines, rat runs, eel weirs, fishing grounds, bird-snaring sights or berry-producing trees (Haami, 2004). He also makes reference in these pepeha to ngakinga karaka (karaka grounds) and mahinga karaka (karaka gardens, or harvest locations) for example:

Ko Waiaute he mahinga kai he mahinga karaka nā Aupaki
Wai-aute is a garden and a karaka preparing place that belonged to Aupaki

Ko Te Tuhi he pā karaka i a Te Pū-Hā
Te Tuhi is a clump of karaka trees belonging to Te Pū-Hā

Ko Manu-hāro he mahinga karaka nga Hika-toa
Manu-hāro is a karaka cultivation of Hika-toa

Hemsley (1903) believed karaka, both in a wild and formerly cultivated state, thrived only in the warmer parts of New Zealand and Featon & Featon (1889) regard all karaka occurrences in the South Island as the remains of cultivation. Kirk (1889) states that it is very rare in the South Island, being restricted to a few localities in the Nelson, Marlborough and Canterbury districts.

At the end of the 19th century, karaka was noted as a species that grew abundantly near the sea, forming groves, and where it grew inland probably resulted from propagation by Māori for food use (Featon & Featon, 1889). The fruit was important as a food to Māori and at this time it formed a 'staple article of subsistence' (Featon & Featon, 1889). Karaka was of particular importance as a food to Māori in regions of New Zealand where other cultivated crops, such as kūmara (*Ipomoea batatas*) and other introduced sub-tropical plant foods were not grown, for example the region between Wellington and Castlepoint along the Wairarapa coast (Best, 1977). In Palmerston North (Papaioea), in the Manawatu region, a large karaka grove, which formed part of one of several 'foodstores' for Māori along the Manawatu River, still exists today (Anon, 1988).

Buck (1949) makes reference to the presentation of baskets of preserved karaka berries by a party of Ngati Ruanui at a *tangi* (funeral). Featon & Featon (1889) mention the use of a chaplet of the leaves of karaka to adorn the heads of Māori when approaching the graves of their ancestors. It was also used medicinally to heal wounds. The leaves were placed, shiny side down, over wounds to heal them (Macdonald, 1973). If the leaf was turned upside down it had a drawing effect and this was used to treat boils (Macdonald, 1973). This was a standard approach in Māori traditional medicine) (Riley, 1994). Given the importance of karaka to Māori, its documented use and historical associations with places of settlement, karaka provides a unique opportunity to document the process of plant domestication in its incipient stages.

1.8 INCIPIENT DOMESTICATION

1.8.1 DOMESTICATION DEFINED

Domestication can be defined as an evolving mutualism between human groups and plant or animal populations (Zeder, 2006) which has selective advantages for both: humans fulfill their resource needs and crop-plants have a reproductive advantage over their wild progenitors. Human selection on the phenotype of managed or cultivated plant populations causes changes in the genotype of the population making them more useful to humans (Clement, 1999). Domestication does not occur in an instant, rather it is a 'cumulative process', the nature of which is determined by the biological species and human society involved (Zeder *et al.*, 2006). Not all domestication events take the same course. Different domesticates and different societies will follow different 'developmental trajectories' of domestication. Genetic markers permit genome-wide investigation of genetic diversity in crops and their progenitors. While it is mainly neutral or non-coding loci and organellar genomes that are the focus of much genetic research into domestication (Zeder *et al.*, 2006), in most cases of domestication novel biological forms have arisen through selection of transcription factors (Sun *et al.*, 2009).

1.8.2 GENETIC DIVERSITY

Whether selection is intended or not, genetic diversity in crop plants is expected to reduce over time (Emshwiller, 2006). When a small number of individual plants are

selected and removed from their wild habitat and placed in a new habitat the diversity in the new population is reduced as founders represent only a small amount of the genetic diversity of the wild population. This is broadly termed 'bottlenecking' and more specifically 'the founder effect' (Ladizinsky, 1985). The number of founding individuals and the duration of the bottleneck determine the characteristics of the genetic bottleneck (Emshwiller, 2006). However, crosses between wild populations and 'cultigens' during incipient domestication through the wild-weed crop complex may lessen the founder effect (Debouck, 1999). Wild relatives of crop plants can be considered to be reasonable representatives of ancestral, pre-domestication population of the crop and can be used as a reference to contrast the genetic diversity in the domesticated crop to provide evidence of genetic bottlenecking (Doebley *et al.*, 2006).

Most studies contributing to our knowledge of domestication have been carried out on annual crops (Doebley, 2004); (Matsuoka *et al.*, 2002)(maize); (Huang *et al.*, 2012) (rice); (Peleg *et al.*, 2011) (wheat); (Labate *et al.*, 2009)(tomato); (Wills, 2006) (sunflower) and very little is known about the genetic processes involved in the domestication of long-lived perennials. Genetic variation in trees is structured very differently from annuals due to their inherent biological differences, including the length of their sexual cycle, breeding system, level of genetic diversity in the wild and their ability to hybridise (Miller, 2008). In a comprehensive review of perennial domestication Miller and Gross (2011) reviewed several studies of domestication in annuals and perennials, and determined that genetic bottlenecking in annual fruit crops retains 5.5-119.5% (averaging 59.9%) of the variation at neutral loci in the wild relatives of those crops, whereas perennial crops retain, on average, 98.4% (64.8-126.9%). The bottleneck in perennials is much wider due to a combination of the number of sexual cycles since the crop plant and its wild progenitor diverged, multiple distinct ancestral populations (both geographically and genetically) and hybridisation.

The natural range of karaka is believed to be Northland, therefore, translocated populations beyond this region have no opportunity to hybridise with their wild relatives. This is an advantage for this kind of study because other crop species are often cultivated in the vicinity of their wild relatives. Hybridisation allows for gene flow between cultivated populations and their wild progenitor, which ultimately contributes

to genetic variation in the cultivated populations. This can also complicate attempts at determining if domesticates have single or multiple origins. The observed patterns in genetic diversity among sympatric species can be a result of incomplete lineage sorting or hybridisation (Miller, 2008), (Petersen *et al.*, 2012) or perhaps a lack of resolving power in the molecular markers employed in the study (Petersen *et al.*, 2012). When a plant begins its journey towards full domestication is there an extreme loss of genetic diversity early on in the process due to the selection of a small number of individuals selected from the source population? Or is it more likely that the genetic diversity present in the source population is retained initially, but then lost gradually over time?

1.8.3 Domestication model

Using Clement's model of domestication (1999), the extent to which karaka has been domesticated can be evaluated, at most, as being incipiently domesticated. According to Clement incipient domestication can be described as follows:

> *" A population that has been modified by human selection and intervention (at the very least being promoted), but whose average phenotype is still within the range of variation found in the wild population for the trait(s) subject to selection. The variance of this average is probably smaller than that of the original wild population, however, as selection has started to reduce genetic variability."*

Hence, the extent to which karaka found at cultivated sites are morphologically distinct, (and also represent only a subset of the genetic diversity of naturally distributed karaka) can be considered the extent to which karaka has been domesticated. However, garden and orchard plants often do not show characteristic morphological changes allowing them to be recognized as domesticates in the archaeological record (Leach & Stowe, 2005). Selection for non-morphological characteristics, such as sweeter-tasting or non-toxic fruit, does not necessarily alter the morphology of a particular plant. Leach (2005) states that it is not only morphological characters that give us the clues to domestication; a species' appearance outside of its natural range can also be an indicator of anthropogenic intervention.

FIGURE 1.9: Map of the Pacific region showing the location of New Zealand and its major offshore Islands (Kermadec Islands, Chatham Islands, Three Kings Islands, labelled) as well as all Pacific Islands mentioned in this thesis. Map courtesy of Andrew Clarke, University ofWarwick, http://www.clarkeresearch.org/

FIGURE 1.10*:* Map of New Zealand showing all governmental regions and places names referred to in this thesis. The black line identifies the phytogeographic boundary at 38˚S. Map courtesy of Daniel Dalet (www.d-maps.com)

1.9 REFERENCES

Anon. 1988. *The Karaka Grove*. Massey University: Palmerston North.

Anon. 1992. Les plantes de cueillette/Gathered edible forest plants. *Tam Tam* **27:** p. 2.

Bannister, P., Bibby, T., and Jameson, P. E. 1996. An investigation of recalcitrance in seeds of three native New Zealand tree species. *New Zealand Journal of Botany* **34:** 583-590.

Best, E. 1972. *Tuhoe: the children of the mist*. 2nd edition. Memoirs of the Polynesian Society. A.H. and A.W.Reed for The Polynesian Society: New Plymouth.

Best, E. 1977. *Forest lore of the Māori*. E.C. Keating: Wellington.

Blench, R. 2004. Fruits and arboriculture in the Indo-Pacific region. *Bulletin of the Indo-Pacific Prehistory Association***:** 31.

Briggs, J. D., and Leigh, J. H. 1996. *Rare or threatened Australian plants*. 4th edition. CSIRO: Melbourne.

Buck, P. 1949. *The coming of the Māori*. Maori Purposes Fund Board: Wellington.

Burrows, C. J. 1996. Germination behaviour of seeds of the New Zealand woody species *Alectryon excelsus*, *Corynocarpus laevigatus*, and *Kunzea ericoides*. *New Zealand Journal of Botany* **34:** 489-498.

Byrami, M., Ogden, J., Horrocks, M., Deng, Y., Shane, P., and Palmer, J. 2002. A palynological study of Polynesian and European effects on vegetation in Coromandel, New Zealand, showing the variability between four records from a single swamp. *Journal of the Royal Society of New Zealand* **32:** 507-531.

Cabalion, P., and Poisson, J. 1987. *Corynocarpus similis* [Hemsley], Plante alimentaire et toxique de Vanuatu (ex-Nouvelles-Hebrides). *Journal of Ethnopharmacology* **21:** 189- 191.

Campbell, J. D. 2002. Angiosperm fruit and leaf fossils from Miocene silcrete, Landslip Hill, northern Southland, New Zealand. *Journal of the Royal Society of New Zealand* **32:** 149-154.

Campbell, K. L. 2006. *A study of home ranges, movements, diet and habitat use of kereru (Hemiphaga novaeseelandiae) in the southeastern sector of Banks Peninsula, New Zealand [PhD Thesis]*. Lincoln University, Canterbury, New Zealand.

Carlquist, S., and Miller, R. B. 2001. Wood anatomy of Corynocarpaceae is consistent with Cucurbitalean placement. *Systematic Botany* **26:** 54-65.

Center, P. G. R. 1996a. *Papua New Guinea: country report for the FAO International Technical Conference on Plant Genetic Resource*. 4th International Technical Conference on Plant Genetic Resources. FAO: Leipzig.

———. 1996b. *Solomon Islands: country report for the FAO International Technical Conference on Plant Genetic Resource*. 4th International Technical Conference on Plant Genetic Resources. FAO: Leipzig.

Christian, F. W. 1925. Karaka tree of Mangaia. *Journal of the Polynesian Society* **34:** 94.

Clarke, A. 2007. *The great sacred forest of Tane: Te wao tapu nui a Tane - a natural pre-history of Aotearoa New Zealand*. 1st edition. Reed Publishing: Auckland.

Clarke, W., and Thaman, R. R. (Eds). 1993. *Agroforestry in the Pacific islands: systems for sustainability*. United Nations University Press: Tokyo.

Clement, C. 1999. 1492 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. *Economic Botany* **53:** 188-202.

Clout, M. N., and Hay, J. R. 1989. The importance of birds as browsers, pollinators and seed dispersers in New Zealand forests. *New Zealand Journal of Ecology* **12:** 27-33.

Colenso, W. 1868. Geographic and economic botany of the North Island of New Zealand. *Transactions and Proceedings of the New Zealand Institute* **1:** 232-283.

Colenso, W. 1880. On the vegetable food of the ancient New Zealanders. *Transactions of the Royal Society of New Zealand* **13** Republished by Kiwi Publishers, Christchurch, 2001.

Cook, R. A., Sutherland, R., and Zhu, H. 1999. *Cretaceous - Cenozoic geology and petroleum systems of the Great South Basin, New Zealand. Monograph 20*. Institute of Geological & Nuclear Sciences: Lower Hutt.

Costall, J., Carter, R., Shimada, Y., Anthony, D., and Rapson, G. 2006. The endemic tree *Corynocarpus laevigatus* (karaka) as a weedy invader in forest remnants of southern North Island, New Zealand. *New Zealand Journal of Botany* **44:** 5-22.

Dawson, M. I. 1997. Chromosome numbers in *Corynocarpus* (Corynocarpaceae). *New Zealand Journal of Botany* **35:** 255-258.

Debouck, D. G. 1999. Diversity in *Phaseolus* species in relation to the common bean. *In* Singh, S. P. (Ed.), *Common bean improvement in the twenty–first century*, pp. 25-52. Kluwer Academic Publishers: Dordrecht.

Dijkgraaf, A. C. 2002. *Phenology and frugivory of large-fruited species in northern New Zealand and the impacts of introduced mammals [PhD thesis]*. University of Auckland, Auckland.

Doebley, J. 2004. The genetics of maize evolution. *Annual Review of Genetics* **38:** 37-59.

Doebley, J. F., Gaut, B. S., and Smith, B. D. 2006. The molecular genetics of crop domestication. *Cell* **127:** 1309-1321.

Drake, D., Whistler, W., Motley, T., and Imada, C. 1996. Rain forest vegetation of 'Eua Island, Kingdom of Tonga. *New Zealand Journal of Botany* **34:** 65-77.

Emshwiller, E. 2006. Genetic data and plant domestication. *In* Zeder, M. A., Bradley, D. G., Emshwiller, E., and Smith, B. D. (Eds), *Documenting domestication: the intersection of genetics and archaeology*, pp. 99-122. University of California Press: Berkley, Los Angeles.

Engler, A. 1897. Corynocarpaceae. *In* Engler, A. and Prantl, K. (Eds), *Die natürlichen Pflanzenfamilien - Nachträge zum II-IV. Teil.*, pp. 215–217. Engelmann: Leipzig.

Featon, E. H., and Featon, S. 1889. *The art album of New Zealand flora*. Bock & Cousins: Wellington.

French, B. 2006. *Food crops of Papua New Guinea* Published by the Author: Sheffield, Tasmania.

French, B. R. 1994. *Some suggestions and ideas for improving food crop production in Papua New Guinea*. available online at http://foodplantsinternational.com/?sec=resources.

Garnier, B. J. 1958. *The climate of New Zealand*. Edward Arnold: London.

Garnock-Jones, P. J., Brockie, R. E., and FitzJohn, R. G. 2007. Gynodioecy, sexual dimorphism and erratic fruiting in *Corynocarpus laevigatus* (Corynocarpaceae). *Australian Journal of Botany* **55:** 803-808.

Haami, B. 2004. *Pūtea whakairo: Māori and the written word*. Huia Publishers in association with the Ministry for Culture and Heritage: Wellington, N.Z. .

Hemsley, W. B. 1903. On the genus *Corynocarpus*, Forst., with descriptions of two new species. *Annals of Botany* **17:** 743-760.

Herzer, R. H., Chaproniere, G. C. H., Edwards, A. R., Hollis, C. J., Pelletier, B., Raine, J. I., Scott, G. H., Stagpoole, V., Strong, C. P., Symonds, P., Wilson, G. J., and Zhu, H. 1997. Seismic stratigraphy and structural history of the Reinga Basin and its margins, southern Norfolk Ridge system. *New Zealand Journal of Geology and Geophysics* **40:** 425-451.

Houston, J. 1965. *Maori life in old Taranaki*. AH & AW Reed: Wellington.

Huang, X., *et al* 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490:** 5.

KDP. 2008. *Kēreru Discovery Project: Kēreru Discovery Project* [electronic source]. http://www.kererudiscovery.org.nz/diet_full.php. Last updated: 2008. Accessed: 9/11/2009.

Kirk, T. 1889. *The forest flora of New Zealand*. Government Printer: Wellington.

Labate, J. A., Robertson, L. D., and Baldo, A. M. 2009. Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). *Heredity* **103:** 257-267.

Ladizinsky, G. 1985. Founder effect in crop-plant evolution. *Economic Botany* **39:** 191-199.

Leach, H. M. 1984. *1,000 years of gardening in New Zealand*. Reed: Welington.

Leach, H. M., and Stowe, C. J. 2005. Oceanic arboriculture at the margins: the case of the karaka (*Corynocarpus laevigatus*) in Aotearoa. *Journal of the Polynesian Society* **114:** 7-27.

Lebot, V., Walter, A., Sam, C. 2008. The domestication of fruit and nut species in Vanuatu, Oceania. *In* F. K. Akinnifesi, F. K., Leakey, R. R. B., Ajayi, O. C., Sileshi, G., Tchoundjeu, Z., Matakala, P., and Kwesiga, F. R. (Eds), *Indigenous fruit trees in the tropics – domestication, utilization and commercialization* pp. 120-136. CAB International: Wallingford, UK.

Lord, J. M., Markey, A. M., and Marshall, J. 2002. Have frugivores influenced the evolution of fruit traits in New Zealand? *In* Levey, D. J., Silva, W. J., and Galetti, M. (Eds), *Seed dispersal and frugivory: ecology, evolution, and conservation*, pp. 55–68. CAB International: Wallingford.

MacArthur, R., and Wilson, E. 1967. *The theory of island biogeography* Princeton University Press.: Princeton, N.J.

Macdonald, C. 1973. *Medicines of the Māori: from their trees, shrubs and other plants, together with foods from the same sources*. William Collins: Auckland.

Matorohanga, T. 1995. Kupe. *In* Henry, T. (Ed.), *Voyaging chiefs of Havai'i*. Kalamaku Press: Honolulu.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez, G. J., Buckler, E., and Doebley, J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 6080-6084.

Matthews, M. L., and Endress, P. K. 2004. Comparative floral structure and systematics in Cucurbitales (Corynocarpaceae, Coriariaceae, Tetramelaceae, Datiscaceae, Begoniaceae, Cucurbitaceae, Anisophylleaceae) *Botanical Journal of the Linnean Society* **145:** 129–185.

McGlone, M. 1985. Plant biogeography and the late Cenozoic history of New Zealand. *New Zealand Journal of Botany* **23:** 723-749.

Miller, A. 2008. Characterization of the *Spondias purpurea* lineage in Mesoamerica based on nuclear and chloroplast sequence data. *Journal of the Torrey Botanical Society*. **135:** 463-474.

Miller, A., and Gross, B. 2011. From forest to field: Perennial fruit crop domestication. *American Journal of Botany* **98** 1389-1414.

Molloy, B. P. J. 1990. The origin, relationships, and use of karaka or kopi (*Corynocarpus laevigatus*). *In* Kapoor, W. H. a. P. (Ed.), *Nga Mahi Maori o te Wao Nui a Tane: contributions to an International Workshop on Ethnobotany, Te Rehua Marae*, pp. 48–53. Botany Division, Department of Scientific and Industrial Research: Christchurch, New Zealand.

Ogden, J. 1989. On the coenospecies concept and tree migrations during the oscillations of the Pleistocene climate. *Journal of the Royal Society of New Zealand* **19:** 249-262.

Palmer-Jones, T., and Line, L. J. S. 1962. Poisoning of honeybees by nectar from the karaka tree (*Corynocarpus laevigata* J. R. et G. Forst). *New Zealand Journal of Agricultural Research* **5:** 433-436.

Parton, K., Bruere, A. N., and Chambers, J. P. 2001. Miscellaneous plants, shrubs and trees. *In Veterinary clinical toxicology (Publication No. 208.)*, 2nd edition. Veterinary Continuing Education, Massey University: Palmerston North.

Peleg, Z., Fahima, T., Korol, A. B., Abbo, S., and Saranga, Y. 2011. Genetic analysis of wheat domestication and evolution under domestication. *Journal of Experimental Botany* **62:** 5051-5061.

Petersen, J., Parker, I., and Potter, D. 2012. Origins and close relatives of a semi-domesticated neotropical fruit tree: *Chrysophyllum cainito* (Sapotaceae). *American Journal of Botany* **99:** 585-604

Platt, G. 2003. Observations on karaka (*Corynocarpus laevigatus*) and its fruit. *Auckland Botanical Society Journal* **58:** 29-31.

Powlesland, R. G., Dilks, P. J., Flux, I. A., and Grant, A. D. 1997. Impact of food abundance, diet and food quality on the breeding of the fruit pigeon, Parea *Hemiphaga novaeseelandiae chathamensis*, on Chatham Island, New Zealand. *Ibis* **139:** 353-365.

Riley, M. 1994. *Maori Healing and Herbal: New Zealand ethnobotanical sourcebook*. Viking Sevenseas: Paraparaumu.

Sawyer, J., McFadgen, B., and Hughes, P. 2003. Karaka (*Corynocarpus laevigatus* JR et G. Forst.) in Wellington Conservancy (excluding Chatham Islands). *DOC Science Internal Series 101*.

Shaw, S. D., and Billing, T. 2006. Karaka (*Corynocarpus laevigatus*) toxicosis in North Island brown kiwi (*Apteryx mantelli*). *Veterinary Clinics of North America: Exotic Animal Practice* **9:** 545–549.

Skey, W. 1871. Preliminary notes on the isolation of the bitter substance of the nut of the karaka tree (*Corynocarpus laevigata*). *Transactions and Proceedings of the New Zealand Institute* **4:** 316-321.

Smith, S. P. 1891. Notes on the geographical knowledge of the Māoris. *In* Hector, J. (Ed.), *Australasian Association for the Advancement of Science Conference*, Christchurch, New Zealand, Vol. 3, pp. 280-310. Wellington: George Didsbury.

———. 1893. 28 - Notes and queries. *Journal of the Polynesian Society* **2:** 125-127

———. 1900. The Aotea Canoe - The migration of Turi to Aotearoa (New Zealand). *Journal of the Polynesian Society* **9:** 211-233.

———. 1903. *Niuē-fekai (or Savage) Island and its People*. Whitcombe & Tombs: Wellington.

Stocker, G. C., and Irvine, A. K. 1983. Seed dispersal by cassowaries (*Casuarius casuarius*) in North Queensland's rainforests. *Biotropica* **15:** 170-176.

Stowe, C. J. 2003. *The ecology and ethnobotany of karaka (Corynocarpus laevigatus) [MSc. thesis]*. University of Otago, Dunedin.

Sun, L., Xing, S., Zhang, J., Yang, J., Wang, X., and Dong, Y. 2009. Function of the transcription factors in plant domestication and stress resistance. *Genomics and Applied Biology* **28:** 9.

van Steenis, C. G. G. J. 1951. Corynocarpaceae. *In Flora Malesiana Series 1: Spermatophyta*, pp. 262–264. Noordhoff–Kolff: Jakarta.

Wagstaff, S., and Dawson, M. 2000. Classification, origin, and patterns of diversification of *Corynocarpus* (Corynocarpaceae) inferred from DNA sequences. *Systematic Botany***:** 134-149.

Whistler, W. 1984. Annotated list of Samoan plant names. *Economic Botany* **38:** 464-487.

Williams, C. 1982. Nutritional properties of some fruits eaten by the possum *Trichosurus vulpecula* in a New Zealand broadleaf-podocarp forest. *New Zealand Journal of Ecology* **5:** 16-20.

Williams, P. A., Karl, B. J., Bannister, P., and Lee, W. G. 2000. Small mammals as potential seed dispersers in New Zealand. *Austral Ecology* **25:** 523-532.

Wills, D. 2006. Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.). *Journal of Heredity* **97:** 6.

Wilmshurst, J. M., Higham, T. F. G., Allen, H., and Johns, D. 2004. Early Maori settlement impacts in northern coastal Taranaki, New Zealand. *New Zealand Journal of Ecology* **28:** 167-179.

Wood, J., Wilmshurst, J., Worthy, T., and Cooper, A. 2012a. First coprolite evidence for the diet of *Anomalopteryx didiformis*, an extinct forest ratite from New Zealand. *New Zealand Journal of Ecology* **36:** 164--170.

Wood, J., Rawlence, N., Rogers, G., Austin, J., Worthy, T., and Cooper, A. 2008. Coprolite deposits reveal the diet and ecology of the extinct New Zealand megaherbivore moa (Aves, Dinornithiformes). *Quaternary Science Reviews* **27:** 2593-2602.

Wood, J., Wilmshurst, J., Wagstaff, S., Worthy, T., Rawlence, N., and Cooper, A. 2012b. High-Resolution Coproecology: Using Coprolites to Reconstruct the Habits and Habitats of New Zealand's Extinct Upland Moa (*Megalapteryx didinus*). *PLOS One* **7:** e40025.

Wotton, D., Clout, M., and Kelly, D. 2008. Seed retention times in the New Zealand pigeon (*Hemiphaga novaezeelandiae novaeseelandiae*). *New Zealand Journal of Ecology* **32:** 1-6.

Wotton, D. M., and Ladley, J. J. 2008. Fruit size preference in the New Zealand pigeon (*Hemiphaga novaeseelandiae*). *Austral Ecology* **33:** 341-347.

Zeder, M. A. 2006. Central questions in the domestication of plants and animals. *Evolutionary Anthropology* **15:** 105-117.

Zeder, M. A., Emshwiller, E., Smith, B. D., and Bradley, D. G. 2006. Documenting domestication: the intersection of genetics and archaeology. *Trends in Genetics* **22:** 139-155.

2

ORIGINS OF KARAKA IN NEW ZEALAND

2.1 CHAPTER OVERVIEW

This chapter is presented in the format of a scientific journal paper, ready for submission in New Zealand Journal of Botany. It is intended as a stand-alone chapter and for this reason some parts may overlap with sections from other chapters. It begins by briefly discussing the study species, and then introduces a review of the literature on the vegetation history of lowland species in New Zealand. This provides a framework for understanding the natural distribution of karaka before human settlement of Aotearoa/New Zealand. Analyses of chloroplast and nuclear loci of other species of *Corynocarpus* provided some insight into the relationships within the genus as a whole. The molecular systematics of karaka are discussed as well as the results of re-sequencing some already tested accessions and markers (Wagstaff & Dawson, 2000) with accessions from the Three Kings Islands, not previously sampled.

2.2 A NOTE ON ATTRIBUTION

This chapter is mostly my own work. However, the work was undertaken with assistance from Trish McLenachan, the Laboratory Manager for the PLEB Laboratory in the Institute of Fundamental Sciences at Massey University in Palmerston North. Trish carried out some of the *WAXY*, *rbcL* and *trnL-trnF* PCR, prepared them for sequencing, and edited the sequences.

2.3 ABSTRACT

This chapter reports genetic analyses of nuclear and chloroplast markers used to test hypotheses of the inter- and intraspecific relationships of karaka in New Zealand. A previous study used nuclear ITS and chloroplast *rbc*L DNA sequences to reconstruct phylogenetic relationships for the genus *Corynocarpus*. The results described here extend the taxon sampling for karaka ITS sequences, and complement these with results for a low copy number nuclear DNA marker *WAXY*. The previously published discrepancy in findings from *rbc*L and ITS analyses suggested conflict in the phylogenetic information at these two loci. This was further investigated by re-sequencing and determining additional *rbc*L sequences, as well as characterisation of chloroplast *trnL-trnF* sequences. Our results show a clearer picture of the relationships between the species with the use of additional nuclear and chloroplast markers, previously untested in *Corynocarpus*. The results indicate karaka was already part of the flora of these islands long before the human settlement of New Zealand.

2.4 INTRODUCTION

2.4.1 CORYNOCARPACEAE

The Corynocarpaceae family consists of one genus and five species found in tropical to warm temperate areas in the southwest Pacific. *Corynocarpus similis* is found in Vanuatu, the Solomon Islands, New Britain, New Ireland, and the Bismarck Archipelago; *C. cribbianus* is found on the island of New Guinea (French, 2006) and northeastern Queensland (van Steenis, 1951). *C. rupestris* occurs in isolated locations in Australia and has two subspecies. *C. rupestris* subsp. *rupestris*; also known as Glenugie Karaka, it occurs in the Clarence Valley near Coffs Harbour, near Grafton and in the Tenterfield area of New South Wales and is listed as vulnerable (Briggs & Leigh, 1996). *C. rupestris* subsp. *arborescens* occurs in southeast Queensland (Guymer, 1984 cited in Wagstaff & Dawson, 2000). *C. dissimilis* is endemic to New Caledonia (Hemsley, 1903). *C. laevigatus* is confined to mainland New Zealand (Aotearoa) and its offshore islands, Rekohu/Chatham Islands (hereafter Chatham Islands) and the Kermadec Islands (Molloy, 1990) (Figure 2.1). Based on ITS and *rbcL* DNA sequence analyses, Wagstaff

and Dawson (2000) suggested a palaeotropical center of origin for the Corynocarpaceae followed by two independent radiations into cooler climates with the first comprising *C. cribbianus* and *C. rupestris* extending through New Guinea to central Australia, and the second comprising *C. similis*, *C. dissimilis* and *C. laevigatus* reaching New Zealand, through New Caledonia several million years ago.

2.4.2 Vegetation history of lowland species in New Zealand

There are no major current geographical barriers to the spread of species, in general, in New Zealand (McGlone, 1985; McGlone *et al.*, 1993; McGlone *et al.*, 2001). However, climate appears to have played a significant role in determining which species are present (Lee *et al.*, 2001) and their distribution in the two islands. Cockayne (1928) and Wardle (1963) both agree there is a significant phytogeographic boundary at approximately 38-39°S (corresponding to Waipiro Bay in the Gisborne region to Kawhia Harbour in the Waikato region, see Figure 1.9 in Chapter 1). Several New Zealand plant species are distributed across the North Island to a southern limit of approximately 38˚S, for example, *Metrosideros excelsus*, *Litsaea calicaris*, *Beilschmiedia tarairi* and *Agathis australis* (Eagle, 2006). This boundary is situated where the warmer climate of the northern region meets the cooler climate of the southern region (Garnier, 1958).

During the last glacial maximum (LGM) it has been suggested that many plant species were restricted to refugia[1] in the northern North Island, northern South Island and southern South Island (Wardle, 1963). These regions contain high levels of endemism, a characteristic of glacial refugia (Petit *et al.*, 2003). In the ecological zone north of 39˚S latitude, the total number of endemics is 125 with endemics making up 5.7% of the total flora of that region; north of 38˚S the figures are 95 and 11% respectively (Wardle, 1963). Of these endemics, a large proportion is woody plants and tall trees. Northland has been a tectonically stable region of New Zealand retaining a diverse flora and has acted as a refuge for some components of the flora during the Pleistocene and major refugia in this region occurred north of latitude 38-39˚S (McGlone, 1985). Near-

[1] Glacial refugium – a place where a population may be restricted to during glacial events. They often contain high levels of genetic diversity

complete conifer-broadleaf forest persisted in the far north of Northland during the last glacial maximum (from ca. 29 ka[2] to ca. 19 ka) when much of the forest cover south towards Auckland was more open and dominated by *Nothofagus* and shrubby genera (Newnham *et al.*, in press). Pollen profiles suggest the rest of New Zealand was dominated by scrubland, with discrete pockets of woodland and woody shrubs, indicating several woody species survived the LGM *in situ* (Newnham *et al.*, in press). Recent evidence of beetle fossils from a site in Westland, in New Zealand's South Island, suggest this area was vegetated by a closed-canopy woodland, which is in direct contrast to palynological[3] interpretation for the same area (Burge & Shulmeister, 2007). This led to the suggestion that perhaps much more of New Zealand south of Northland was forested (Burge & Shulmeister, 2007). Newnham *et al.* (in press) suggest the difference is simply semantics, and perhaps the woodland Burge and Shulmeister (2007) describe is woody shrub and small woody trees, rather than typical tall forest (Newnham *et al.*, in press).

A phylogeographic study of five species of *Metrosideros* in New Zealand (Gardner *et al.*, 2004) using chloroplast markers, suggests the genus exhibits a 'classic' glacial refugia pattern, with levels of genetic diversity higher in the postulated glacial refugia areas of Wardle (1963). In this case Northland and the Nelson region in the South Island, where the climate was warmer have higher levels of endemism. Similarly, *Veronica speciosa*, now a threatened species, historically occurred from Scots Point in Northland, to Urenui in Taranaki, though its current distribution is much smaller (Armstrong & De Lange, 2005). However, southern populations are hypothesised to be more recent in origin than northern populations (Armstrong & De Lange, 2005), suggesting a possible contraction to the refuge of the northern North Island during the LGM and expansion south from there at the end of this period.

In contrast to these examples of survival in glacial refugia, genetic diversity of *Asplenium hookerianum*, a fern associated with lowland forests, appears to indicate the species survived *in situ* through the LGM (Shepherd *et al.*, 2007). Support for this claim comes from multiple widely-dispersed populations with endemic haplotypes, in regions other than those postulated to be glacial refugia (Shepherd *et al.*, 2007). Similarly,

[2] ka – thousand years ago
[3] Palynology – the study of pollen grains, especially those found in archaeological or glacial deposits

Pseudopanax ferox, another predominantly lowland tree species, occurring in both islands, also appears to have persisted during the LGM *in situ* (Shepherd & Perrie, 2011). Nuclear microsatellite data from *P. ferox* detected four distinct genetic clusters (Northland, Auckland and Moawhango; Rimutaka; Durville, Takaka and Wairoa Valley; rest of South Island), each containing private alleles. Evidence from beetle fossil assemblages near Westport on the South Island suggests the traditional view, based on pollen diagrams, that shrub and grasslands dominated the South Island, is not robust (Labate *et al.*, 2009). The beetle evidence challenged the interpretation of LGM flora based on pollen diagrams and suggested closed canopy woodlands could have been more prolific during the LGM. However, McGlone *et al.* (2005) argue such a clear existence of woodland could not have gone un-noticed in the pollen records and suggests the beetle evidence supports, rather than challenges, his and his colleagues' hypothesis of the widespread survival of small and patchy wooded areas.

2.4.3 WHAT WAS THE PRE-HUMAN DISTRIBUTION OF KARAKA IN NEW ZEALAND, BASED UPON WHAT IS KNOWN OF OTHER LOWLAND SPECIES?

Uplifting of the southern end of the Norfolk Ridge during the Oligocene extended the New Caledonia landmass to 32˚S and land connections via island chains (Herzer *et al.*, 1997) could have facilitated dispersal of the ancestors of karaka into New Zealand from New Caledonia (Stowe, 2003). Fossilised karaka-like kernels were discovered at Landslip Hill in Southland, New Zealand, dating to the early Miocene (24 mya[4]) (Campbell, 2002) which may indicate the arrival in New Zealand in the mid-Tertiary. However, no definite identification of these fossils was made beyond the possible genus level. Macrofossil remains of the other species (*Avicennia*, *Pomaderris*, and *Pouteria*) found at Landslip Hill indicate a deltaic-coastal ecosystem similar in nature to the vegetation of modern northern New Zealand and New Caledonia (Campbell, 2002). In the late Oligocene-Early Miocene, the area around Gore, Southland, would have been at a latitude of more than 50˚S (Cook *et al.*, 1999). It is unlikely that these plants would survive a similar modern day latitude, suggesting global temperatures were warmer in the Mid-Cenozoic (Campbell, 2002). The mid-Pliocene saw a gradual reduction in the number of taxa of tropical and subtropical affinities in the northern South Island and by

[4] Mya – million years ago

the Pleistocene most of these taxa had disappeared from the flora (McGlone, 1985; Lee *et al.*, 2001).

Species that are present in the pollen record do not necessarily correspond to the diversity of the actual composition of the historical flora. Wind-pollinated (anemophilous) plants can produce somewhere in the region of between 10,000 and 70,000 grains of pollen per anther, resulting in their dominance of the pollen fossil record (Olsen, 2004). Animal-pollinated and insect-pollinated (zoophilous and entomphilous) plants produce only about 1000 grains per anther and are often contribute a minor component of the pollen fossil record because of their numbers, but also because they are often covered in oils and remain stuck to the anther until picked up by an animal or insect (Olsen, 2004). Karaka is an entomophilous tree and its pollen is severely under-represented in the palynological record (Dodson, 1976). Mildenhall (1994) suggests karaka was either a recent introduction to Chatham Islands, or karaka pollen simply does not preserve well. Trees with fragile pollen that is easily degraded will always be under-represented or even missing entirely from the fossil record (Hicks, 2006). Holt (2009) noted that pollen of karaka was not recorded from any sampling site on Chatham Islands during their palynological study, even though karaka is now a major part of lowland broadleaf woodlands on the island group. The most suitable sites for pollen studies are often in peatlands, lake beds and basins where sediments have been accumulating for much longer (Macphail & McQueen, 1983). Firstly, these sites may not be areas where karaka was naturally found. Secondly, the detection of pollen depends upon the proximity of the source plants to the site, or to a water source for transport (Macphail & McQueen, 1983).

Karaka pollen was found in pollen cores at two sites in the Mimi and Waitoetoe catchments in Taranaki (Wilmshurst *et al.*, 2004). Its sudden appearance in the sections of the cores corresponding to the deforestation period and early Māori settlement period suggests karaka did not grow historically in Taranaki, and was probably brought to the region by Māori and planted in recently deforested clearings (Wilmshurst *et al.*, 2004). In the Mimi and Waitoetoe catchments, karaka are still present in small groves today (Wilmshurst *et al.*, 2004). Karaka pollen was also found in the Coromandel by Byrami (2002), corresponding to the same deforestation period previously mentioned.

Karaka's association with the species found at Landslip Hill (*Avicennia*, *Pomaderris*, and *Pouteria*) (Campbell, 2002), which are now confined to Northland, suggests that the range of karaka prior to human arrival in New Zealand was probably also Northland (Stowe, 2003). Stowe (2003) used a combination of climate profiling, and the association of karaka with archaeological sites, to uncover the extent to which the current distribution of karaka is determined, 1) by the environment, and 2) by human-mediated dispersal. His comprehensive study grouped karaka accessions into two types: cultural and unknown. Cultural karaka were those strongly associated with archaeological sites such as pa, middens, kumara pits, terraces and walls, and found growing (or recorded as growing at the time of the archaeological study) within 500 m of a registered archaeological site. Of 805 records of the occurrence of karaka, 82% were classed as cultural and the remainder was unknown. In Northland, the putative natural range of karaka, cultural and non-cultural karaka occurred with the same frequency, whereas elsewhere in the country, the cultural trees far outnumbered those classed as unknown, with the largest difference being in Auckland (~250:25). This adds weight to the suggestion that Northland could be the natural range for karaka, although Stowe (2003) suggests it could be as far south as Taranaki and Wanganui, and as far east as the Coromandel, due to the number of unknown karaka occurring across this region. This correlates with the postulated 38-39˚S phytogeographic boundary of Wardle (1985). Stowe (2003) attributes the high level of spatial association between karaka and archaeological sites as an indicator of settlement and the cultivation of food.

Climate profiling was also used to determine the natural and translocated range of the species (Stowe, 2003). There were significant differences in the climate profiles of cultural and unknown karaka accessions with the climate profile of cultural accessions being similar to that of kumara, and the climate profile of unknown accessions comparable to other broadleaved trees of tropical affinities currently restricted to the northern North Island (eg. *Litsaea calicaris*, *Weinmannia silvicola,* and *Beilschmiedia tarairi*). Stowe (2003) concluded that prior to the arrival of humans in New Zealand, karaka probably occurred from the mid to northern North Island and since human settlement has been translocated to all other regions where it now occurs.

Karaka is a climax broadleaf forest species often found growing with puriri (*Vitex lucens*), taraire (*Beilschmiedia tarairi*) and kohekohe (*Dysoxylum spectabile*). Platt

(2003) considers it reasonable to assume that where these four trees co-exist karaka trees are natural components of the surrounding flora (though this is not necessarily supported by our own observations). Today, karaka grows mainly in coastal regions from Cape Reinga to Banks Peninsula although populations do occur inland, particularly in the North Island (see Figure 1).

2.4.4 MOLECULAR SYSTEMATICS OF KARAKA

Wagstaff and Dawson (2000) reported the first molecular systematic study of karaka and closely related species, specifically undertaking phylogenetic analysis of nuclear DNA (nrDNA) marker ITS (internal transcribed spacer) and chloroplast marker *rbcL*. Only ITS sequences provided phylogenetic resolution between karaka and other species within the genus *Corynocarpus*. Both chloroplast DNA (cpDNA) markers such as *rbcL* and nuclear ITS have proven useful for interspecific phylogenetic reconstruction of several plant genera (Wagstaff & Garnock-Jones, 1998; Mitchell & Heenan, 2000; Stöckler *et al.*, 2002; Hörandl *et al.*, 2005; Knapp *et al.*, 2007). However, they are often at their limit for phylogenetic resolution in some genera, as suggested by the analyses on *Corynocarpus* reported by Wagstaff and Dawson (2000).

An alternative to the ITS region are low-copy nuclear genes such as the granule-bound starch synthase gene *WAXY*, arginine decarboxylase gene (*Adc*) and coenzyme A ligase (4CL) (Sang, 2002). In this chapter, we tested the phylogenetic origins of karaka further with analyses that included additional accessions of karaka for ITS and *rbc*L (including the Three Kings Islands (hereafter The Three Kings) as well as sequences determined for the chloroplast *trnL-trnF* region and *WAXY*. The *trnL-trnF* region was chosen because previous experience of others has suggested it's wide application in plant systematics (Shaw *et al.*, 2005).

A second aim of this chapter was to sequence the same chloroplast and nuclear loci in karaka in order to test their utility for examining dispersal hypotheses for the distribution of the species in New Zealand.

2.5 METHODS

2.5.1 SAMPLE COLLECTION

One accession from each of the five species and subspecies of *Corynocarpus* endemic to regions outside New Zealand (*C. cribbianus, C. dissimilis, C. similis, C.rupestris ssp rupestris and C. rupestris ssp arborescens*) was available from Wagstaff and Dawson (2000)*:* Forty-two accessions of *Corynocarpus laevigatus* were sampled for sequence analysis of the ITS region. Ten of these 42 accessions were used for sequence analysis of the nuclear gene *WAXY* that encodes a granule-bound starch synthase; the chloroplast gene ribulose 1, 5-bisphosphate carboxylase/oxygenase (*rbcL*); and the chloroplast intergenic region of *trnL-trnF* (the *trnL* intron, *trnL* 3′- exon and *trnL-trnF* intergenic spacer). These accessions encompass the known range of *C. laevigatus* in New Zealand. Herbarium samples and DNA leaf samples were collected in the field, except for a *C. laevigatus* accession (1162), which was a seed-propagated tree sampled in cultivation. DNA was obtained from silica-dried or fresh leaf tissue. The remainder of each branchlet was kept as a voucher specimen, with representatives accessioned into the Wellington (WELT) and Auckland (AKL) herbaria.

2.5.2 DNA EXTRACTION, POLYMERASE CHAIN REACTION AMPLIFICATION AND SEQUENCING

Fresh or silica-dried leaf tissue from *C. laevigatus* samples was snap-frozen in liquid nitrogen and powdered using a disposable grinder or milled from silica-dried leaves using a MagnaLyser with 2 mm zirconia beads (Biospec, Bartlesville, USA). Total genomic DNA was extracted from leaf tissue using a modified CTAB protocol (Doyle & Doyle, 1987) and re-suspended in de-ionized H_2O.

To characterise polymorphisms in the cpDNA of the genus *Corynocarpus*, parts of the cpDNA genome were amplified using conserved primers. DNA sequences for the plastid locus *rbcL*, 1324 bp in length and corresponding to position 59919-60842 of the karaka chloroplast genome, were determined for each of the species in using the primers rbcLAsF1 and rbcLAsR1 (Hasebe *et al.*, 1994). A section of the plastid locus *trnL-trnF*, 1019 bp in length, corresponding to position 51410-52428 of the karaka chloroplast

genome, was amplified using the primers TabC and TabF (Taberlet *et al.*, 1991), and analysed for sequence variation. Sequences of nuclear genes were also analysed, including *WAXY*, using the primers *WAXY* 10F and *WAXY* 13R (Olmstead, unpublished), and ITS using primers ITS5 (White *et al.*, 1990) and ITS28cc (Wagstaff & Garnock-Jones, 1998),

Polymerase chain reactions (PCRs) of chloroplast loci were performed using the protocols outlined in Shaw *et al.* (2005) on a T3 Thermocycler (Biometra) in 10 µl reactions containing ~50 ng DNA, 2.5 mM MgCl2, 0.5 U Red Hot Taq DNA polymerase (Thermo Fisher Scientific), 2.5 µl 10× Reaction buffer, 200 µM of each dNTP and 0.5mM of both forward and reverse primer. PCR were carried out using the following protocol: template denaturation at 94°C for 3 min followed by 35 cycles of denaturation at 94°C for 30 s, primer annealing at 50°C for 30 s, and primer extension at 68°C for 45 s per kb of sequence; followed by a final extension step of 10 min at 68°C. Amplification products were purified by digestion with 0.2 U shrimp alkaline phosphatase (SAP, USB Corp.) and 1 U exonuclease I (ExoI, USB Corp.) at 37°C for 30 min, followed by inactivation of the enzymes at 80°C for 15 min.

Sequencing was performed in both directions with the ABI Big Dye™ Terminator Version 3.1 Ready Reaction Cycle Sequencing kit in a Biometra thermal cycler following the manufacturer's protocol. Unincorporated fluorescent dNTPs were removed using CleanSEQ (Agencourt) following the manufacturer's protocol, and capillary separation was subsequently undertaken at the Massey Genome Service, Palmerston North. Sequences were edited and aligned using Sequencher 4.9 (GeneCodes Corporation).

FIGURE 2.1 : *Corynocarpus laevigatus* distribution showing (A) the distribution of karaka across New Zealand including the Three Kings Islands off the northern tip of New Zealand's North Island, the Kermadec Islands to the north and the Chatham Islands to the west; (B) the distribution of karaka in the northern South Island; and (C) North Island of New Zealand showing locations of inland karaka populations. All three maps are based on records in the AK, CHR, NZFRI, and WELT herbaria. Abbreviations follow Holmgren *et al.* (1990).

2.5.3 DATA ANALYSIS

Sequences were trimmed at the 3' and 5' ends to remove ambiguous sequence, primers sequences and the ends of sequences that extend beyond the assembled reference sequence. ITS, *WAXY*, *rbcL* and *trnL-trn-F* sequences were trimmed to 624 bp, 449 bp, 1140 bp and 854 bp respectively. The sequences were aligned using MUSCLE (www.ebi.ac.uk/Tools/msa/muscle) and converted to a nexus file using CLUSTALX (Larkin *et al.*, 2007). Indels and ambiguous bases were removed using PAUP* V4.0 (Swofford, 2003), resulting in a 555 bp sequence for further analysis.

2.5.4 DATING ITS SEQUENCE DIVERGENCE BETWEEN THE THREE KINGS AND MAINLAND KARAKA

Distance analyses were performed on the ITS dataset. The ITS nexus alignment of 555 bp was manually converted to a PHYLIP file. PHYML 3.0 (Guindon *et al.*, 2010) was then used to reconstruct a maximum-likelihood tree of *Corynocarpus* accessions assuming a Jukes-Cantor invariable sites model (Steel *et al.*, 2000), with the proportion of variable sites estimated from the data. This model was chosen as it is the simplest and better suited when there is very little genetic diversity between sequences. The Jukes-Cantor model assumes all sites can vary and when unvaried sites are present in two sequences it will underestimate the amount of change that has occurred at variable sites. SplitsTree4 (Huson & Bryant, 2006) was used to calculate the number of substitutions per site between the Three Kings and mainland New Zealand karaka.

Kay *et al.* (2006) studied rates of substitution for ITS sequences in plants. For woody perennials, the substitution rate varied between 0.38×10^{-9} – 7.83×10^{-9} substitutions per site, per year. The divergence time between the Three Kings and New Zealand karaka was calculated from evolutionary distance and mutation rate.

$$\delta = \mu \times \tau$$

where ∂ = the patristic distance in the PHYML maximum likelihood tree between the Three Kings and mainland New Zealand; μ = rate; t = time.

2.6 RESULTS

2.6.1 ITS SEQUENCES

ITS sequences were determined for all *Corynocarpus* species from Wagstaff and Dawson (2000) and a further 42 accessions of karaka, including accessions from the Three Kings, the Kermadec Islands and the Chatham Islands. Fifty variable sites were present in the alignment of 42 sequences (Table 2.1). Figure 2.2 shows a NEIGHBORNET splits graph (Huson & Bryant, 2006) indicating the inferred relationships. The graph is largely tree-like, with a small amount of internal reticulation. It shows that for *Corynocarpus* the greatest diversity occurs between recognised species. Midpoint rooting (not shown) indicates that *C. laevigatus* is derived form the ancestor of extant species currently found in New Caledonia and Vanuatu (C. *dissimilis* in New Caledonia and *C. similis* in Vanuatu). Within the species *C. laevigatus* (karaka), the greatest divergence is between the Three Kings Islands accessions and mainland New Zealand/Chatham/Kermadec Island accessions.

2.6.1.1 DATING ITS SEQUENCE DIVERGENCE BETWEEN THREE KINGS AND MAINLAND KARAKA

From the alignment data, PHYML (Guindon *et al.*, 2010) computed the estimated number of variable sites in the ITS region of karaka to be 0.35. SplitsTree4 (Huson & Bryant, 2006) was used to calculate the number of substitutions per site, which was 0.0086.

$$\frac{0.0086}{7.38\times10^{-9}} = 11.65mya \qquad \frac{0.0086}{0.38\times10^{-9}} = 22.63mya$$

Table 2.1: Fifty variable sites defining *Corynocarpus ITS* haplotypes, with their sequence alignment position indicated

Species	Nucleotide position																												
	18	51	56	87	132	144	145	158	194	195	198	344	353	357	358	361	381	400	441	445	446	469	472	475	516	518	521	526
C. rup. ssp. arborescens	A	C	A	G	T	C	A	C	A	A	C	G	G	G	C	A	A	C	C	A	G	G	G	A	G	G	T	C
C. rup. ssp. rupestris	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C. dissimilis	.	.	.	.	C	G	C	T	C	T	.	.	.	C	.	G	T	.	.	G	A	A	.	T	.	C	.	.
C. similis	.	.	G	.	.	.	.	.	.	.	G	A	.	.	.	G	T	T	.	G	A	.	T	.	.	.	C	.
C. cribbianus	.	.	.	.	.	.	.	.	.	.	G	.	T	.	.	G	.	.	.	.	.	.	.	.	C	.	.	A
C. laevigatus	.	.	.	A	.	T	.	.	.	.	G	.	.	.	T	G	T	.	A	G	A	.	.	T	.	.	.	.
C. laevigatus	T	A	.	A	.	.	.	.	.	.	G	.	.	.	T	G	T	.	A	G	A	.	.	T	.	.	.	.

2.6.2 *WAXY* SEQUENCES

Sequences for the *WAXY* locus were determined for five accessions of karaka including representatives from the Three Kings (RA154), mainland New Zealand (RA84 and RA517), Chatham Islands (RA83) and Kermadec Islands (RA117) and all other species of Corynocarpus. Twenty four variable sites were present in the alignment of nine sequences (Table 2.2).

Table 2.2: Twenty four variable sites defining *Corynocarpus WAXY* haplotypes, with their sequence alignment position indicated.

Species	*Nucleotide position*								
	33	38	60	92	164	211	283	292	374
C. laevigatus (Three Kings)	C	T	C	G	T	G	T	G	T
C. laevigatus	.	.	G	.	.	.	.	.	.
C. dissimilis	G	.	G	.	.	T	G	.	.
C. rupestris ssp. arboreus	.	C	G	A	.	T	G	A	.
C. rupestris ssp. rupestris	.	C	G	A	.	T	G	A	.
C. similis	.	C	G	A	C	T	G	.	C

NEIGHBORNET analyses using SplitsTree4 (Huson & Bryant, 2006) indicates that the Three Kings Islands karaka is genetically distinct from karaka from mainland New Zealand, Chatham and Kermadec Islands Islands (Figure 2.3).

In this case, the mainland New Zealand, Chatham Islands and Kermadec Islands haplotype appears ancestral to the type found in Three Kings Islands. The relationships inferred between karaka and the other Pacific species were very similar for both the *WAXY* and ITS loci. For *WAXY*, as with ITS, assuming a mid point root, the closest relative of karaka is *C. dissimilis*, and the greatest genetic diversity is between *Corynocarpus* species.

2.6.3 *RBCL* SEQUENCES

Sequences for the *rbcL* locus were determined for nine accessions of karaka including representatives from the Three Kings (RA154), mainland New Zealand (RA82, RA84, 5002, RA517), Chatham Islands (RA83) and Kermadec Islands (RA117, 1162, 96.160 (Wagstaff & Dawson, 2000)) and all other species of *Corynocarpus*. Twenty-eight variable sites were present in the alignment of 21 sequences (Table 2.3).

Table 2.3: Twenty eight variable sites defining *Corynocarpus rbcL* haplotypes, with their sequence alignment position indicated.

Species	*Nucleotide position*									
	307	**555**	**635**	**666**	**667**	**718**	**809**	**832**	**840**	**1101**
Corynocarpus laevigatus	C	C	C	G	T	G	A	C	A	A
Corynocarpus rup. ssp. rupestris	.	.	G	.	C	.	G	G	C	T
Corynocarpus cribbianus	.	.	G	.	C	.	G	G	C	T
Corynocarpus rup. ssp arboreus	.	.	G	.	C	.	G	G	C	T
Corynocarpus dissimilis	A	.	G	A	C	.	.	.	.	T
Corynocarpus similis	.	A	.	.	C	C	G	.	.	T

A NEIGHBORNET split graph that includes all available *rbcL* sequences is shown in Figure 2.4. This includes those determined by Wagstaff and Dawson (2000) as well as earlier sequences by Martin and Dowd (1994) and Savolainen *et al.* (1994). Excluding the sequences by Martin and Dowd (1994) and Savolainen *et al.* (1994) greatly simplified the splits graph (Figure 2.5), suggesting that the substitution pattern in these sequences were anomalous with respect to one another. As reticulations still existed after their removal, the sequences previously sequenced by Wagstaff and Dawson (2000) were re-sequenced, which confirmed they were correct. Thus the reticulation present in Figure 2.4 cannot be easily explained as a sequencing artifact. Table 2.5 shows the nucleotide

sites in the data that are incompatible and which lead to the reticulations shown in Figure 2.4. As discussed, a multiple substitution, possibly at site 814, leads to the reticulate splits graph. This occurrence explains why ITS and *rbcL* trees reconstructed by Wagstaff and Dawson (2000) differed. No intraspecific variation was observed within *C. laevigatus* for *rbcL*.

2.6.4 *TRNL-TRNF* SEQUENCES

To test inferences from the nuclear and *rbcL* markers, a cross section of accessions were sequenced for the *trnL – trnF* region of the chloroplast genome. Sequences for the *trnL-trnF* locus were determined for one accessions of karaka (96.160 (Wagstaff & Dawson, 2000)) and all other species of *Corynocarpus*. Twelve variable sites were present in the alignment of nine sequences (Table 2.4).

Table 2.4: Twelve variable sites defining *Corynocarpus trnL-trnF* haplotypes, with their sequence alignment position indicated

Species	*Nucleotide position*				
	20	**439**	**500**	**612**	**646**
C. laevigatus	T	A	A	C	A
C. rupestris ssp. arboreus	A	.	.	.	T
C. rupestris ssp. rupestris	A	T	.	.	T
C. cribbinanus	A	.	.	.	T
C. similis	A	.	T	A	.
C. dissimilis	A	.	.	.	T

Figure 2.6 shows a tree-like NEIGHBORNET splits graph. Within this graph there is no intraspecific variation among *C. laevigatus* and with this marker the most genetically similar species to *Corynocarpus laevigatus* is *Corynocarpus similis*.

NEIGHBORNET splits graphs (Huson & Bryant, 2006) for these loci appear on the following pages.

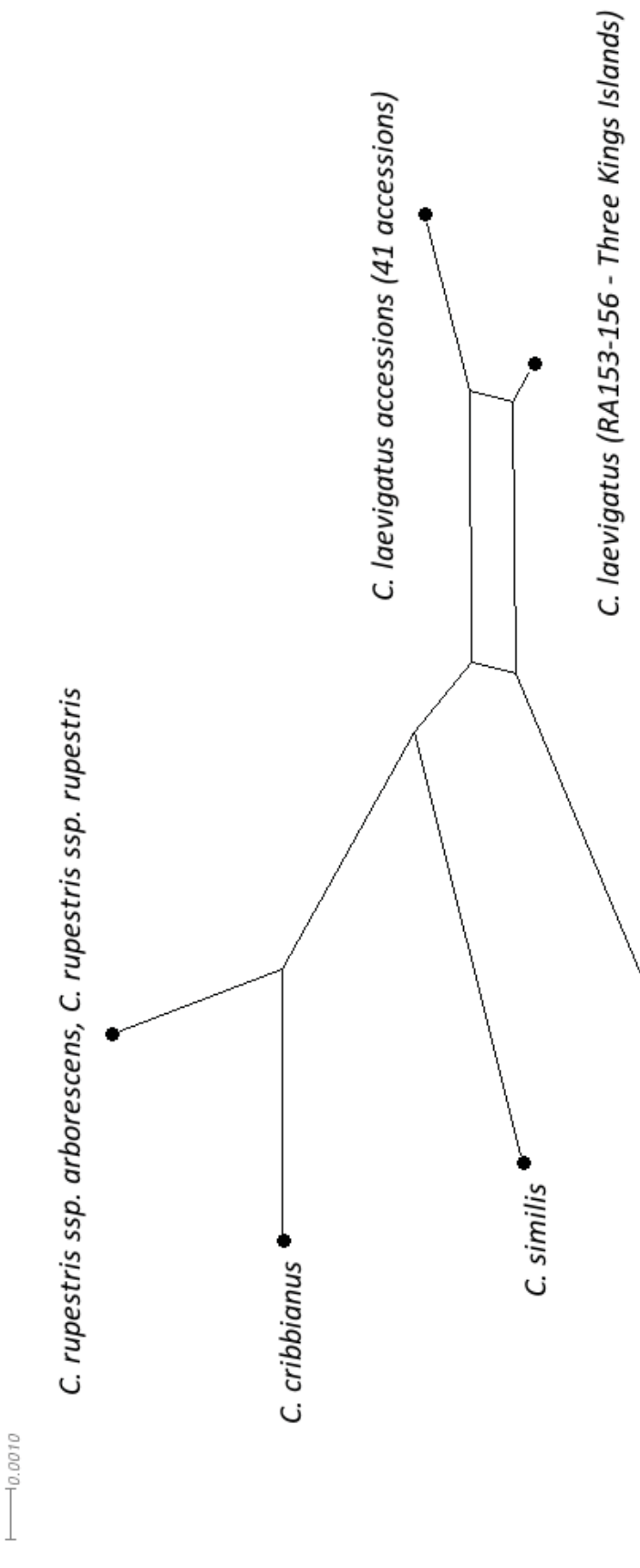

FIGURE 2.2: NEIGHBORNET splits graph computed using SplitsTree4 (Huson & Bryant, 2006) of aligned *ITS* DNA sequences from all *Corynocarpus* species. *C. laevigatus* accessions - RA04, RA66, RA82, RA83, RA84, RA103, RA117, RA118, RA141, RA180, RA186, RA209, RA212, RA218, RA305, RA320, RA340, RA368, RA440, RA436, RA464, RA469, RA473, RA563, RA510, RA517, 1015, 1109, 1162, 1296, 1311, 1313, 1321, 1345, 1445, 1981, 4650, 5002, 5314, 5726 and *C. laevigatus* (GenBank numbers AF149004.1, AF149001.1, AF149008.1, AF149007.1 and AF149002.1). Most support links *C. laevigatus* accessions including that of the Three Kings. There is a small amount of support in the data linking the Three Kings accessions and *Corynocarpus dissimilis*.

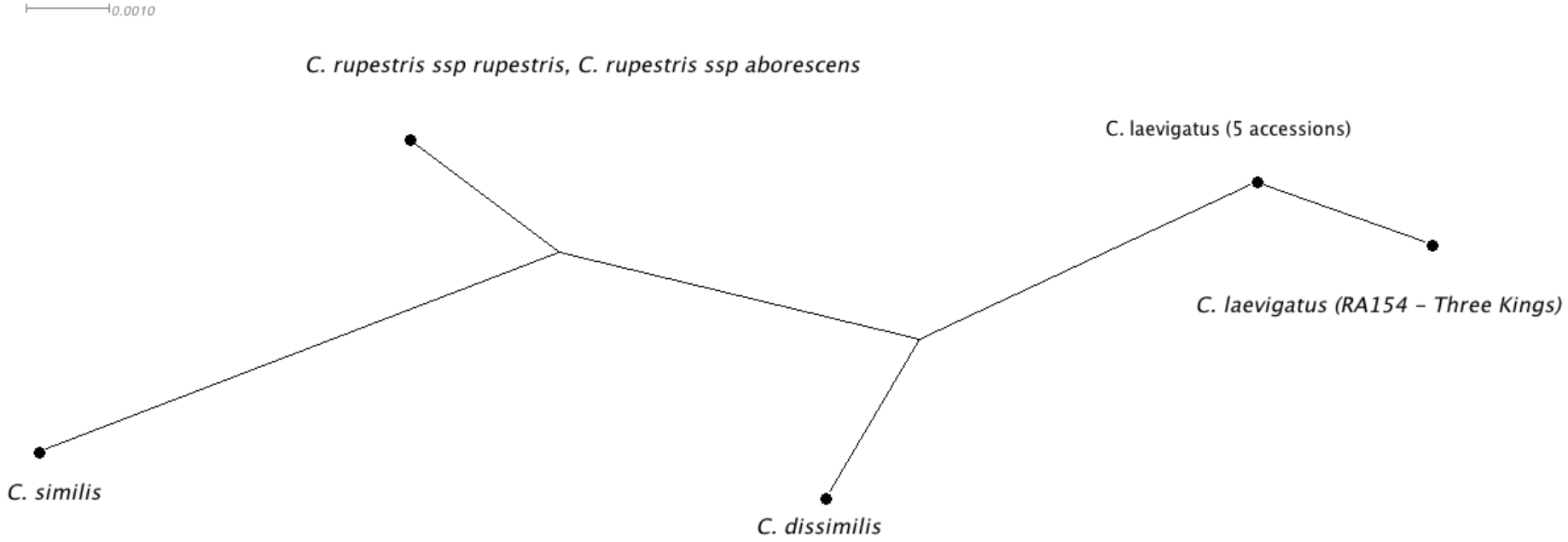

FIGURE 2.3: NEIGHBORNET splits graph computed using SplitsTree4 (Huson & Bryant, 2006) of aligned WAXY sequences from all *Corynocarpus* species except *C. cribbianus for which there was no sequence data.*

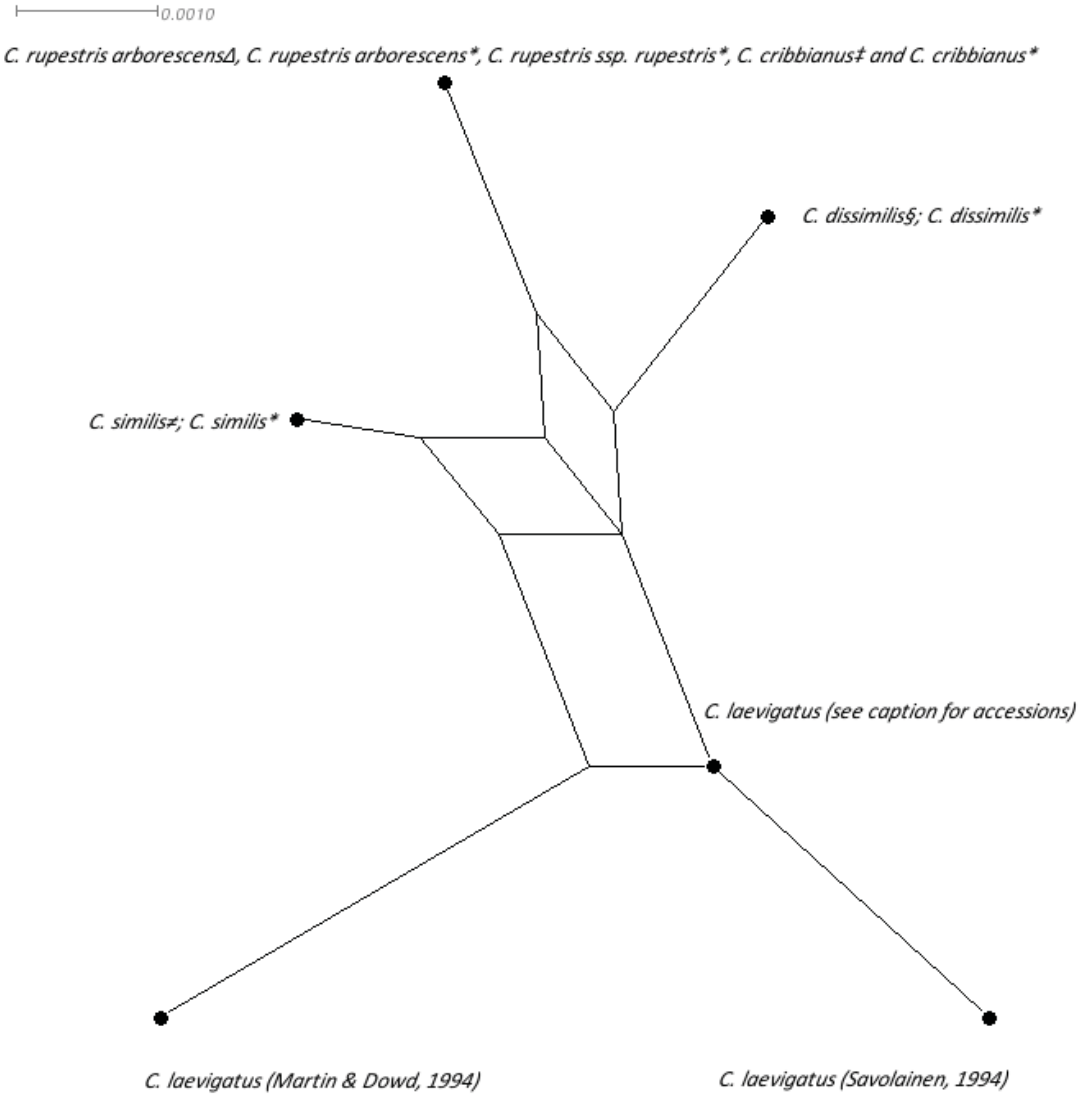

FIGURE 2.4: NEIGHBORNET splits graph computed using SplitsTree4 (Huson & Bryant, 2006) of aligned *rbcL* cpDNA sequences from all *Corynocarpus* species. This splits graph includes *rbcL* sequences of *C. laevigatus* accessions from Martin and Dowd (1994) and Savolainen *et al.* (1994). Δ = GenBank accession no. AF148995; ‡ = GenBank accession no. AF148996; § = GenBank accession number AF148998; ≠ = GenBank accession number AF148997. *C. laevigatus* accessions - *re-sequenced DNA from Wagstaff and Dawson (2000), RA82, RA83, RA84, RA117, RA154, RA517, 1162 (GenBank accession no. HQ207704), AF148994, 96.160 (re-sequenced DNA from Wagstaff (Wagstaff & Garnock-Jones, 1998)).

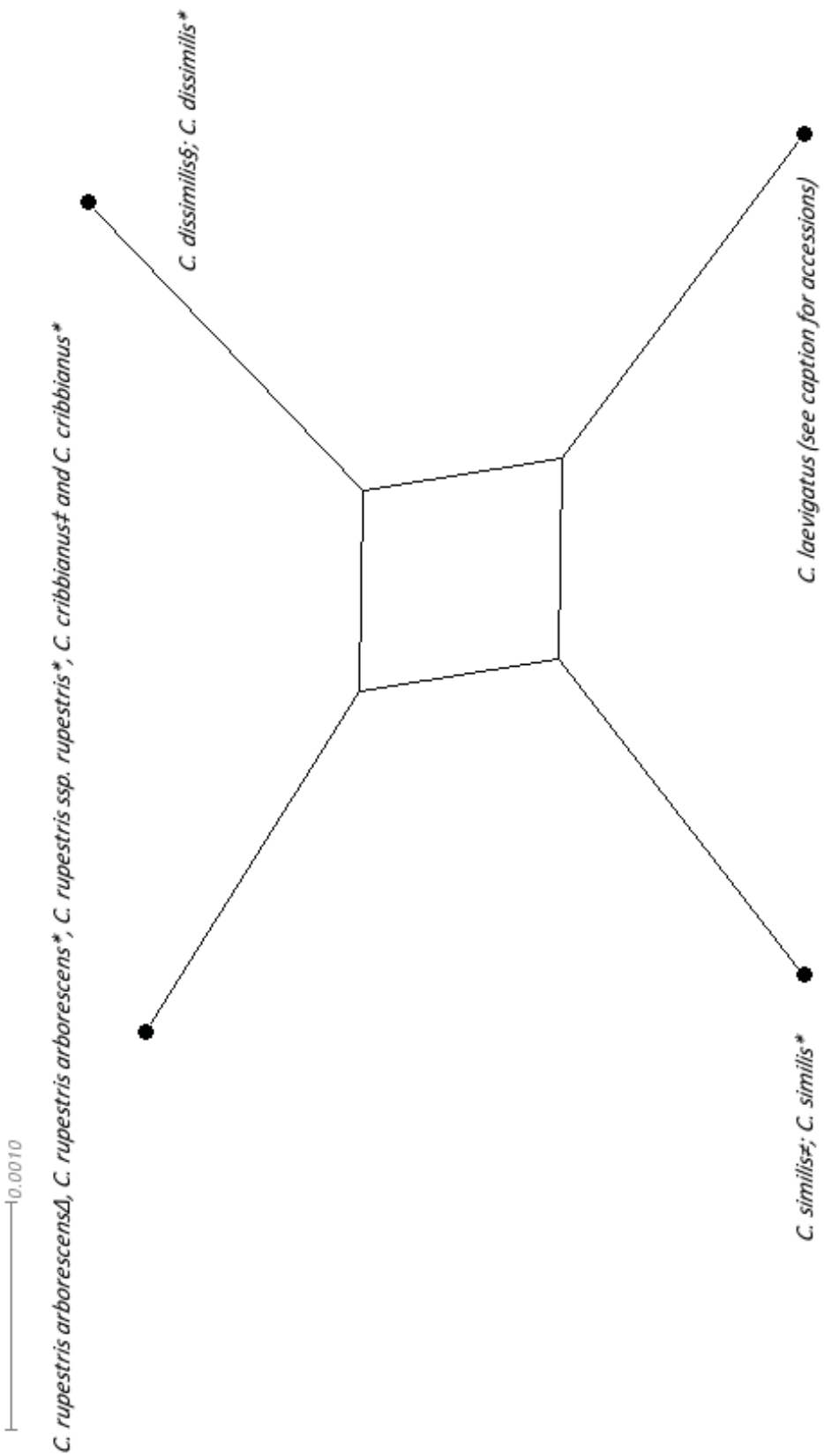
C. dissimilis§; C. dissimilis*
C. laevigatus (see caption for accessions)
C. similis≠; C. similis*
C. rupestris arborescensΔ, C. rupestris arborescens*, C. rupestris ssp. rupestris*, C. cribbianus‡ and C. cribbianus*
0.0010

Table 2.5: Incompatible parsimony site patterns which correspond to position 640 (column 3) and position 814 (column 7) in the rbcL alignment.

Species and accession numbers	Parsimony sites									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
96_102 (*C. rupestris ssp. arborescens*)	C	C	G	G	C	G	G	G	C	T
96_102 (*C. rupestris ssp. arborescens**)	C	C	G	G	C	G	G	G	C	T
96_104 (*C. rupestris ssp. rupestris**)	C	C	G	G	C	G	G	G	C	T
96_105 (*C. dissimilis*)	A	C	G	A	C	G	A	C	A	T
96_105 (*C. dissimilis**)	A	C	G	A	C	G	A	C	A	T
96_106 (*C. cribbianus*)	C	C	G	G	C	G	G	G	C	T
96_106 (*C. cribbianus**)	C	C	G	G	C	G	G	G	C	T
96_138 (*C. similis*)	C	A	C	G	C	C	G	C	A	T
96_138 (*C. similis**)	C	A	C	G	C	C	G	C	A	T
96_160 (*C. laevigatus*)	C	C	C	G	T	G	A	C	A	A
96_160 (*C. laevigatus**)	C	C	C	G	T	G	A	C	A	A
C. laevigatus (Martin & Dowd, 1994)	C	C	C	G	T	C	A	C	A	A
C. laevigatus (Savolainen, et al, 1994)	C	C	C	G	T	G	A	C	A	A
RA82 (*C. laevigatus,* Okains Bay, Banks Peninsula)	C	C	C	G	T	G	A	C	A	A
RA83 (*C. laevigatus*, Chatham Islands)	C	C	C	G	T	G	A	C	A	A
RA84 (*C. laevigatus*, Fanal Island)	C	C	C	G	T	G	A	C	A	A
RA117 (*C. laevigatus*, Kermadec Islands)	C	C	C	G	T	G	A	C	A	A
RA154 (*C. laevigatus*, Three Kings Islands)	C	C	C	G	T	G	A	C	A	A
RA517 (*C. laevigatus*, Mahia Peninsula)	C	C	C	G	T	G	A	C	A	A
5002 (*C. laevigatus*, southern Wairarapa)	C	C	C	G	T	G	A	C	A	A
1162 (*C. laevigatus*, Kermadec Islands)	C	C	C	G	T	G	A	C	A	A

* denotes the re-sequenced accessions of Wagstaff and Dawson (2000)

FIGURE 2.5: (*previous page*) NEIGHBORNET splits graph computed using SplitsTree4 (Huson & Bryant, 2006) of aligned *rbcL* cpDNA sequences from all *Corynocarpus* species. This splits graph excludes *rbcL* sequences of *C. laevigatus* accessions from Martin and Dowd (1994) and Savolainen *et al.* (1994). Δ = GenBank accession no. AF148995; ‡ = GenBank accession no. AF148996; § = GenBank accession number AF148998 ; ≠ = GenBank accession number AF148997. *C. laevigatus* accessions - *re-sequenced DNA from Wagstaff and Dawson (2000), RA82, RA83, RA84, RA117, RA154, RA517, 1162 (GenBank accession no. HQ207704), AF148994, 96.160 (-

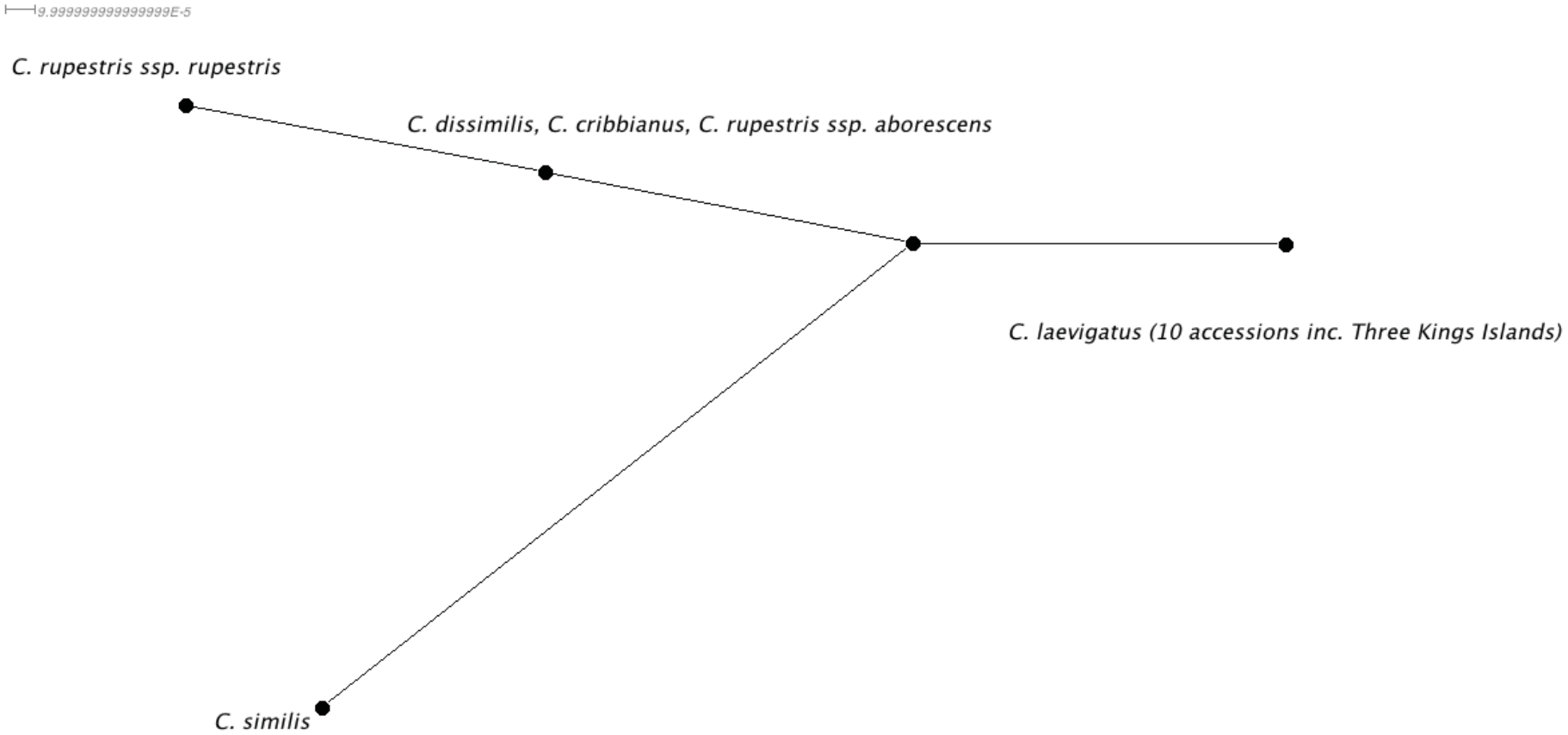

FIGURE 2.6: NEIGHBORNET splits graph computed using SplitsTree4 (Huson & Bryant, 2006) of aligned *trnL-trnF* cpDNA sequences from all *Corynocarpus* species. *C. laevigatus* accessions as follows: RA82, RA83, RA84, RA117, RA154, RA160, RA482, RA517, 1162, 5002 (accession list in Table A2.1, Appendix 2).

2.7 DISCUSSION

2.7.1 ITS AND WAXY SEQUENCES

The analyses of ITS reported here extend the findings of Wagstaff and Dawson (2000), most significantly by including accessions from the Three Kings Islands. An unexpected finding was that sequencing accessions from this island group uncovered genetic variants of karaka not previously recognised. This was observed in both ITS and *Waxy* sequences. Based on our limited sampling these variants appear to be confined to karaka from the Three Kings Islands, indicating that they have been genetically isolated from karaka in the rest of New Zealand. This observation suggests the ancestral nuclear genotype of karaka is not extant (which is often not the case with closely-related species), but existed in northern New Zealand or on a landmass to the north of New Zealand.

The archipelago of the Three Kings Islands (also known as Ngā Motu Karaka) comprises 13 islands ranging from small rock stacks to four main islands, of which the largest is Manawatāwhi (also known as Great Island, King Island and Ohau) at just over 4km^2. Manawatāwhi was inhabited in 1642 when Tasman visited New Zealand (Cheeseman, 1887) and uninhabited until the early 19th century, when members of the iwi (tribe) Te Aupouri, of Northland, moved to the islands. They were noted as living in a state of destitution in 1836 (Cheeseman, 1887), though the exact date the islands became uninhabited is not certain. In a list of plant species growing on Three Kings Islands, karaka is not mentioned by Cheeseman (1887).

The Three Kings flora and fauna has a long history of isolation. The divergence time between insect lineages on The Three Kings Islands and sister groups in the rest of New Zealand range from 2.24–24 mya (Buckley & Leschen, 2013). These dates were obtained from comparative phylogenetic analysis of six insect lineages occurring both on the Three Kings Islands and widespread in New Zealand. Buckley and Leschen (2013) suggest there has been emergent land on the Three Kings Ridge since the Miocene, 24 mya. The lower divergence time of 2.24 mya suggests there was no land connection between the Three Kings Islands and New Zealand during the Pleistocene, when sea levels were lower due to glaciation (Buckley & Leschen, 2013).

The genetic variation detected within karaka allowed calculation of a preliminary estimate for the time of a common karaka ancestor using ITS sequences. The occurrence of four substitutions between the Three Kings Islands and mainland New Zealand haplotypes suggests they diverged 11.65-22.63mya, suggesting that karaka already existed in New Zealand long before humans settled here. Therefore the hypothesis that karaka was introduced to New Zealand from New Caledonia and Vanuatu by the ancestors of Māori (Stevenson, 1978) can be rejected. The oldest dates, using the ITS region for dating, is consistent with the fossil evidence of Campbell (2002) which suggested the presence of karaka in New Zealand ~24 mya and with the recent findings in insect lineages for divergence of some insect taxa from their mainland New Zealand sister taxa (Buckley & Leschen, 2013). Interestingly, accessions from the Kermadec Islands and the Chatham Islands exhibit ITS haplotypes identical to those of mainland New Zealand.

An important point of interest concerns the identical nuclear genotypes found on the Chatham Islands, Kermadec Islands and mainland New Zealand, for both *ITS* and *Waxy* analyses. While preliminary analyses (not shown) showed it was problematic to outgroup root *ITS* (as well as chloroplast phylogenies) using the *Coriaria* and *Tetrameles* sequences available on Genbank, the identical haplotypes in the above locations suggests that recent long distance dispersal links plants in these localities. Whether or not this has been the result of human mediated translocation or natural process cannot be determined from the markers analysed. Recent transoceanic dispersal from New Zealand to outlying landmasses has been a feature of NZ plant biodiversity and natural processes (Winkworth *et al.*, 2002; Gardner *et al.*, 2004; Heenan *et al.*, 2010).

2.7.2 *RBCL* AND TRNL-TRNF SEQUENCES

The results reported here help explain the *rbcL* tree polytomies reported in the study by Wagstaff and Dawson (2000). Multiple substitutions within the *rcbL* gene between very closely related taxa suggests this marker is not ideal for reconstructing *Corynocarpus* relationships. More promising is *trnL-trnF* which produced a tree-like splits graph of phylogenetic relationships. It may be necessary to resequence the *rbcL* sequences to exclude the possibility of sequencing errors.

2.8 CONCLUSION

Nuclear markers suggest a closer relationship between *Corynocarpus laevigatus* and *Corynocarpus dissimilis* whereas the interpretation from chloroplast markers is less clear. This is indicated by the *rbcL* and *trnL-trnF* networks, which both show a reticulation suggesting both *Corynocarpus laevigatus* and *Corynocarpus similis* being more closely related and *Corynocarpus laevigatus* and *Corynocarpus dissimilis* being more closely related. Differences in the results obtained using nuclear *ITS* and Waxy and those obtained using chloroplast *rbcL* and *trnL-trnF* may be explained by the mode of transmission of the two genomes, each of which has a different effect on population structure. Chloroplast genomes are generally inherited maternally in angiosperms and are moved only by seed dispersal. The effective population size of chloroplast genomes is much smaller than that of the nuclear genome of the same organism (Hamilton, 2009). The reduction of the effective population size is caused by two factors: chloroplast genomes are haploid, that is, there is only one copy of the genome in a single chloroplast. Added to this, chloroplasts are only inherited maternally resulting in half the effective population size of the genomes inherited from all possible parents. Thus, chloroplasts have one quarter (0.5×0.5) the effective population size of the nuclear genome of the same organism.

Nevertheless, in all cases, all markers suggest a close relationship between *Corynocarpus laevigatus* and *Corynocarpus* species to the north of New Zealand (*Corynocarpus dissimilis* in New Caledonia and *Corynocarpus similis* in Vanuatu), which supports the work carried out by Wagstaff and Dawson (2000).

All the standard makers employed in this chapter displayed limited phylogenetic resolution. Intraspecific variation within karaka was found to be too low for studying translocation histories within New Zealand. However, based on the data presented in this chapter, these investigations can exclude the Three Kings Islands karaka as a source population for translocated karaka in New Zealand.

Additional sampling and reanalysis of ITS and analysis of *WAXY* and *trnL-trnF* sequences provided further support for the hypothesis of Wagstaff and Dawson (2000) that karaka is derived from the ancestor of more northern *Corynocarpus* species.

Consistent with the findings of Campbell (2002), the findings estimate an age of 11-22my for the ancestor of Three Kings and mainland New Zealand karaka.

Multiple substitutions in *rbcL* appear to make this marker less useful for analyses of *Corynocarpus* phylogeography.

These standard markers indicated limited intraspecific resolution suggesting they would not be suitable for studying translocation histories in New Zealand. For this reason, work described in subsequent chapters of this thesis sought to develop and investigate other chloroplast DNA markers, as well as a rapid means for their assessment.

2.9 REFERENCES

Armstrong, T., and De Lange, P. J. 2005. Conservation genetics of *Hebe speciosa* (Plantaginaceae) an endangered New Zealand shrub. *Botanical Journal of the Linnean Society* **149:** 229–239.

Briggs, J. D., and Leigh, J. H. 1996. *Rare or threatened Australian plants*. 4th edition. CSIRO: Melbourne.

Buckley, T. R., and Leschen, R. A. B. 2013. Comparative phylogenetic analysis reveals long-term isolation of lineages on the Three Kings Islands, New Zealand. *Biological Journal of the Linnean Society* **108:** 361-377.

Burge, P. I., and Shulmeister, J. 2007. Re-envisioning the structure of last glacial vegetation in New Zealand using beetle fossils. *Quaternary Research* **68:** 121-132.

Byrami, M., Ogden, J., Horrocks, M., Deng, Y., Shane, P., and Palmer, J. 2002. A palynological study of Polynesian and European effects on vegetation in Coromandel, New Zealand, showing the variability between four records from a single swamp. *Journal of the Royal Society of New Zealand* **32:** 507-531.

Campbell, J. D. 2002. Angiosperm fruit and leaf fossils from Miocene silcrete, Landslip Hill, northern Southland, New Zealand. *Journal of the Royal Society of New Zealand* **32:** 149-154.

Cheeseman, T. 1887. Art. XXII - Notes on the Three Kings Islands. *Transactions and Proceedings of the Royal Society of New Zealand* **20:** 142-149.

Cook, R. A., Sutherland, R., and Zhu, H. 1999. *Cretaceous - Cenozoic geology and petroleum systems of the Great South Basin, New Zealand. Monograph 20*. Institute of Geological & Nuclear Sciences: Lower Hutt.

Dodson, J. R. 1976. Modern pollen spectra from Chatham Island, New Zealand. *New Zealand Journal of Botany* **14:** 341-347.

Doyle, J. J., and Doyle, J. L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue *Phytochemistry Bulletin* **19:** 11–15.

Eagle, A. 2006. *Eagle's Complete Trees and Shrubs of New Zealand*. Te Papa Press: Wellington.

French, B. 2006. *Food crops of Papua New Guinea* Published by the Author: Sheffield, Tasmania.

Gardner, R., De Lange, P., Keeling, D., and Bowala, T. 2004. A late Quaternary phylogeography for *Metrosideros* (Myrtaceae) in New Zealand inferred from chloroplast DNA haplotypes *Biological Journal of the Linnean Society* **83:** 399-412.

Garnier, B. J. 1958. *The climate of New Zealand*. Edward Arnold: London.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59:** 307-321.

Hamilton, M. B. 2009. Genetic drift and effective population size. *In Population genetics*. Wiley-Blackwell: Chichester, UK.

Hasebe, M., Omori, T., Nakazawa, M., Sano, T., Kato, M., and Iwatsuki, K. 1994. *rbcL* gene-sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 5730-5734.

Heenan, P., Mitchell, A., de Lange, P., Keeling, J., and Paterson, A. 2010. Late-Cenozoic origin and diversification of Chatham Island endemic plant species revelaed by analyses of sequence data. *New Zealand Journal of Botany* **48:** 83-136.

Hemsley, W. B. 1903. On the genus *Corynocarpus*, Forst., with descriptions of two new species. *Annals of Botany* **17:** 743-760.

Herzer, R. H., Chaproniere, G. C. H., Edwards, A. R., Hollis, C. J., Pelletier, B., Raine, J. I., Scott, G. H., Stagpoole, V., Strong, C. P., Symonds, P., Wilson, G. J., and Zhu, H. 1997. Seismic stratigraphy and structural history of the Reinga Basin and its margins, southern Norfolk Ridge system. *New Zealand Journal of Geology and Geophysics* **40:** 425-451.

Hicks, S. 2006. When no pollen does not mean no trees. *Vegetation History and Archaeobotany* **15:** 253-261.

Holmgren, P., Holmgren, N., and Barnett, L. 1990. Index herbariorum. Part 1, The herbaria of the world. *Regnum Vegetabile* **120:** 1-693.

Holt, K. 2009. *The Quarternary History of Chatham Island, New Zealand. Ph.D. thesis*. Massey University, Palmerston North.

Hörandl, E., Paun, O., Johansson, J. T., Lehnebach, C., Armstrong, T., Chen, L., and Lockhart, P. 2005. Phylogenetic relationships and evolutionary traits in *Ranunculus* s.l. (Ranunculaceae) inferred from ITS sequence analysis. *Molecular Phylogenetics and Evolution* **36:** 305-327.

Huson, D. H., and Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23:** 254-267.

Kay, K., Whittall, J., and Hodges, S. 2006. A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evolutionary Biology* **6:** 1471-2148.

Knapp, M., Mudaliar, R., Havell, D., Wagstaff, S. J., and Lockhart, P. J. 2007. The drowning of New Zealand and the problem of *Agathis*. *Systematic Biology* **56:** 862-870.

Labate, J. A., Robertson, L. D., and Baldo, A. M. 2009. Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). *Heredity* **103:** 257-267.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Lee, D. E., Lee, W. G., and Mortimer, N. 2001. Where and why have all the flowers gone? Depletion and turnover in the New Zealand Cenozoic angiosperm flora in relation to palaeogeography and climate. *Australian Journal of Botany* **49:** 341–356.

Macphail, M. K., and McQueen, D. R. 1983. The value of New Zealand pollen and spores as indicators of Cenozoic vegetation and climates. *Tuatara* **26:** 37-56.

McGlone, M. 1985. Plant biogeography and the late Cenozoic history of New Zealand. *New Zealand Journal of Botany* **23:** 723-749.

McGlone, M., Duncan, R. P., and Heenan, P. B. 2001. Endemism, species selection and the origin and distribution of the vascular plant flora of New Zealand *Journal of Biogeography* **26:** 199-216.

McGlone, M., Salinger, M. J., Moar, N. T., and Kutzbach, J. E. 1993. Paleovegetation studies of New Zealand's climate since the Last Glacial Maximum. *In* H.E., J., Kutzbach, J.E., Webb III, T., Ruddiman, W.F., Street-Perrott, F.A. and Bartlein, P.J. (Ed.), *Global climates since the last glacial maximum*, pp. 294-317. University of Minnesota Press: Minneapolis, .

Mildenhall, D. C. 1994. *Palynological reconnaissance of Early Cretaceous to Holocene sediments, Chatham Islands, New Zealand*. Institute of Geological & Nuclear Sciences Ltd: Lower Hutt, New Zealand.

Mitchell, A. D., and Heenan, P. B. 2000. Systematic Relationships of New Zealand Endemic Brassicaceae Inferred from nrDNA ITS Sequence Data. *Systematic Botany* **25:** 98-105.

Molloy, B. P. J. 1990. The origin, relationships, and use of karaka or kopi (*Corynocarpus laevigatus*). *In* Kapoor, W. H. a. P. (Ed.), *Nga Mahi Maori o te Wao Nui a Tane: contributions to an International Workshop on Ethnobotany, Te Rehua Marae*, pp. 48–53. Botany Division, Department of Scientific and Industrial Research: Christchurch, New Zealand.

Newnham, R., McGlone, M., Moar, N., Wilmshurst, J., and Vandergoes, M. in press. The vegetation cover of New Zealand at the Last Glacial Maximum. *Quaternary Science Reviews*.

Olsen, K. 2004. SNPs, SSRs and inferences on cassava's origin. *Plant Molecular Biology* **56:** 517-526.

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Muller-Starck, G. M., Demesure-Musch, B., Palme, A., Martin, J. P., Rendell, S., and Vendramin, G. G. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300:** 1563-1565.

Platt, G. 2003. Observations on karaka (*Corynocarpus laevigatus*) and its fruit. *Auckland Botanical Society Journal* **58:** 29-31.

Sang, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* **37:** 121-147.

Savolainen, V., Manen, J. F., Douzery, E., and Spichiger, R. 1994. Molecular phylogeny of families related to Celastrales based on *rbcL* 5' flanking sequences. *Mol Phylogenet Evol* **3:** 27-37.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., and Liu, W. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92:** 142–166.

Shepherd, L. D., and Perrie, L. R. 2011. Microsatellite DNA analyses of a highly disjunct New Zealand tree reveal strong differentiation and imply a formerly more continuous distribution. *Molecular Ecology* **20:** 1389-1400.

Shepherd, L. D., Perrie, L. R., and Brownsey, P. J. 2007. Fire and ice: volcanic and glacial impacts on the phylogeography of the New Zealand forest fern *Asplenium hookerianum*. *Mol Ecol* **16:** 4536-4549.

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G. J. 2005. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 14694-14699.

Steel, M., Huson, D., and Lockhart, P. J. 2000. Invariable sites models and their use in phylogeny reconstruction. *Systematic Biology* **49:** 225-232.

Stevenson, G. 1978. Botanical evidence linking the New Zealand Maoris with New Caledonia and the New Hebrides. *Nature* **276:** 704 - 705.

Stöckler, K., Daniel, I., and Lockhart, P. 2002. New Zealand Kauri (*Agathis australis* (D.Don) Lindl., Araucariaceae) survives Oligocene drowning. *Systematic Biology* **51:** 827-832.

Stowe, C. J. 2003. *The ecology and ethnobotany of karaka (Corynocarpus laevigatus) [MSc. thesis]*. University of Otago, Dunedin.

Swofford, D. 2003. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), version 4.0*. Sinauer Associates: Sunderland, Massachusetts.

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17:** 1105-1109.

van Steenis, C. G. G. J. 1951. Corynocarpaceae. *In Flora Malesiana Series 1: Spermatophyta*, pp. 262–264. Noordhoff–Kolff: Jakarta.

Wagstaff, S., and Dawson, M. 2000. Classification, origin, and patterns of diversification of *Corynocarpus* (Corynocarpaceae) inferred from DNA sequences. *Systematic Botany***:** 134-149.

Wagstaff, S. J., and Garnock-Jones, P. J. 1998. Evolution and biogeography of the Hebe complex (Scrophulariaceae) inferred from ITS sequences. *New Zealand Journal of Botany* **36:** 425-437.

Wagstaff, S. J., and Garnock-Jones, P. J. 1998. Evolution and biogeography of the Hebe complex (Scrophulariaceae) inferred from ITS sequences. *New Zealand Journal of Botany* **36:** 425-437.

Wardle, P. 1963. Evolution and distribution of the New Zealand flora, as affected by Quaternary climates. *Evolution*.

———. 1985. Environmental influences on the vegetation of New Zealand. *New Zealand Journal of Botany*.

White, T., Bruns, T., Lee, S., and Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M., Gelfand, D., Shinsky, J., and White, T. (Eds), *PCR Protocols: A Guide to Methods and Applications*, pp. 315-322. Academic Press.

Wilmshurst, J. M., Higham, T. F. G., Allen, H., and Johns, D. 2004. Early Maori settlement impacts in northern coastal Taranaki, New Zealand. *New Zealand Journal of Ecology* **28:** 167-179.

Winkworth, R. C., Wagstaff, S. J., Glenny, D., and Lockhart, P. J. 2002. Plant dispersal NEWS from New Zealand. *Trends in Ecology & Evolution*.

3

WHOLE GENOME SEQUENCING OF ENRICHED CHLOROPLAST DNA USING THE ILLUMINA GAII PLATFORM

PREAMBLE

The standard makers employed in the previous chapter displayed limited phylogenetic resolution. Intraspecific variation within karaka was found to be too low for studying translocation histories within NZ. However, based on the data presented in this chapter, these investigations can exclude the Three Kings Islands karaka as a source population for translocated karaka in New Zealand. For this reason, work described in subsequent chapters of this thesis sought to develop and investigate other chloroplast DNA markers, as well as rapid means for their assessment.

3.1 UTILITY/EVALUATION OF CHLOROPLAST AS A MOLECULE FOR A HIGH-RESOLUTION STUDY OF TRANSLOCATION

Molecular markers are used to characterize the basis of genetic variation in or between taxa and the application of this data provides answers to ecological, historical, evolutionary or phylogenetic questions. Molecular markers have long been used to investigate the process of plant domestication and for resolving genetic relationships between domesticates and their wild progenitors. The number of domestication events and their location(s) can also be inferred from molecular data. Common markers used for these studies include microsatellites (SSRs) (Howe *et al.*, 2003; González-Jara *et al.*), single nucleotide polymorphisms (SNPs) (Olsen, 2004; Labate *et al.*, 2009), amplified fragment length polymorphisms, (Allaby & Brown, 2003; Spooner *et al.*, 2005) and inter simple sequence repeats markers (ISSRs) (Clarke, 2006).

As already discussed, the natural range of karaka may have been restricted to Northland and all populations growing south of this region could be the result of translocation by humans. It is likely only a small subset of the natural population would have been translocated as part of the cultivation of karaka, resulting in population bottlenecks, which reduce genetic variation. To determine the nature of these translocations and their effect on the population structure of the species it is necessary to develop molecular markers. These markers are derived from sequence changes in small stretches of DNA that show polymorphisms between individuals and are used to infer relationships between organisms.

Plants contain three genomes: nuclear, mitochondrial and chloroplast. Each has a different mode of transmission, which influences their pattern of population structuring. Nuclear genes follow Mendelian inheritance and are usually biparentally inherited. Chloroplasts and mitochondria have non-Mendelian inheritance, also known as *extranuclear* or *cytoplasmic inheritance* (Röhr *et al.*, 1999). The effective population size of chloroplast genomes is much smaller than that of the nuclear genome of the same organism (Hamilton, 2009). The reduction of the effective population size is caused by two factors: chloroplast genomes are haploid, that is, there is only one copy of the genome in a single chloroplast. Added to this, chloroplasts (generally) are only inherited maternally resulting in half the effective population size of the genomes inherited from all possible parents. Thus, chloroplasts have one quarter (0.5 x 0.5) the effective population size of the nuclear genome of the same organism. This can be useful for studies of recent divergence in species due to genetic drift where a large effective population size in the nuclear genome would show much less divergence (Hamilton, 2009).

The use of organellar DNA is universal in phylogenetic studies and for data to be correctly analysed and interpreted, the mode of inheritance of chloroplasts needs to be known (Harris & Ingram, 1991). However, in most cases, because a lack of published data exists for most species, assumptions must be made on the mode of inheritance based on available knowledge of closely related species or genera. The mode of inheritance of chloroplasts is varied in seed plants, ranging from strictly maternal (inherited through the female parent) to strictly paternal (inherited through the male

parent) (Harris & Ingram, 1991; Reboud & Zeyl, 1994; Mogensen, 1996; Röhr *et al.*, 1999).

It is assumed that chloroplast genomes are generally uniparentally inherited, predominantly maternally, in angiosperms (Corriveau & Coleman, 1988; Reboud & Zeyl, 1994; Birky, 1995) and are thus moved only by seed dispersal. Although maternal inheritance of chloroplasts is more common in angiosperms, there is some evidence of paternal transmission in some species, e.g., kiwifruit (Chat *et al.*, 1999) and *Turnera ulmifolia* (Shore & Triassi, 1998), *Passiflora* (Hansen *et al.*, 2007) and *Medicago sativa* (Schumann & Hancock, 1989). Documentation of chloroplast heteroplasmy is rare, perhaps due, in part, to the dogma of strict maternal inheritance in angiosperms (Ellis *et al.*, 2008). It is therefore advisable to check the chloroplast inheritance if such markers are to be used for analyses assuming strict maternal inheritance of this genome. (Raspé, 2001).

Although the inheritance of cpDNA has not yet been investigated in the Corynocarpaceae and published data on chloroplast inheritance in Corynocarpaceae are non-existent, it has been assumed to be maternal. Chloroplast inheritance in a closely-related family, Cucurbitaceae, has shown maternal inheritance (Havey *et al.*, 1998).

Historically, molecular diversity studies have used markers such as allozymes, isozymes, AFLP, RAPD and SSRs, the choice of marker depending upon the organism and the question being asked. However, the advent of high-throughput second-generation sequencing has shifted the focus to nucleotide-based surveys detecting patterns of polymorphisms across whole genomes. Nuclear DNA sequences have the advantage of providing evidence of both the maternal and paternal lineages. Levels of polymorphism for genomic DNA can be more suitable for analyses of intraspecific variation than organellar DNA (Doebley, 1992). However, the size and complexity of chloroplast genomes means they can contain structural and point mutations that can be used to study population-level processes (Cronn *et al.*, 2008). Conservatism of cpDNA generally can result in low levels of intraspecific variation, often reducing its usefulness for studies at this taxonomic level (Doebley, 1992).

Plastids are generally maternally inherited in angiosperms and, therefore, moved by seeds only. Because translocation of karaka was mainly by seed (there is one example of whole tree dispersal in Taupo) cpDNA markers would provide information on the natural distribution of karaka and the magnitude of bottlenecking in the translocated populations. cpDNA markers provide information on past changes in species distribution that is unaffected by subsequent pollen movements. Intraspecific polymorphisms in the chloroplast genome can be difficult to discover in recently diverged populations (McCauley, 1995). One way to search for polymorphisms is by comparative sequencing of stretches of PCR-amplified non-coding cpDNA (Weising *et al.*, 2005) to mine for single nucleotide polymorphisms (SNPs). SNPs, as their name suggests, are single base change differences between homologous DNA sequences usually with two possible nucleotides at a given position. Mutation mechanisms result either in transitions or transversion. Transitions are either purine-purine (A⇔G) or pyrimidine-pyrimidine (C⇔T) exchanges. Transversions are either purine-pyrimidine or pyrimidine-purine exchanges (e.g. A⇔C, A⇔T, G⇔C, G⇔T) exchanges (Vignal *et al.*, 2002).

The first step in the development of cp markers was characterisation of the chloroplast genome as a reference for different strategies in molecular marker identification. This chapter describes a protocol developed for the isolation of chloroplasts and the sequencing of their genomes using the Illumina Genome Analyser II. This protocol was also shown to be effective in the characterisation of chloroplast genomes in other elements of the New Zealand flora. Two papers were published which made use of this protocol: Zhong *et al.* (2011) (Appendix 5) and Goremykin *et al.* (2012) (Appendix 6).

The protocol, which follows section 3.2, was published in Plant Methods in September 2010:

Atherton, R. A., McComish, B. J., Shepherd, L. D., Berry, L. A., Albert, N. W., and Lockhart, P. J. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods* **6:**

3.2 REFERENCES

Allaby, R. G., and Brown, T. A. 2003. AFLP data and the origins of domesticated crops. *Genome* **46:** 448-453.

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

Chat, J., Chalak, L., and Petit, R. J. 1999. Strict paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in intraspecific crosses of kiwifruit. *Theoretical and Applied Genetics* **99:** 314-322.

Clarke, A. 2006. Reconstructing the Origins and Dispersal of the Polynesian Bottle Gourd (Lagenaria siceraria). *Molecular Biology and Evolution* **23:** 893-900.

Corriveau, J. L., and Coleman, A. W. 1988. Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for Over 200 Angiosperm Species. *American Journal of Botany* **75:** 1443-1458.

Cronn, R., Liston, A., Parks, M., Gernandt, D., Shen, R., and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* **36:** e122-e122.

Doebley, J. 1992. Molecular systematics and crop evolution. *In* Soltis, P. S., Soltis, D., Doyle, J.J. (Ed.), *Molecular systematics of plants*, pp. 202–222. Chapman & Hall: London.

Ellis, J. R., Bentley, K. E., and McCauley, D. E. 2008. Detection of rare paternal chloroplast inheritance in controlled crosses of the endangered sunflower Helianthus verticillatus. *Heredity (Edinb)* **100:** 574-580.

González-Jara, P., Moreno-Letelier, A., Fraile, A., Piñero, D., and García-Arenal, F. 2011. Impact of Human Management on the Genetic Variation of Wild Pepper, *Capsicum annuum* var. *glabriusculum*. *PLOS One* **6:** e28715.

Goremykin, V., Nikiforova, S., Biggs, P., Zhong, B., DeLange, P., Martin, W., Woetzel, S., Atherton, R., McLenachan, T., and Lockhart, P. 2012. The evolutionary root of flowering plants. *Systematic Biology* **62:** 50–61.

Hamilton, M. B. 2009. Genetic drift and effective population size. *In Population genetics*. Wiley-Blackwell: Chichester, UK.

Hansen, A. K., Escobar, L. K., Gilbert, L. E., and Jansen, R. K. 2007. Paternal, maternal, and biparental inheritance of the chloroplast genome in Passiflora (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany* **94:** 42-46.

Harris, S. A., and Ingram, R. 1991. Chloroplast DNA and Biosystematics: The Effects of Intraspecific Diversity and Plastid Transmission. *Taxon* **40:** 393-412.

Havey, M. J., McCreight, J. D., Rhodes, B., and Taurick, G. 1998. Differential transmission of the Cucumis organellar genomes. *Theoretical and Applied Genetics* **97:** 122-128.

Howe, C., Barbrook, A., Koumandou, V., Ellen, R., Nisbet, R., Symington, H., and Wightman, T. 2003. Evolution of the Chloroplast Genome. *Philosophical Transactions of the Royal Society of London B Biol Sci.* **358:** 99-107.

Labate, J. A., Robertson, L. D., and Baldo, A. M. 2009. Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). *Heredity* **103:** 257-267.

McCauley, D. E. 1995. The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology & Evolution* **10:** 198-202.

Mogensen, H. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83:** 383-404.

Olsen, K. 2004. SNPs, SSRs and inferences on cassava's origin. *Plant Molecular Biology* **56:** 517-526.

Raspé, O. 2001. Inheritance of the Chloroplast Genome in Sorbus aucuparia L. (Rosaceae). *Journal of Heredity* **92:** 507-509.

Reboud, X., and Zeyl, C. 1994. Organelle inheritance in plants. *Heredity* **72:** 9.

Röhr, H., Kües, U., and Stahl, U. 1999. Recombination: organelle DNA of plants and fungi: inheritance and recombination. *In* Esser, K., Kadereit, J. W., Lüttge, U., and Runge, M. (Eds), *Progress in Botany*, vol. 60, Progress in Botany, pp. 39-87. Springer Berlin Heidelberg.

Schumann, C. M., and Hancock, J. F. 1989. Paternal inheritance of plastids in Medicago sativa. *Theoretical and Applied Genetics* **78:** 863-866.

Shore, J., and Triassi, M. 1998. Paternally biased cpDNA inheritance in Turnera ulmifolia (Turneraceae). *American Journal of Botany* **85:** 328.

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G. J. 2005. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 14694-14699.

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

Weising, K., Nybom, H., Wolff, K., and Kahl, G. 2005. Applications of DNA fingerprinting in plant science. *In DNA fingerprinting in plants: principles, methods, and applications*, p. 472. CRC Press: Boca Raton.

Zhong, B., Deusch, O., Goremykin, V., Penny, D., Biggs, P., Atherton, R., Nikiforova, S., and Lockhart, P. 2011. Systematic error in seed plant phylogenomics. *Genome Biology and Evolution* **3**.

Atherton *et al. Plant Methods* 2010, **6**:22
http://www.plantmethods.com/content/6/1/22

PLANT METHODS

METHODOLOGY Open Access

Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform

Robin A Atherton[1,2*†], Bennet J McComish[1,2†], Lara D Shepherd[1], Lorraine A Berry[3], Nick W Albert[1], Peter J Lockhart[4]

Abstract

Background: Complete chloroplast genome sequences provide a valuable source of molecular markers for studies in molecular ecology and evolution of plants. To obtain complete genome sequences, recent studies have made use of the polymerase chain reaction to amplify overlapping fragments from conserved gene loci. However, this approach is time consuming and can be more difficult to implement where gene organisation differs among plants. An alternative approach is to first isolate chloroplasts and then use the capacity of high-throughput sequencing to obtain complete genome sequences. We report our findings from studies of the latter approach, which used a simple chloroplast isolation procedure, multiply-primed rolling circle amplification of chloroplast DNA, Illumina Genome Analyzer II sequencing, and de novo assembly of paired-end sequence reads.

Results: A modified rapid chloroplast isolation protocol was used to obtain plant DNA that was enriched for chloroplast DNA, but nevertheless contained nuclear and mitochondrial DNA. Multiply-primed rolling circle amplification of this mixed template produced sufficient quantities of chloroplast DNA, even when the amount of starting material was small, and improved the template quality for Illumina Genome Analyzer II (hereafter Illumina GAII) sequencing. We demonstrate, using independent samples of karaka (*Corynocarpus laevigatus*), that there is high fidelity in the sequence obtained from this template. Although less than 20% of our sequenced reads could be mapped to chloroplast genome, it was relatively easy to assemble complete chloroplast genome sequences from the mixture of nuclear, mitochondrial and chloroplast reads.

Conclusions: We report successful whole genome sequencing of chloroplast DNA from karaka, obtained efficiently and with high fidelity.

Background

Chloroplast genomes provide a wealth of information for studies in molecular ecology and evolution. Their conservative gene content and organisation have enabled researchers to isolate homologous loci for comparative studies over different evolutionary time-scales [1-7].

Obtaining the DNA sequence for chloroplast genomes can be achieved by using the polymerase chain reaction (PCR) to amplify chloroplast DNA fragments from genomic DNA (gDNA) extracts. However, this can involve up to 35 amplifications of overlapping chloroplast DNA PCR products [2,8]. While this approach is time consuming [8], it has been preferred over protocols that attempt to first separate chloroplasts from other cellular material. Reasons for this appear to be that chloroplast isolation can be troublesome in some species [9] and because rapid chloroplast isolation protocols often produce template which is still contaminated by large quantities of nuclear DNA [10]. Nevertheless, given the depth of sequencing coverage with the Illumina GAII sequencing platform, we were interested to investigate whether this alternative approach could be used for sequencing whole chloroplast genomes without the need for whole genome PCR amplification. Here we report findings which demonstrate that, even with small amounts of chloroplast DNA, and in the presence of large amounts of nuclear DNA, Illumina short read sequencing provides a practical approach for obtaining complete chloroplast genome sequences.

* Correspondence: r.a.atherton@massey.ac.nz
† Contributed equally
[1]Institute of Molecular BioSciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand
Full list of author information is available at the end of the article

Atherton et al. Plant Methods 2010, 6:22
http://www.plantmethods.com/content/6/1/22

Methods

Chloroplast isolation

Fresh leaf material was obtained from two cultivated karaka trees originating in Rekohu/Chatham Islands and the Kermadec Islands, New Zealand. Leaf material was collected and processed immediately for the sample from the Chatham Islands and within 3 h for the sample from the Kermadec Islands. Leaf samples weighing 2.5-5 g were excised from living trees and processed as follows. Chloroplasts were isolated using a protocol originally designed for isolating chloroplasts from *Arabidopsis thaliana* [11] with minor modifications: (i) the leaf material was homogenised using an Ultra-Turrax homogeniser with an N18 rotor (Janke & Kunkel IKA, Hamburg, Germany); (ii) the homogenate was passed through a double layer of washed and autoclaved nappy (diaper) liner (Johnson & Johnson Ltd.) rather than through Miracloth (Calbiochem); and (iii) the final centrifugation step was carried out using a Sorvall SS32 angled rotor, rather than a swinging bucket rotor. After the final centrifugation step, DNA was extracted from pooled chloroplasts for each tree sample using a DNEasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. Genomic DNA (gDNA) was extracted from silica-dried karaka leaf material from the same accessions using a DNEasy Plant Mini Kit (Qiagen).

Multiply-primed rolling circle amplification

Multiply-primed rolling circle amplification (RCA) was used to produce an abundance of purified chloroplast DNA template in preparation for sequencing [12]. This technique involves isothermal, strand-displacing amplification using multiple primers and is capable of yielding a large amount of product from very little starting DNA template [13]. Phi29, the DNA polymerase used in multiply-primed RCA, is reported to have a very low level of amplification bias making the template suitable for whole genome sequencing [14]. Chloroplast-enriched DNA (cpDNA) from both karaka samples was amplified in this way using a REPLI-g™ Mini Kit (Qiagen) following the manufacturer's instructions, with the exception that samples were incubated at room temperature for 9 min rather than the recommended 3 min. This extension time consistently produced better results with different plant samples. The kit produced ~5 μg of product for each sample.

Confirmation of chloroplast DNA enrichment

Genomic DNA (gDNA), chloroplast-enriched DNA (cpDNA) and RCA amplified chloroplast-enriched DNA (RCAcpDNA) from the Chatham Island sample were quantified fluorometrically using the Quant-iT™ dsDNA HS assay kit on a Qubit™ Quantitation Platform (Invitrogen). The concentration of the gDNA, cpDNA and RCAcpDNA was 110 ngμL^{-1}, 20 ngμL^{-1} and 104 ngμL^{-1}, respectively, in a total volume of 50 μL of AE buffer (Qiagen). The purity of gDNA, cpDNA and RCAcpDNA samples was determined by A_{260}/A_{280} and A_{260}/A_{230} ratios on a NanoDrop (NanoDrop Technologies) spectrophotometer. Enrichment for chloroplast DNA was determined by quantitative real-time PCR (qPCR) with gDNA, cpDNA and RCAcpDNA templates; the quantity of the plastid gene *psbB* was determined relative to nuclear encoded 18S *rRNA* by comparative quantification [15]. Gene-specific primers were designed for *psbB* (*psbB* F 5'GGGGGTTGGAGTATCACAGG3'; *psbB* R 5'CCAAGAAGCACAAGCCAGAA3', 103 bp amplicon) using Primer3 [16] and primers for 18S are described by Zhu and Altmann [17]. qPCR was performed using Lightcycler480 SYBR Green1 Master (Roche Diagnostics) reagents in a Rotor Gene 3000 instrument (Corbett Research) with four technical replicates per sample. Template DNA was diluted 20-fold for cpDNA, and 100-fold for gDNA and RCAcpDNA samples for qPCR. The qPCR cycling conditions were: 95°C 10 min, (95°C 10 s, 60°C 15 s, 72°C 20 s) × 40 cycles with fluorescent detection at 72°C and during the final melt. Melt curve analysis confirmed the amplification of a single product.

Illumina GAII sequencing

The RCAcpDNA samples from both accessions were sequenced by Massey Genome Service (Massey University, Palmerston North, New Zealand). A 75 bp paired-end run was performed on the Illumina GAII with the two samples described here in a single lane along with four other samples from a separate experiment. Samples were prepared for sequencing as follows: genomic DNA libraries were prepared by fragmenting purified genomic DNA using a nebulisation kit (Invitrogen), paired-end index adaptor ligation (Illumina) and 18 cycles of PCR enrichment using the Illumina Paired-End Genomic DNA library preparation kit, Illumina Multiplex Oligonucleotide library preparation kit and Illumina Multiplex Paired-End Genomic DNA library preparation protocol. The enriched libraries were quantified using an ND-1000 NanoDrop spectrophotometer (NanoDrop Technologies) and quality checked by Agilent 2100 Bioanalyzer, DNA 1000 Labchip kit assay. The libraries were then diluted to a 10 nM concentration using EB buffer (Qiagen) and quantified for optimal cluster density using the LightCycler® 480 system Absolute Quantification protocol (Roche Diagnostics) and the LightCycler® 480 SYBR Green I Master kit (Roche Diagnostics). The libraries were pooled at equal molarity and amplified in one flow cell lane on the Illumina Cluster Station instrument at a density of 140,000 clusters per tile and a molarity of 13 pM using the Illumina Paired-

End Cluster Generation kit v2. The amplified libraries were sequenced on the Illumina GAII instrument using 4 Illumina 36 cycle SBS sequencing kits (v3), Illumina Multiplex Sequencing Primers and PhiX control kit v2, on a 75 bp paired-end indexed run. After sequencing, the resulting images were analysed with the proprietary Illumina pipeline v1.3. Reads for each of the indexed samples were then separated using a custom Perl script.

Assembly

Reads from each indexed sample were trimmed to remove poor quality sequence at the 3' end. To determine the optimum trim length, initial de novo assemblies were made for read sets of different length (untrimmed reads, and reads trimmed to 70, 65, 55, 50 bp). These assemblies were carried out using Velvet 0.7 [18] with a range of hash lengths from 33 to 63 and a minimum k-mer coverage of 5×. For these initial assemblies, the data were treated as single reads, that is, the paired-end information was not used. Maximum contig lengths and N50 values were tabulated and the hash lengths that gave the highest N50 for each trimmed set of reads were selected for further optimisation. A second round of assembly was carried out on each trimmed set of reads using the hash length determined above and varying the coverage cut-off parameter from 1 to 100. Finally, paired-end assembly was carried out for each of these read-length/hash-length combinations using the coverage cut-off value that gave the highest N50 value. For these paired-end assemblies, expected coverage was set to the length-weighted median of the coverage values obtained in the initial single read assemblies, and the insert length was estimated as 240 bp. Assembled contigs were aligned to the *Cucumis sativus* chloroplast genome [GenBank: NC_007144; GenBank: DQ119058] using Geneious 4.7 [19].

Four short regions of ambiguous sequence were checked by PCR amplification using the following primers, custom designed using Primer3 [16] unless referenced: Corlaerps2-rpoc2F (TATAGGGTGCCATTCG AGGA), Corlaerps2-rpoc2R GTATCAACAACGGC-CAATCC; CorlaendhAF (GGAATAGGATGGAGA-TAAGAAAGAC), CorlaendhAR (CACGATTCCG ATCCAGAGTA); psbJ ATAGGTACTGTARCYGG-TAT [20], petA AACARTTYGARAAGGTTCAATT [20]; psbAR (CGCGTCTCTCTAAAATTGCAGTCAT) [21], CorlaepsbA-R (ATCCGACTAGTTCCGGGTTC). Figure 1 shows the relative position of the priming sites on the karaka chloroplast genome. The PCR cycling conditions were modified slightly from an existing published protocol [20] as follows: template denaturation at 80°C for 5 min followed by 32 cycles of denaturation at 95°C for 1 min, primer annealing at 50°C for 1 min, followed by a ramp of 0.3°C/s to 65°C, and primer extension at 65°C for 4 min; followed by a final extension step of 5 min at 65°C. Amplified PCR products were sequenced using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) and an ABI 3730 automated capillary sequencer at Massey Genome Service (Massey University, Palmerston North, New Zealand). The resulting sequences were visualised and edited using Sequencher 4.9 software for Mac (Gene Codes Corporation, Ann Arbor, MI). Using Geneious [19], the four ambiguous regions of the assembled genome were edited, where necessary, to match the Sanger sequences.

Mapping and annotation

In order to check the de novo assembly, reads were aligned against the assembled genome using BWA [22] with default parameters. Only 19.6% of reads were successfully aligned, but this was sufficient to give a mean coverage of 400×. This mapping enabled us to resolve some short regions of ambiguous sequence in the assembly. The final complete chloroplast genome sequence was annotated using DOGMA [23] and through comparison to published complete chloroplast genome sequences available through GenBank [24].

Results

DNA sequencing template for karaka

The relative quantity of chloroplast DNA in samples of total gDNA, cpDNA and RCAcpDNA preparations was determined by quantitative PCR (Figure 2). The enriched cpDNA sample had 2.6-fold higher levels of chloroplast DNA compared to a standard gDNA preparation prior to RCA and 2.2-fold after RCA. The purity of the DNA preparations was assessed by spectrophotometric A_{260}/A_{280} and A_{260}/A_{230} ratios. The gDNA and cpDNA samples had low ratios, indicating the presence of protein (A_{260}/A_{280} = 1.66 and 1.69, respectively) and other contaminants such as carbohydrates and phenolics (A_{260}/A_{230} = 1.49 and 1.28, respectively). RCA of the cpDNA-enriched sample substantially increased the quantity and quality of template DNA (A_{260}/A_{280} = 1.75, A_{260}/A_{230} = 2.20).

Sequencing and assembly of the karaka chloroplast genomes

Paired-end sequencing of the RCAcpDNA template in a single lane on an Illumina GAII flow cell produced 1.84 and 1.76 million reads for the Chatham Islands sample and Kermadec Islands sample respectively. The Chatham Islands sample was assembled de novo as described in the methods section. The Kermadec Island sample was then mapped to this assembly.

The most useful assembly was achieved for the Chatham Islands sample, with reads trimmed to 50 bp, coverage cut-off of 9 and expected coverage of 40. While

Atherton *et al. Plant Methods* 2010, **6**:22
http://www.plantmethods.com/content/6/1/22

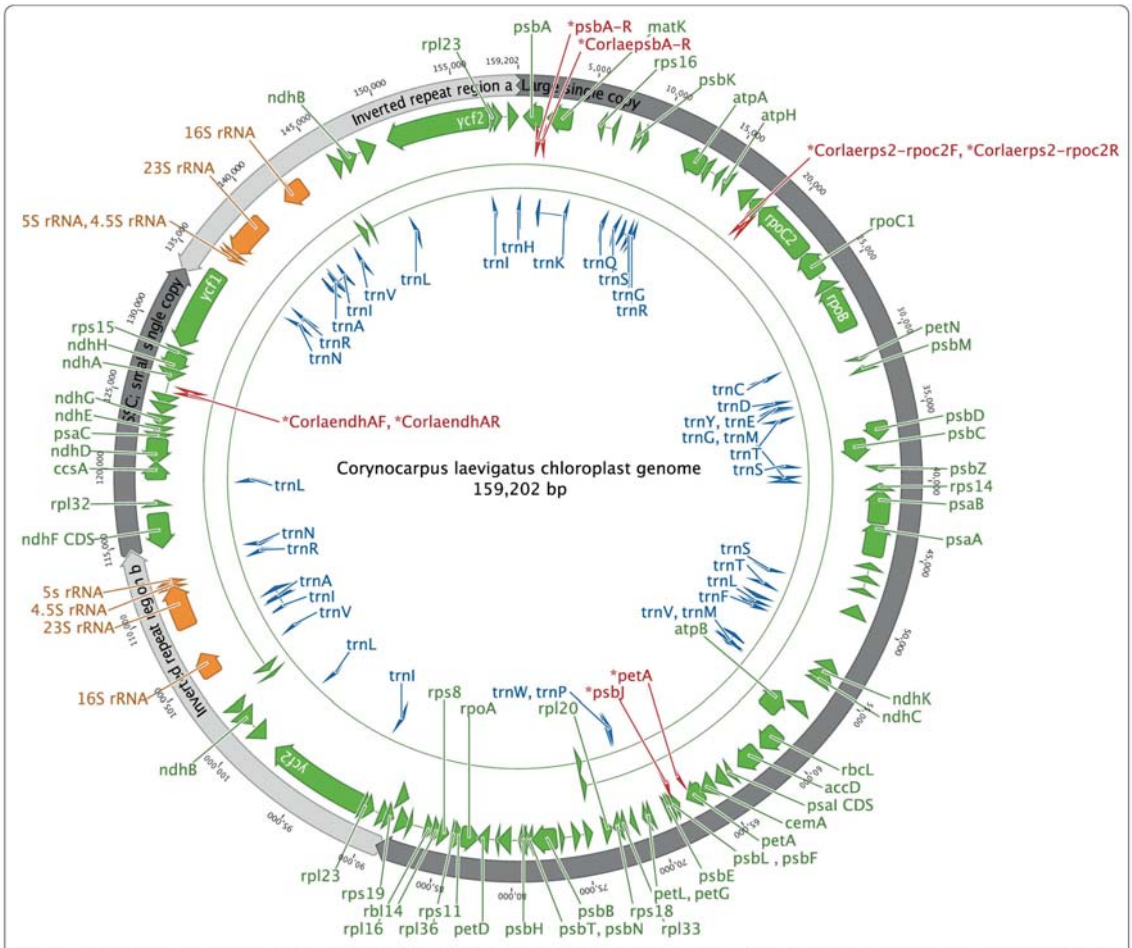

Figure 1 Complete gene map of the entire *Corynocarpus laevigatus* chloroplast genome. Arrowheads depict the direction of transcription. Binding sites for the primers used to generate the PCR products for resolving homopolymers are indicated in red. This figure was produced using Geneious [19].

some of our other assemblies had higher overall N50 values or longer maximum contig lengths, these improved statistics did not reflect any real improvement in the assembly, as the large single-copy region was merged with part of the inverted repeat to form a single long contig at the expense of a more fragmented assembly of the remainder of the inverted repeat and small single-copy region.

The optimal assembly parameters produced a total of 13 contigs, four of which could be mapped to the *Cucumis sativus* chloroplast genome. These four contigs ranged from 7,857 bp to 88,955 bp in length and covered the entire chloroplast genome. The nine remaining contigs were much shorter, ranging from 81 bp to 669 bp. These were checked against the GenBank nucleotide database using the web-based BlastN algorithm [25], and the only significant alignments found were to nuclear ribosomal DNA sequences.

Of the four contigs mapping to the *C. sativus* chloroplast genome, one mapped to the inverted repeat region, one to the large single-copy region, and two to the small single-copy region. The overlaps between contigs at all four junctions between inverted repeat and single-copy regions were 40 bp long, indicating that contig extension was interrupted by the ambiguity of the overlap rather than by insufficient coverage. The overlap between the two contigs that formed the small single-copy region consisted of a polyA-polyT homopolymer. The contig corresponding to the large single-copy region contained six short (1-49 bp) stretches of

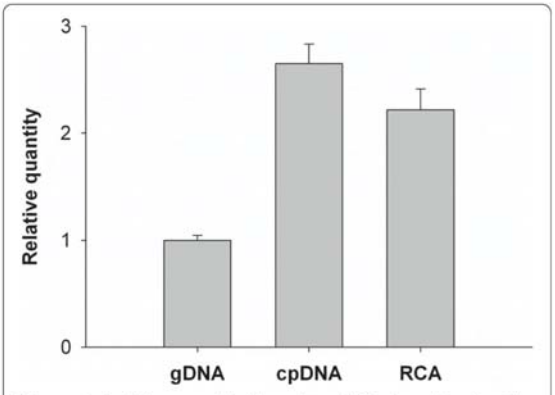

Figure 2 Enrichment of chloroplast DNA from karaka. The relative quantity of chloroplast DNA was determined in preparations of genomic DNA (gDNA), chloroplast-enriched DNA (cpDNA) and RCA amplified chloroplast-enriched cpDNA (RCA). Quantitative PCR amplification of the plastid encoded *psbB* gene was compared to amplification of nuclear 18S *rRNA*. The mean of four technical replicates ± SE has been reported as fold-differences relative to gDNA.

ambiguous bases where Velvet was unable to resolve the sequence due to mono- or dinucleotide repeats. Three of these stretches were resolved by mapping the reads to the assembled sequence as described in the methods. The other three stretches, along with the overlap between the two contigs that made up the small single-copy region, were checked by PCR amplification and Sanger sequencing.

The final assembled chloroplast genome sequence (shown in Figure 1) was checked by mapping the original Illumina reads against the assembled sequence. A total of 344,475 of 1.76 million reads (19.6%) were successfully aligned, suggesting that approximately 80% of the DNA sequenced was of nuclear or mitochondrial origin.

Discussion

We have shown that the modified chloroplast isolation protocol produced DNA template sufficiently enriched for chloroplast sequence to allow de novo assembly of the chloroplast genome. Comparison of the two genomes indicated high fidelity with less than 0.002% error. Whilst RCA of the cpDNA marginally reduced the final ratio of cpDNA/gDNA in the enriched sample, the purity of the DNA was of a higher quality for Illumina sequencing.

The coverage cut-off parameter of the Velvet assembler was crucial for successful assembly, as it allowed the chloroplast sequence reads to be assembled without interference from nuclear sequence. Although over 80% of reads failed to align to our assembled chloroplast genome, and are likely to be of nuclear origin, the much greater size of the nuclear genome means that these reads were present at much lower coverage than the chloroplast reads. A notable exception is nuclear ribosomal DNA, which is present in many copies in the nuclear genome, thus its coverage was comparable to that of the chloroplast genome in our enriched sample.

The lower copy number of the nuclear genome compared to the chloroplast genomes means that nuclear copies of chloroplast DNA sequences are very unlikely to affect our assemblies. In contrast, nuclear-encoded chloroplast DNA may be more difficult to distinguish from chloroplast-encoded sequences if amplified by chloroplast DNA primers. Thus, this is potentially another advantage of the approach we have used for determining complete chloroplast genome sequences.

Finally, although de novo assembly was a feature of our protocol, the availability of a related reference genome did help with our final assembly, allowing us to separate contigs derived from chloroplast DNA from the few short contigs of nuclear origin. This was helpful for determining the arrangement of chloroplast contigs.

Conclusions

We have successfully applied a whole genome sequencing approach to determine the complete chloroplast genome sequence of karaka. We have also applied this approach more recently to a range of New Zealand seed plants (gymnosperms and angiosperms: herbaceous and woody plants), sequencing up to three chloroplast genomes per GAII flow cell lane. Thus we are confident that the approach that we describe here for karaka provides a fast and efficient protocol for obtaining whole chloroplast genome sequences for seed plants.

The fully annotated chloroplast genome sequence of karaka (*Corynocarpus laevigatus*) from the Chatham Islands sample has been deposited in the GenBank database under accession number HQ207704.

Acknowledgements

We would like to acknowledge the help and support of the Hokotehi Moriori Trust Board of Rekohu/Chatham Islands and Ngāti Kuri and Te Aupouri, tangata whenua (indigenous inhabitants) of the Kermadec Islands. We would also like to acknowledge Te Ati Awa (Taranaki Whānui) in the Wellington and Hutt Valley regions of New Zealand, with whom we have ongoing consultations. We thank Peter de Lange (Department of Conservation), Rewi Elliot and Eleanor Burton (Otari Wilton's Bush, Wellington) for assistance with sample collection. Samples were collected under Department of Conservation permit WA-23814-FLO and Otari Wilton's Bush permit 145. Thanks also to Maurice Collins (MGS) for pipeline analysis, Trish McLenachan (IMBS, Massey University) and Jan Binnie (IMBS, Massey University) for laboratory assistance. This study was funded by the New Zealand Marsden Fund (contract numbers MAU050; MAU0709) and the Allan Wilson Centre for Molecular Ecology and Evolution. Robin Atherton acknowledges financial support from a Massey University Doctoral Scholarship, JP Skipworth Scholarship for Plant Biology, Hokotehi Moriori Trust Board, Intellectual Property in Cultural Heritage (IPinCH) and the Allan Wilson Centre for Molecular Ecology and Evolution. Bennet McComish

acknowledges support from an Allan Wilson Centre for Molecular Ecology and Evolution Doctoral Scholarship. Peter Lockhart is supported by a James Cook Research Fellowship from the Royal Society of New Zealand.

Author details
[1]Institute of Molecular BioSciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [2]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [3]Massey Genome Service, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [4]Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand.

Authors' contributions
RAA helped with the design of the study, isolated chloroplasts, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments and submitted data to GenBank, as well as drafting, contributing to and editing the manuscript; BJM carried out de novo assembly and annotation of the genome, contributed to and edited the manuscript; LDS contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study. LB carried out sample preparation and sequencing on the Illumina GAII; NWA designed and performed qPCR assays and contributed to and edited the manuscript; PJL conceived of and designed the study, contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 25 June 2010 Accepted: 28 September 2010
Published: 28 September 2010

References

1. Gruenheit N, Lockhart PJ, Steel M, Martin W: **Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites.** *Molecular Biology and Evolution* 2008, **25(7)**:1512-1520.
2. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH: **Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm.** *Molecular Biology and Evolution* 2003, **20(9)**:1499-1505.
3. Knapp M, Stockler K, Havell D, Delsuc F, Sebastiani F, Lockhart PJ: **Relaxed molecular clock provides evidence for long-distance dispersal of *Nothofagus* (southern beech).** *Plos Biology* 2005, **3(1)**:38-43.
4. Stehlik I, Blattner FR, Holderegger R, Bachmann K: **Nunatak survival of the high Alpine plant *Eritrichium nanum* (L.) Gaudin in the central Alps during the ice ages.** *Molecular Ecology* 2002, **11(10)**:2027-2036.
5. Ingvarsson PK, Ribstein S, Taylor DR: **Molecular evolution of insertions and deletion in the chloroplast genome of *Silene*.** *Molecular Biology and Evolution* 2003, **20(11)**:1737-1740.
6. Golenberg EM, Clegg MT, Durbin ML, Ma DP: **Evolution of a noncoding region of the chloroplast genome.** *Molecular Phylogenetics and Evolution* 1993, **2(1)**:13.
7. Powell W, Morgante M, Andre C, Mcnicol JW, Machray GC, Doyle JJ, Tingey SV, Rafalski JA: **Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome.** *Current Biology* 1995, **5(9)**:1023-1029.
8. Cronn R, Liston A, Parks M, Gernandt D, Shen R, Mockler T: **Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology.** *Nucleic Acids Research* 2008, **36(19)**:e122-e122.
9. Dhingra A, Folta KM: **ASAP: Amplification, sequencing & annotation of plastomes.** *BMC Genomics* 2005, **6**:176.
10. Jansen RK, Raubeson LA, Boore JL, DePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, *et al*: **Methods for obtaining and analyzing whole chloroplast genome sequences.** *Molecular Evolution: Producing the Biochemical Data, Part B* 2005, **395**:348-384.
11. Aronsson H, Jarvis P: **A simple method for isolating import-competent *Arabidopsis* chloroplasts.** *FEBS Letters* 2002, **529(2-3)**:215-220.
12. Jansen R, Raubeson L, Boore J, dePamphilis C, Chumley T, Haberle R, Wyman S, Alverson A, Peery R, Herman S, *et al*: **Methods for obtaining and analyzing whole chloroplast genome sequences.** In *Methods in Enzymology: Molecular Evolution: Producing the Biochemical Data, Part B.* Edited by: Zimmer E, Roalson E. San Diego: Elsevier Academic Press Inc; 2005:**348-384**:896.
13. Dean FB, Nelson JR, Giesler TL, Lasken RS: **Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification.** *Genome Research* 2001, **11(6)**:1095-1099.
14. Dean FB, Hosono S, Fang LH, Wu XH, Faruqi AF, Bray-Ward P, Sun ZY, Zong QL, Du YF, Du J, *et al*: **Comprehensive human genome amplification using multiple displacement amplification.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(8)**:5261-5266.
15. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Research* 2001, **29(9)**:e45.
16. Rozen S, Skaletsky HJ: **Primer3 on the www for general users and for biologist programmers.** In *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Edited by: Krawetz S, Misener S. Totowa, NJ: Humana Press; 2000:365-386.
17. Zhu L, Altmann SW: **mRNA and 18S-RNA coapplication-reverse transcription for quantitative gene expression analysis.** *Anal Biochem* 2005, **345**:102-109.
18. Zerbino D, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18(5)**:821-829.
19. Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A: **Geneious v4.7.** 2009 [http://www.geneious.com].
20. Shaw J, Lickey E, Schilling E, Small R: **Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III.** *American Journal of Botany* 2007, **94(3)**:275.
21. Winkworth RC, Grau J, Robertson A, Lockhart PJ: **The origins and evolution of the genus *Myosotis* L. (Boraginaceae).** *Molecular Phylogenetics and Evolution* 2002, **24(2)**:180-193.
22. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25(14)**:1754-1760.
23. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA.** *Bioinformatics* 2004, **20(17)**:3252-3255.
24. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2009, **37**:D26-D31.
25. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25(17)**:3389.

doi:10.1186/1746-4811-6-22
Cite this article as: Atherton *et al.*: **Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform.** *Plant Methods* 2010 **6**:22.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Robin Amber Atherton

Name/Title of Principal Supervisor: Professor Peter J. Lockhart

Name of Published Research Output and full reference:

"Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform."

Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ: Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 2010, 6:22.

http://www.plantmethods.com/content/6/1/22

In which Chapter is the Published Work: Chapter Three

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The candidate developed a chloroplast isolation protocol, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments, prepared amplicons for sequencing and edited sequences. The candidate also submitted data to GenBank, as well as drafting, contributing to and editing the manuscript

Candidate's Signature — 25 July 2013 (Date)

Principal Supervisor's signature — 26 July 2013 (Date)

GRS Version 3– 16 September 2011

4

SNP MARKERS FOR KARAKA ASSAYED USING HIGH RESOLUTION MELT ANALYSIS

4.1 CHAPTER OVERVIEW

This chapter is presented in the format of a scientific journal paper, ready for submission in Molecular Ecology. It begins by briefly discussing the study species, karaka (*Corynocarpus laevigatus* Forst & Forst) and then outlines the considerations and approaches for developing appropriate species-specific molecular markers for the particular scientific questions being asked. The focus is on the use of high-throughput sequencing as a method for scanning the whole chloroplast genome to uncover SNP variation in a species where genetic variation is low. This is followed by the application of high resolution melting (HRM) analysis to genotype an initial 60 accessions of karaka for multiples SNP markers, followed by a further 288, once the method was established.

In the context of these considerations, an evaluation of molecular methods for elucidating the history of karaka in Aotearoa/New Zealand (hereafter New Zealand) is given. A brief section on the relationships of karaka within New Zealand completes this chapter.

4.2 A NOTE ON ATTRIBUTION

This chapter is mostly my own work. However, the work was undertaken in collaboration with several researchers. This collaborative research includes the following people:

1. DAVID CHAGNÉ, PLANT AND FOOD RESEARCH, PALMERSTON NORTH.

David Chagné is a researcher at Plant and Food Research in Palmerston North. David is experienced in the use of HRM analysis for genotyping fruit trees of economic importance, such as apple. David and I worked together to optimise the HRM PCR conditions for 19 SNP markers. David's guidance was invaluable for the development of this method, and particularly with troubleshooting, manual binning of the results and determining profiles of individual markers.

2. TRISH MCLENACHAN, PLEB LAB MANAGER, MASSEY UNIVERSITY, PALMERSTON NORTH.

Trish McLenachan is the Laboratory Manager for the PLEB Laboratory in the Institute of Fundamental Sciences at Massey University in Palmerston North. Trish provided invaluable assistance during the testing stages of this method. Trish carried out some of the PCR reactions to test the validity of the HRM SNP calls and helped with preparation of amplicons for Sanger sequencing[1].

4.3 ABSTRACT

The sequence variability of the karaka chloroplast genome was investigated as a potential source for seed dispersal markers. The markers were then evaluated in terms of their potential for elucidating the history of karaka translocation during Māori settlement of New Zealand. Long-range polymerase chain reaction (LRPCR) products were amplified from the chloroplast genome of 22 individuals and subsequently sequenced using Illumina next-generation sequencing[2] technology, which enabled the identification of 48 putative chloroplast single nucleotide polymorphisms (SNPs). Sanger sequencing on the same 22 accessions validated 16 of these detected SNPs. This chapter evaluated the high resolution melting (HRM) technique as an accurate, sensitive and fast PCR-based method to screen SNP variations in the chloroplast genome of karaka. The newly developed HRM assays were validated and compared to traditional Sanger sequencing on a subset of 60 accessions before applying HRM assays to a larger sample set. A set of six SNP markers defined five haplotypes in 348 accessions, and a seventh SNP defined a

[1] Sanger sequencing – a very accurate capillary electrophoresis-based sequencing method, developed by Sanger *et al.* (1977) and considered the 'gold standard' in sequencing for 25 years. Sanger sequencing is a combination of dideoxy-based termination chemistry, fluorescent labeling, capillary separation, and computerised laser detection of DNA fragments.

[2] The Illumina next-generation sequencing reaction is conducted in a massively parallel fashion on several million different template molecules spread out on a flow cell.

further haplotype when tested against a smaller subset of accessions. Geographic distribution of these six haplotypes was evaluated to provide insight into the extent of human-mediated dispersal of karaka in New Zealand.

4.4 INTRODUCTION

4.4.1 BACKGROUND

New Zealand was the last substantial landmass to be settled by prehistoric people (Anderson, 1991) approximately 800-1000 years ago (Wilmshurst *et al.*, 2008) by the ancestors of Māori. Māori are known to have transported valuable goods such as obsidian and greenstone, which has led to the inference of linkages between some regions (Anderson & McFadgen, 1990; Belich, 1996). However, knowledge of pre-European interactions between iwi (tribes) of different regions is far from complete. Around the world, molecular studies of human-dispersed organisms have proved invaluable for tracing human migration patterns (Matsuoka *et al.*, 2002; Matisoo-Smith & Robins, 2004). Similarly, genetic relationships between populations of karaka have the potential to be used as additional indicators for prehistoric movement of Māori and Moriori (the latter being the indigenous inhabitants of Rekohu/Chatham Islands, hereafter Chatham Islands) around New Zealand.

Pacific voyagers cultivated and translocated a number of crop species around the region (Whistler, 1991). However, owing to New Zealand's cooler climate, it is unlikely many of these tropical crops survived (Leach & Stowe, 2005). Compensating for the loss of introduced crops, Māori cultivated a number of plants they discovered in New Zealand. This study focuses on one such plant: karaka (*Corynocarpus laevigatus*), an important staple winter food. The family Corynocarpaceae consists of five species of trees found in tropical to warm temperate areas in the southwest Pacific. Karaka is confined to mainland New Zealand and its offshore islands, Chatham Islands and the Kermadec Islands (Molloy, 1990). Karaka is a tall, spreading evergreen tree growing to a height of approximately 15 m found mainly in coastal regions throughout New Zealand (Clarke, 2007). The fruit are small drupes, up to 5 cm in length, with smooth skin that turns orange when ripe. The flesh of the drupe covers a tough fibrous endocarp, inside which is a highly prized seed.

Karaka is an entomophilous[3] tree whose pollen is severely under-represented in the palynological record (Dodson, 1976). Mildenhall (1994) suggests karaka was either a recent introduction to Chatham Islands, or it simply does not preserve well. Holt (2009) noted that pollen of karaka was not recorded from any sampling site on Chatham Islands during their palynological study, even though karaka is a major part of lowland broadleaf woodlands on the island group. Karaka pollen was found in pollen cores at two sites in the Mimi and Waitoetoe catchments in Taranaki (Wilmshurst *et al.*, 2004). Its sudden appearance in the sections of the cores corresponding to the deforestation period and early Māori settlement period, suggest karaka did not grow historically in Taranaki, and was probably brought to the region by Māori and planted in recently deforested clearings (Wilmshurst *et al.*, 2004). In the Mimi and Waitoetoe catchments karaka are still present in small groves today (Wilmshurst *et al.*, 2004). Karaka pollen was also found in the Coromandel by Byrami (2002), corresponding to the same deforestation period in Taranaki.

Karaka was of great importance as a food to Māori in regions of New Zealand where introduced cultivated crops, such as kumara (sweet potato, *Ipomoea batatas*) and other sub-tropical plants foods such as hue (bottle gourd, *Lagenaria siceraria*), aute (paper mulberry, *Broussonetia papyrifera*), taro (*Colocasia esulenta),* uwhi (yam, *Dioscorea* species) and tī pore (Pacific cabbage tree, *Cordyline fruticosa*) were difficult to grow (Leach & Stowe, 2005).

Originally karaka was thought to have been restricted to the northern North Island. However, its occurrence in the southern North Island, the South Island, Chatham and Kermadec Islands is strongly associated with Māori and Moriori archaeological sites and considered to have resulted from translocations as part of its cultivation (Leach & Stowe, 2005). Traditional oral histories exist regarding the origins of some of these plant populations (Smith, 1893, 1900). The study of karaka phylogeography and past colonisation is therefore a useful tool to understand human migration during the settlement of New Zealand by Māori and Moriori in the last ten centuries.

[3] Entomophily – pollen is dispersed by insects

4.5 TRANSLOCATION OF KARAKA

Intentional translocations of karaka, in addition to natural range expansion of the species makes it difficult to infer the direction of movement and timing of dispersal events. Historic records are not available for a species that was likely the subject of several human-mediated dispersal events over the last 800-1000 years, which is the time ancestors of modern Māori arrived in New Zealand (Wilmshurst *et al.*, 2011). However, in New Zealand, there are oral histories that tell of the movement of karaka and of its importance as a food source for these founding people. Māori korero (oral histories) talk of the arrival of karaka in New Zealand on waka (voyaging canoes). The Aotea waka purportedly brought karaka to New Zealand (Smith, 1891), planting it at Aotea Harbour. More specifically, korero tells the story of the Aotea translocating karaka to Taranaki on the west coast of New Zealand's North Island, with Turi[4] planting a karaka grove in Patea (Smith, 1900). In another account written by Houston (1965), Turi made the final part of his journey from Aotea to Patea, in Taranaki, by foot. He sent Pungarehu ahead and instructed him to plant karaka seeds [brought on the Aotea canoe] all along the route to provide a plentiful supply of food. There is korero that mentions Kupe[5], too, brought karaka to Taranaki, planting the seed at Patea on the west coast and also at Mahia on the east coast of the North Island (Whatahoro, 1915).

The grove planted at Patea was of a type of karaka known as 'Oturu', the same type occurring at Nuhaka near Mahia, believed to have been taken there by Kupe. The Kurahaupo waka, too, claims to have brought the karaka tree to New Zealand and "the tree became the parent of all the East Coast trees." (Mitira, 1972). Buick (1903) mentions the karaka brought in the Kurahaupo canoe was a smaller kind than the kind the Aotea brought with them. The Nukutere waka brought specimens of karaka and tī (*Cordyline* sp.) with them (Best, 1902) and planted them at Waioeka in the Bay of Plenty region (Best, 1972). Similarly, the Takitumu waka introduced the tree, and Ruawharo took them to locations on Mahia Peninsula: Nukutaurua, Table Cape (Best, 1976) and Mahia (Best, 1977) (see Figure 1.9, Chapter 1 for mapped locations).

[4] Turi was the captain of the Aotea waka whose occupants became the ancestors of the Taranaki, Ngāti Ruanui, Ngā Rauru, and Wanganui tribes of the West Coast of New Zealand.
[5] Kupe – the first voyager to make contact with New Zealand from Hawaiki, the traditional Māori place of origin. He appears in many Māori oral histories

In Moriori oral tradition, Rangimata, one of the founding canoes, landed on the north coast of Chatham Islands at a place called Wairarapa and there karaka, which they also called *wairarapa*, was planted (Shand, 1896).

4.5.1 GENETIC STUDY

If the current karaka distribution was entirely natural, we would expect one of two scenarios, depending on the phylogeographic history of the species: 1) There would be many isolated karaka populations, widely-dispersed and with endemic[6] haplotypes, in regions other than those postulated to be glacial refugia (if there was widespread survival of karaka outside refugia during the last glacial maximum (LGM)); 2) levels of genetic diversity would be highest in the postulated refugia areas of Wardle (1963) (if karaka was restricted to refugia then expanded its distribution following the end of the LGM). If the karaka population was made up of translocated trees, the scene would be quite different. If the karaka currently south of 38°S derive solely from translocations then these populations would likely show reduced genetic variation compared to the putative natural populations further north, and not only would the translocated populations be less diverse than the source population, they would also be a subset of the genetic diversity of the source. However, because the northern North Island refugium occurs in the same region as the suggested natural range of karaka, it may be difficult to distinguish between a translocation origin and natural dispersal from a northern refugium following the LGM.

Additionally, if oral histories were accurate, i.e. multiple translocations, then a mixed distribution would be expected, reflective of those histories.

4.5.2 MOLECULAR METHODS

The approach taken for this work has been to examine single nucleotide polymorphisms (SNP) variation in the chloroplast genome of karaka. The aim was to elucidate the genetic relationship between trees growing in the natural range of karaka and putative translocated trees. A further aim of this work was to determine the extent to which karaka was domesticated, if at all. Karaka is an excellent model for studying this process of settlement because of its significance in the Māori diet.

[6] Endemic - unique to a defined geographic location

Nuclear DNA sequences have the advantage of providing evidence of both the maternal and paternal lineages. Levels of polymorphism for genomic DNA can be more suitable for analyses of intraspecific variation than organellar DNA (Doebley, 1992). However, the size and complexity of chloroplast genomes means they can contain structural and point mutations that can be used to study population-level processes (Cronn *et al.*, 2008).

Assessing genetic diversity in plants has become more sophisticated with the advent of high-throughput sequencing techniques. Although microsatellites are often the favoured option for these studies due to their multi-allelic states, development and genotyping of large numbers of accessions can be expensive. Other marker types have been used to study plant variation including amplified fragment length polymorphisms (AFLP), restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), diversity arrays technology (DArT) and allozymes. However, an optimal marker for this study was one suitable for tracing seed dispersal. Organellar DNA such as chloroplast DNA (cpDNA) can be used to study population-level processes (Cronn *et al.*, 2008). cpDNA sequence variation is also a major source of data for inferring plant phylogenies (Shaw *et al.*, 2005). The (predominantly) maternal inheritance of cpDNA (see Chapter 3.1) is useful to trace gene flow in populations, such as seed dispersal. Therefore, cpDNA markers provide information on past changes in species distribution that is unaffected by subsequent pollen movements.

The chloroplast genome has been used to search for markers for the study of domestication in apple (*Malus*) (Coart *et al.*, 2006), cowpea (*Vigna unguiculata*) (Feleke *et al.*, 2006), *Brassica oleracea* (Zhang *et al.*, 2012), sunflower (*Helianthus annus*) (Wills, 2006) *Cucurbita* (Zheng *et al.*, 2013) and *Linum* (Fu & Allaby, 2010) amongst many others. However, the work in this chapter could be used to determine whether the karaka chloroplast genome has the resolving power and suitability for the study of recent events of a plant species' early domestication history.

The advent of high-throughput second-generation sequencing has enabled surveying the set of nucleotide variations within whole genomes, including cpDNA (Cronn *et al.*, 2008). A method to detect polymorphisms is by comparative sequencing of stretches of PCR-amplified non-coding cpDNA (Weising *et al.*, 2005) to mine for single nucleotide

polymorphisms (SNPs) among accessions of the same species. SNPs are single base change differences between homologous DNA. SNPs are a common DNA sequence variation occurring in both plant nuclear (Newcomb *et al.*, 2006) and plastid genomes (Diekmann *et al.*, 2008), making them ideal for molecular marker development for phylogenetic analysis and genetic diversity studies.

Using SNPs, chloroplast haplotype variation can be detected in individual populations of a species ultimately identifying cytoplasmic gene pools (Diekmann *et al.*, 2008). However, intra-specific polymorphisms in the chloroplast genome can be difficult to identify in recently diverged populations (McCauley, 1995). Even once SNPs have been discovered, methods for large-scale genotyping of hundreds of samples must be developed. High resolution melting (HRM) analysis is a PCR-based technique useful for the genotyping of individuals for base mutations such as SNPs. The technique is based on the melting behavior of the double-stranded DNA PCR product, using high fidelity intercalating dyes. Differing melting curves are the result of a different sequence of bases, or a mutation at one or more base positions. It has been popular in studies of humans and been used successfully for genotyping crops of economic importance such as apple (Birky, 1995; Chagné *et al.*, 2008), cherry (Miller & Gross, 2011), and almond (Wu *et al.*, 2008). Whilst HRM has been used for genotyping using nuclear markers, it has only recently been applied to chloroplast markers where it was used as a screening method to detect SNP variants in the *atpB* gene and the upstream intergenic spacer in *Brassica* (Yan *et al.*, 2012). HRM was used to identify haplotypes in *Arenaria ciliata* and *A. norvegica* for phylogeographic analysis using the chloroplast *rps16* intron (Petersen *et al.*, 2012) and alongside nuclear SNPs to distinguish between species of *Capsicum* for the purpose of species classification (Allaby & Brown, 2003).

Atherton *et al* (2010) (Chapter 3 in this thesis) sequenced the chloroplast genomes of two accessions of karaka using Illumina GAII technology. Whilst the two accessions were from distant offshore locations in New Zealand (Kermadec Islands and Chatham Islands), they only showed two sequence differences (in the *ndhA* intron and the *psbB* gene). This finding might suggest either little or no sequence variation among extant karaka or a recent history of translocation between these populations and mainland populations. The research in this chapter sought to distinguish these alternative hypotheses by assaying for sequence variation in a greater cross-section of karaka

accessions from New Zealand. It describes results obtained by comparing long-range PCR products from multiple accessions. This approach identified additional polymorphisms, which were validated using Sanger sequencing. Having done this, the potential of HRM profiling was evaluated as a method of rapid, low-cost screening of karaka accessions. Initially, 60 accessions were tested and once established, the method was applied to a further 288 accessions, which resulted in six genotyped chloroplast SNP markers.

Figure 4.1 provides an overview of the methods developed and used in this chapter.

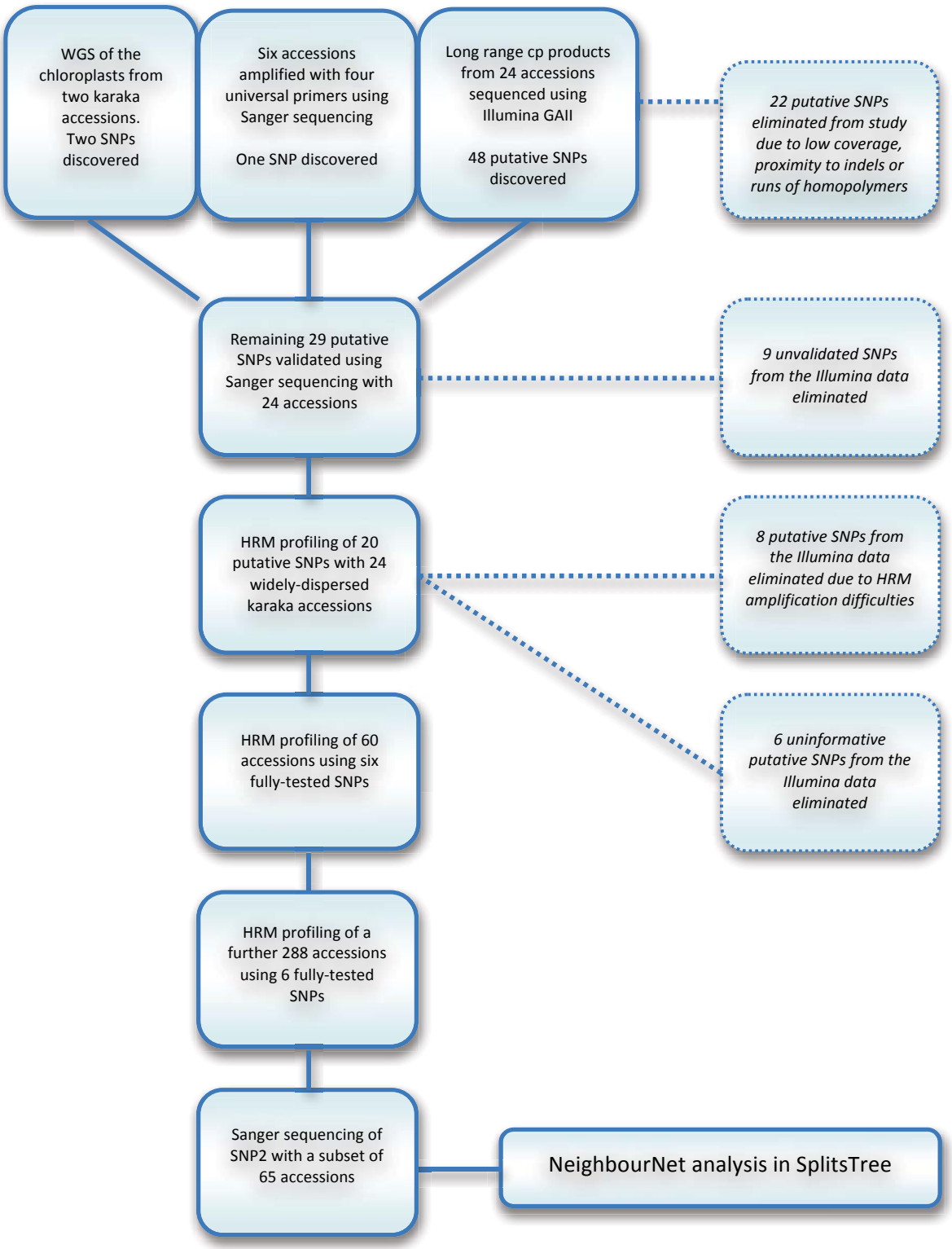

FIGURE 4.1: Methodology used to identify SNP markers in the karaka chloroplast genome. SNPs were determined using three methods: whole genome sequencing, universal primers and long range PCR. Unsuitable markers were eliminated from the study (boxes with dashed lines) at each stage for the reasons stated. In total, 51 putative markers were discovered and seven markers were used in the final suite. (WGS = whole genome sequencing).

4.6 MATERIALS AND METHODS

4.6.1 SAMPLE COLLECTION AND DNA EXTRACTION

To determine the distribution of this species, and thus select representative sampling sites, karaka distribution was mapped using information from Stowe (2003) and the herbarium records of The Auckland Museum, Landcare Research, Waikato University, Victoria University, Massey University and The Museum of New Zealand Te Papa Tongarewa (Fig. 4.2). Further information was gathered from meetings arranged with iwi (Māori tribe) leaders.

Samples were obtained from populations of karaka throughout its distribution in the North and South Islands of New Zealand (Table A2.1 in Appendix 2). Herbarium samples and DNA leaf samples were collected in the field. In some cases, samples were collected on private land with permission from the landowners, and from known provenance trees in cultivation at native plant garden, such as Otari Native Botanic Garden/Wilton's Bush in Wellington (sample 1162). An initial sampling targeted 60 karaka accessions that were a good representation of the species' distribution, including two large offshore island groups (Kermadec and Chatham Islands, ~800 km and 680 km from mainland New Zealand, respectively). Once the method was established, sampling was broadened to include a further 288 accessions.

DNA was obtained from silica-dried leaf tissue or fresh leaf tissue. The fresh leaf tissue was snap-frozen in liquid nitrogen and powdered and silica-dried leaves were milled using a MagnaLyser with 2 mm zirconia beads (Biospec) and genomic DNA extracted using a modified CTAB protocol (Doyle & Doyle, 1987). The remainder of each branchlet was kept as a voucher specimen, with representatives accessioned into the Wellington (WELT) and Auckland (AKL) herbaria.

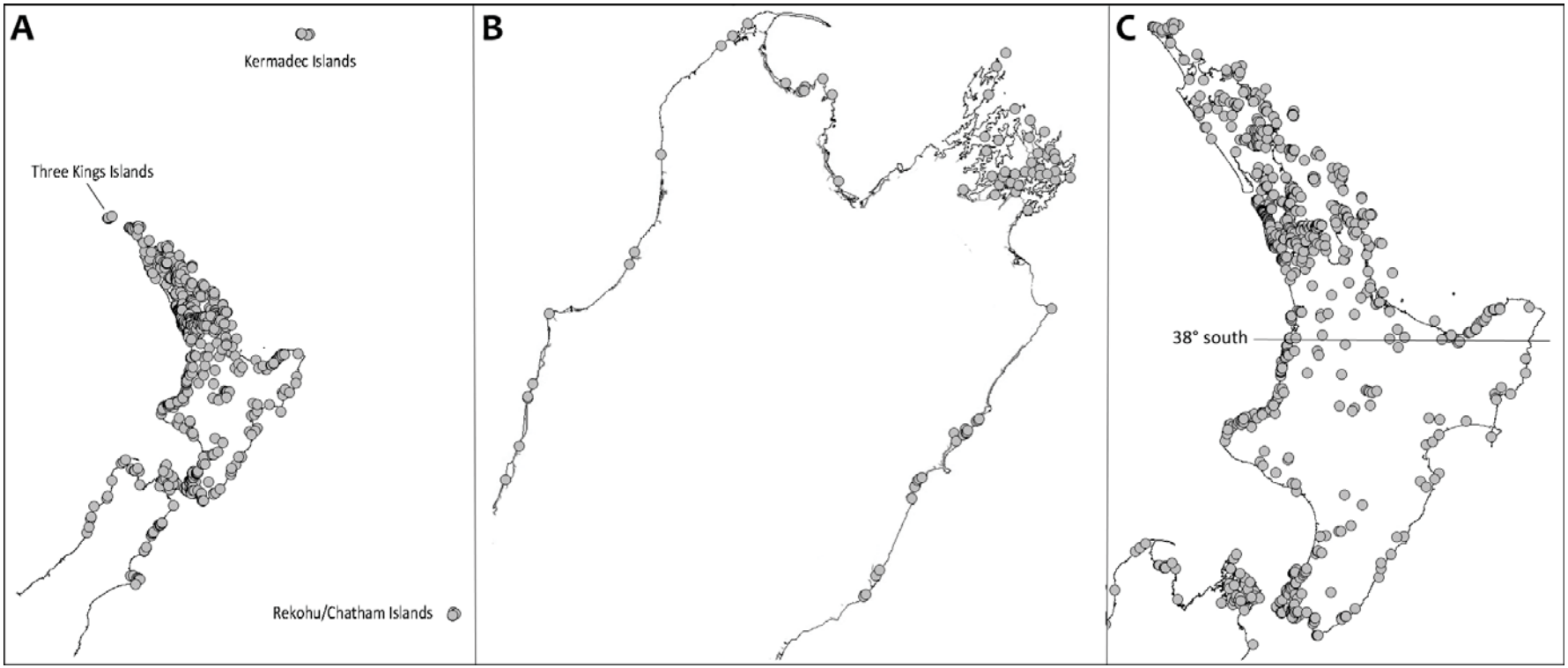

FIGURE 4.2: *Corynocarpus laevigatus* distribution showing (A) the distribution of karaka across New Zealand including the Kermadec Islands to the north and the Chatham Islands to the east; (B) the distribution of karaka in the South Island of New Zealand; and (C) the distribution of karaka in the North Island of New Zealand showing locations of inland populations. All maps are based on records in the AK, CHR, NZFRI, and WELT herbaria (abbreviations follow Holmgren *et al.* (1990)).

4.6.2 PREPARATION OF SHORT-RANGE AMPLICONS FOR SANGER SEQUENCING USING UNIVERSAL PRIMERS

Prior to assaying the chloroplast genome using long-range polymerase chain reaction (PCR), regions of hypervariability studied by others (Taberlet *et al.*, 1991; Hasebe *et al.*, 1994; Demesure *et al.*, 1995; Shaw *et al.*, 2005; Shaw *et al.*, 2007) were first examined using primers found to be conserved across angiosperms (Table A3.1, Appendix 3). Sequences of the chloroplast regions *rpl32-trnL*, *trnQ-5'-rpS16*, *3'trnV-ndhA, psbD-trnT*$^{GCU-R}$ and *trnfM-trnS* were obtained from six geographically isolated samples (Table 4.1) using the universal primers *rpl32-trnL*, *trnQ-5'-rpS16*, *3'trnV-ndhA* and *psbD-trnT*$^{GCU-R}$ (Shaw *et al.*, 2007) and *trnfM-trnS* and PCR programmes in (Shaw *et al.*, 2005). Amplification products were purified by digestion with 0.2 U shrimp alkaline phosphatase (SAP, USB Corp.) and 1 U exonuclease I (ExoI, USB Corp.) at 37°C for 30 min, followed by inactivation of the enzymes at 80°C for 15 min. Sequencing was performed in both directions with the ABI Big Dye™ Terminator Version 3.1 Ready Reaction Cycle Sequencing kit in a Biometra thermal cycler following the manufacturer's protocol. Unincorporated fluorescent dNTPs were removed using CleanSEQ (Agencourt), following the manufacturer's protocol, and capillary separation was subsequently undertaken at the Massey Genome Service, Palmerston North. Sequences were edited and aligned using Sequencher 4.9 (GeneCodes Corporation).

Table 4.1: Geographic location of six karaka samples sequenced with universal primers.

Sample no.	Location	Geographic coordinates							
1012	Mapito Rd, Swamp forest Chatham Islands	S	43	47	15.3	E	176	49	8.07
1035	Bledisloe Park, Massey University, Palmerston North	S	40	22	59.39	E	175	37	9.84
4724	Karekare Beach, Waitakere Ranges	S	59	-	-	E	174	28	-
4799	Matapouri, Whale Bay	S	35	34	-	E	174	30	-
5042	Karikari Peninsula, northern end of Tokerau Beach	S	34	52	-	E	-	173	23
5314	Hokianga Harbour, near the Waimamaku River mouth	S	35	35.25	-	E	173	25	-

4.6.3 SEQUENCING USING SPECIES-SPECIFIC PRIMERS

Given the low variation detected using Sanger sequencing and universal primers, a long-range PCR approach was adopted, similar to that used by Goremykin *et al.* (2003) and Cronn *et al.* (2008) to mine for SNPs. This method involved amplifying large portions of the non-repetitive components of the karaka chloroplast genome. Although this process can be uneconomical for multiple samples (Cronn *et al.*, 2008), an attempt was made, nevertheless, to amplify the large single copy (LSC) and small single copy (SSC) regions of the chloroplast genome from a discovery panel of 22 individuals chosen to broadly sample the geographic diversity of the species (denoted in Table A2.1 in Appendix 2). High molecular weight total DNA was isolated using a DNeasy® Plant Mini Kit (Qiagen). Long-range PCR primers ranging from ~3-12 kilo bases (kb) were designed using Primer3 (Rozen & Skaletsky, 2000), from the consensus chloroplast genome sequence from karaka accession RA83 (GenBank accession number HQ207704.1) (Atherton *et al.*, 2010). Primer pairs were designed to cover the large single copy (LSC) and small single copy (SSC) regions of the chloroplast genome with overlaps of ~500 bp (Table A3.1, Appendix 3).

Using the Expand Long-Range DNTPack PCR system (Roche), according to the manufacturer's protocol, and ~100ng starting DNA, amplifications were performed in 10 μL total reaction volumes, with the following thermocycler conditions: 93°C denaturation 2 min (1 cycle) followed by 93°C denaturation for 30 s, 60°C annealing for 30 s, and 68°C extension for 45 s per kb of sequence (10 cycles). This was followed by 93°C denaturation for 30 s, 60°C annealing for 30 s and 68°C extension for 45 s/kb with an increase of 20 s per cycle (24 cycles) followed by 68°C final extension step for 11 mins.

Reactions were confirmed by gel electrophoresis and PCR products were purified using SAP-EXO (USB Corporation, Cleveland, OH) as described above. PCR products were quantified using a NanoDrop ND-1000 spectrophotometer (ThermoScientific) and equimolar amounts were pooled for each accession to generate 1-5 μg of DNA.

4.6.4 ILLUMINA SEQUENCING, MAPPING AND VISUALISATION OF SNPS

Pooled DNA was prepared for sequencing by Massey University Genome Service (Palmerston North, New Zealand) using the Illumina sample preparation kit (Illumina Inc., San Diego, CA). A 100-bp paired-end read run was performed, in a single lane, on an Illumina Genome Analyser GAII (Illumina Inc., San Diego, CA), according to the manufacturer's instructions. The resulting images were analysed with the proprietary Illumina pipeline (software version 1.4). This resulted in approximately 13.2 million reads. Reads were subjected to dynamic trimming with a cut-off value of p=0.01 (Cox *et al.*, 2010). Dynamic trimming crops each read to its longest contiguous segment based on quality score for those reads.

The reads were aligned to the reference genome using the Burrows-Wheeler Algorithm (BWA) (Li & Durbin, 2009) using default settings (two mismatches permitted and a seed length of 32). BWA is a gapped aligner that allows for short insertions/deletions (indels) when matching the assembled contigs to the reference genome. The reference genome chosen was the full chloroplast genome sequenced as part of the research for Chapter 3 (Atherton *et al.*, 2010). As per standard procedure using BWA, first the reference genome is indexed, whereby repetitive patterns and locations are stored in a database format. This database is held in the memory and the reads are compared against it. The output is an SAI formatted file, which is then converted to the more conventional SAM format. The resulting SAM files can be read by Tablet (Milne *et al.*, 2010) a graphical viewer for next-generation sequence assemblies and alignments. Putative SNPs in the mapped reads were visualised using this software. However, this is not a recommended viewer for SNP detection, though for the small number of SNPs in this study, it was sufficient for this purpose.

For the mix of karaka accessions, the SNP sites were determined based on the number of reads supporting that base call. SNPs were classified according to minor allele frequency (MAF) (the fraction of the total alleles of the given marker that are minor alleles, which is presented as a fraction: MAF = minor allelic count/total allelic count).

4.6.5 SANGER-BASED SNP VALIDATION

Sanger sequencing was used, for the same discovery suite of 22 accessions, to confirm the identity of true SNPs and to identify false-positive SNPs, as described in Whittall *et al.* (2010). Primer pairs flanking putative SNPs were designed using Primer3 (Rozen & Skaletsky, 2000) (Table A3.1, Appendix 3). Primer pairs amplifying multiple putative SNPs within amplicons <1 kb in length were preferentially selected. Individual PCR amplifications were performed in 10 µl volumes using the following PCR protocol: template denaturation at 94°C for 3 min followed by 35 cycles of denaturation at 94°C for 30 s, primer annealing at 50°C for 30 s, and primer extension at 68°C for 45 s per kb of sequence; followed by a final extension step of 10 min at 68°C. Sequences were edited and assembled using Sequencher 4.9 (GeneCodes).

4.6.6 HIGH RESOLUTION MELTING PCR DESIGN AND OPTIMISATION

Primer pairs were designed for HRM (Table A3.1, Appendix 3). SNPs validated by Sanger sequencing were subjected to an HRM validation trial using the same 22 accessions as used in section 4.6.3 to determine suitability of the marker for HRM genotyping.

DNA stocks were diluted using a Nanodrop ND1000 spectrophotometer (ThermoScientific) and DNA template samples were diluted to 10 ng/µl. Pairs of primers flanking each SNP were designed to amplify DNA fragments 50-150 bp using Primer3 (Rozen & Skaletsky, 2000). For primer searching, the length of the primers was set between 18 and 25 bp, the primer annealing temperature (T_a) was set at 57.0 ± 5.0°C and with a minimum GC content of 27%. All primer pairs amplifying a single product of 150 bp or less were used to test for polymorphism between the two SNP variants using HRM analysis methodology as described in Chagné *et al.* (2008) but with minor modifications as follows: PCR was performed in a total volume of 10 µl containing ~10 ng of template DNA, 1× HRM master mix (Roche Applied Science), 2.5 mM $MgCl_2$, 300 nM forward and reverse primers. HRM were performed on a Roche LightCycler® 480 (Roche Applied Science). The PCR parameters used were an initial denaturation step of 95°C for 5 min, followed by 45 cycles of 95 for 10 sec, 55°C for 30s and 72°C for 15s. Following amplification, the samples were heated to 95°C for 1 min (ramp rate 4.4°C/s) and then cooled to 40°C for 1 min.

Melting curves were generated with continuous fluorescence acquisition during the final ramp from 65°C to 95°C at 1.0°C/s with a 40°C cooling 30 sec, and the resultant fluorescence data were processed using the LightCycler480® software (version 1.5.0.39; Roche Applied Science) (Anderson & McFadgen, 1990). Six markers produced distinctive HRM profiles and these were then screened against 60 karaka accessions, chosen for their geographic location or cultural importance (Table A2.1, Appendix 2). HRM genotyping results were validated by Sanger sequencing, performed with the ABI PRISM Big Dye Terminator cycle sequencing kit version 3.1 on an ABI 3730 DNA sequencer at Massey Genome Service, Palmerston North. These SNP regions were also sequenced in all other species and subspecies of *Corynocarpus* previously studied (Wagstaff 2006).

4.6.7 Phylogenetic analyses

HRM profiles for each accession at each of six loci were tabulated and converted to a nexus file (Appendix 9 on CD). Data from genotyped accessions were analysed using NeighborNet in SplitsTree (v4.0) (Huson & Bryant, 2006). Conflict in the data, potentially indicating genotyping errors, was visualised in this network, with conflict represented as a reticulation in the network. Sanger sequencing was used as the gold standard for the detection of SNPs in this study as it has the advantage of being able to identify the exact mutation. Therefore, where potential errors existed, Sanger sequencing was used to check HRM base calls.

4.6.8 Using previously discarded SNPs for further resolution in the data set

A subset of the accessions selected for HRM analysis was used to determine the utility of a further SNP that was not suitable for HRM due to amplification difficulties. A section of the *ndhA* gene was amplified using the primers CorLaeSNP002F and CorLaeSNP002R. PCR conditions followed those in Shaw *et al.* (2005). Sanger sequencing was used to further distinguish between chlorotypes 1, 2 and 3, with the intention of increasing the resolution in the HRM data set. A subset of the 348 accessions (Table A6.1, Appendix 6) was selected for genotyping with SNP2. Amplicons were sequenced using Sanger sequencing, performed with the ABI PRISM Big Dye Terminator v.3.0 cycle sequencing kit version 3.1 on an ABI 3730 DNA sequencer at Massey Genome Service, Palmerston North, according to the manufacturer's protocols.

4.6.9 COMPARISON WITH SPATIAL AND CLIMATE DATA OF THE DISTRIBUTION OF KARAKA

Stowe (2003) used a combination of climate profiling, and the association of karaka with archaeological sites, to uncover the extent to which the current distribution of karaka is determined 1) by the environment, and 2) by human-mediated dispersal. His comprehensive study grouped karaka accessions into two types: cultural and unknown. Cultural karaka were those strongly associated with archaeological sites such as pa, middens, kumara pits, terraces and walls, and found growing (or recorded as growing at the time of the archaeological study) within 500 m of a registered archaeological site. Of 805 records of the occurrence of karaka, 82% were classed as cultural and the remainder was unknown.

Climate profiling was also used to determine the natural and translocated range of the species (Stowe, 2003). There were significant differences in the climate profiles of cultural and unknown karaka accessions with the climate profile of cultural accessions being similar to that of kumara, and the climate profile of unknown accessions comparable to other broadleaved trees of tropical affinities currently restricted to the northern North Island (eg. *Litsaea calicaris*, *Weinmannia silvicola,* and *Beilschmiedia tarairi*).

Data from Stowe (2003) were plotted onto a map of New Zealand and compared to chloroplast haplotype distribution (Figures 4.6 and 4.7, respectively).

4.7 RESULTS

Sequencing and subsequent analysis of long-range PCR products, across 22 karaka accessions, showed low sequence variation but identified six SNPs and five distinct chlorotypes. All Chatham Island accessions assayed were identical and matched accessions from several mainland locations. The Kermadec Island was also only represented by one chlorotype. HRM analyses provided a rapid approach, however, in some cases assignments were not unambiguous.

4.7.1 INITIAL CHLOROPLAST INVESTIGATIONS USING UNIVERSAL PRIMERS

Three of the six universal primer pairs successfully amplified the regions *rpl32-trnL*, *trnQ-5'-rps16* and *psbD-trnT*$^{GCU-R}$ in the test accession (accession number 1035). The *rpL32-trnL* region was discarded owing to long mononucleotide runs. The *psbD-trnT*$^{GCU-R}$ and *trnQ-5'-rps16* intergenic regions were tested against five further geographically isolated karaka samples. One SNP was found in the *trnQ-5'-rps16* intergenic region at position 7418 in the karaka chloroplast genome. The *psbD-trnT*$^{GCU-R}$ showed no variation.

4.7.2 SNPS

In total, 48 putative SNPs were discovered using long-range PCR followed by Illumina GAII sequencing (Figure 4.3). Of the 48 SNPs, 24 were classified as very common (MAF 0.5), seven as common (MAF 0.1) and ten as rare SNPs (MAF < 0.1). Fourteen SNPs were discovered in three regions: the *psbE-petL* intergenic region (5), *rpl32-trnL* gene (5) and *ycf1* gene (4). Twenty-six SNPs were located in intergenic regions and 22 SNPs were located in genes (see Table 4.1, Appendix 4). In total, 16 SNPs were validated using Sanger sequencing. The *petL-petG* region contained what appeared to be six common putative SNPs. Upon amplification with universal primers (Shaw *et al.*, 2007) against 14 accessions, this region did not contain the variation suggested by the Illumina data. Sixteen SNPs were validated using Sanger sequencing and selected for HRM testing along with the three SNPs from our initial chloroplast investigations. Of the nineteen SNPs selected for HRM analysis, fifteen were transversions[7] and four were transitions. Of the fifteen transversions, twelve had an minor allelic frequency[8] (MAF) of >1.0, one of 0.5 and two of <0.1, when visualised in Tablet. Of the four transitions, two had an MAF of >1.0, one of 0.5 and one of <0.1.

4.7.3 HRM MARKER OPTIMISATION

Of the nineteen SNPs subjected to HRM amplification, six showed distinct melting profiles. The remaining thirteen were excluded for two reasons: either the marker was

[7] Transversions are interchanges of purine for pyrimidine bases (A⇔C, C⇔G, G⇔T and A⇔T) and transitions are interchanges of purines (A⇔G) or of pyrimidines (C⇔T).
[8] Minor allelic frequency (MAF) – minor allelic count/total allelic count. The number of reads which show a polymorphism/total number of reads.

uninformative, or there was inconsistent or low PCR amplification and presence of primer-dimers interacting with the amplicon melting profiles. The inconsistencies in PCR amplification could be attributed to DNA quality or quantity. The six SNPs were SNP1, 3, 8, 16, 41 and 49, which mostly corresponded to very common SNP (MAF > 0.5) in the detection set. DNA of all 60 accessions were 'spiked-in' 50:50 with one accession (RA123) to allow the formation of heteroduplexes, which usually exhibit a very distinctive shape and therefore melt curves would be easier to distinguish between. This was the method used for SNPs 3, 8, 16 and 49.

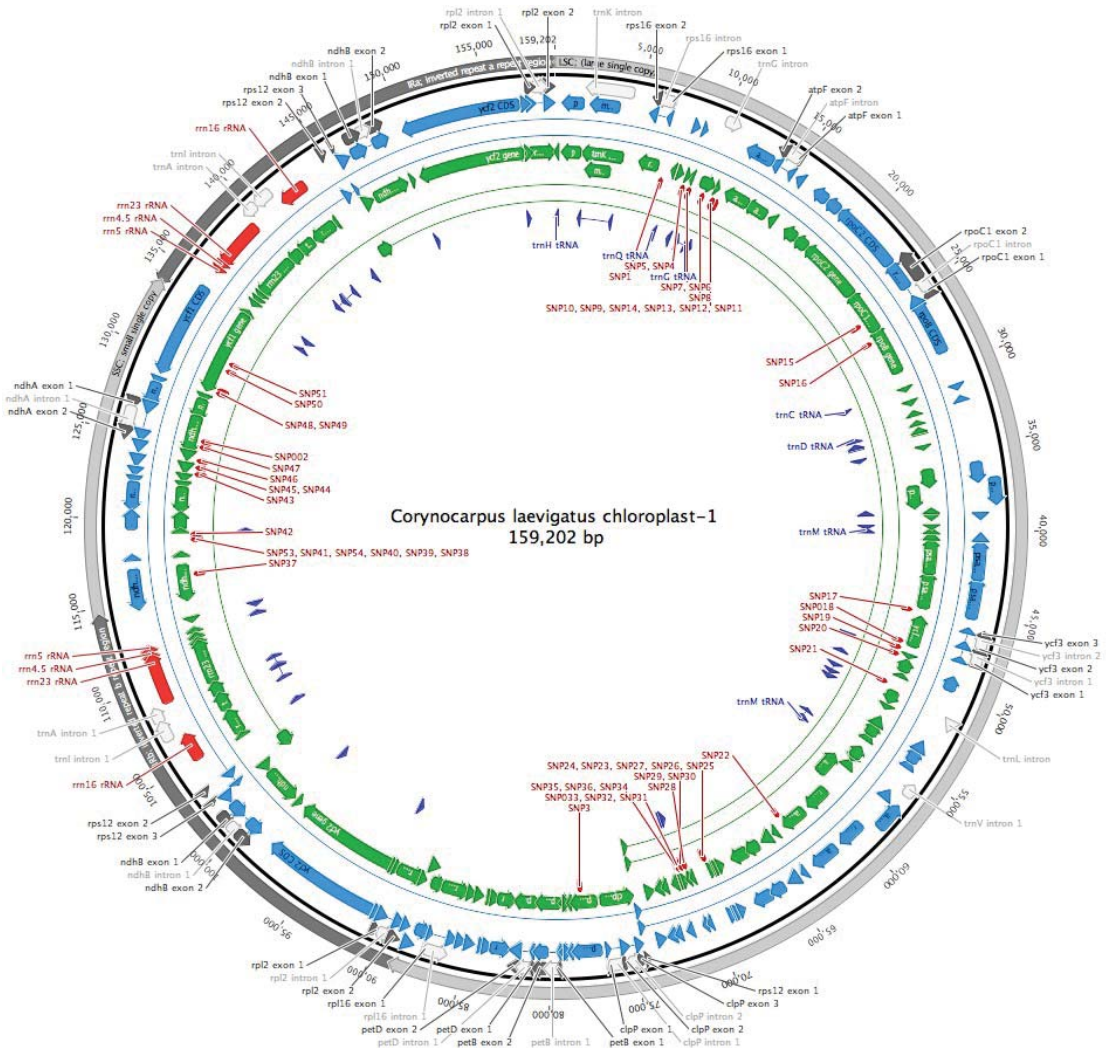

FIGURE 4.3: The chloroplast genome of karaka showing the position of 51 putative SNPs (red), genes (green), coding sequence (blue), trnA (dark blue), inverted repeat regions (sky blue). Figure produced using Geneious (Drummond *et al.*, 2009)**.**

4.7.4 HRM SCREENING OF THE KARAKA POPULATION

High resolution melting analysis of short PCR (<150 bp) products was used to genotype six SNPs in the karaka chloroplast genome. HRM was used effectively to identify single SNP differences in DNA sequences, which were assessed by viewing changes in the shape of their melting curve profiles. High-resolution melting variants were as follows (Figures 4.4a and 4.4b). SNP1 was a G→A transition and the G allele melted at 0.5˚C higher than the A allele. SNP3 was an A→C transversion, with a temperature difference of 0.5˚C. The spiking-in for this SNP was successful as the A allele, which was spiked with a C allele, displayed a typical heterozygous double melting peak. SNP8, was a G→A transition, with a temperature difference of 0.55˚C. The spiking-in for this SNP was successful as the G allele, which was spiked with an A allele was heterozygous, as above. SNP016 was a T→G transversion. HRM melting curves suggested four possible sequence variations. These were not confirmed by Sanger sequencing, which distinguished only two possible bases, T or G. Upon further investigation, those melting curves grouped as variant 1 and 3 were verified by Sanger sequencing to be a G and those grouped as variant 2 and 4 were a T. The spiking-in method for this SNP was successful as the T allele, which was spiked with a G allele, was heterozygous. SNP041, an A→C transversion, displayed two clear HRM profiles when viewed in the difference plot, however, profiles were difficult to distinguish using melting peaks data. Although temperature difference between the two melting types was very small (0.25˚C), the use of the difference plot was sufficient to manually bin SNP results. SNP049 displayed two clear profiles in both the difference plot and the melting peaks plot, with a temperature difference of 0.5˚C. The spiking-in for this SNP was successful, as the A allele, which was spiked with a T allele was heterozygous.

HRM data was unavailable for analysis for several accessions due to the failure of some of the PCR or ambiguities in the analysis of melting peaks. Of 2088 PCR reactions carried out for the HRM analysis, 199 (9.53%) failed to amplify and 186 (8.9%) gave dubious results for some loci, placing them in false haplotypes. Missing data and errors were resolved using Sanger sequencing. Table A5.1 in Appendix 5 compares HRM and Sanger sequencing results for each of 60 accessions using all 6 SNPs. Table 4.2 summarises these results.

4.7.5 HRM method compared with Sanger sequencing

HRM genotyping is a well-developed method for the analysis of nuclear sequence data but has had little application in chloroplast sequence analysis. Sanger sequencing is considered the 'gold standard' in molecular biology, generating accurate and reliable sequence data (0.001% errors per bp of sequence) and was therefore used as a standard against which the efficacy of HRM genotyping could be measured. Where HRM data was unavailable or suspected to be a mis-call, Sanger sequencing was used to determine the correct base-call. The extent of concordance between Sanger sequencing and HRM genotyping differed with each marker. The highest (98.08%) and lowest (75.44%) success rates were achieved with SNP16 and SNP3 respectively (Table 4.2). A full set of the data comparing HRM with Sanger sequencing for 60 accessions can be found in Table A5.1, Appendix 5.

Table 4.2: An evaluation of the concordance of HRM profiling with Sanger sequencing

SNP	No. of accessions	Matches	Mismatches	Concordance (%)
1	50	45	5	90
3	57	43	14	75.44
8	38	34	4	89.47
16	52	51	1	98.08
41	55	45	10	81.82
49	56	54	2	96.43

Concordance between HRM and Sanger was measured by dividing the matches by the number of accessions and multiplying by 100

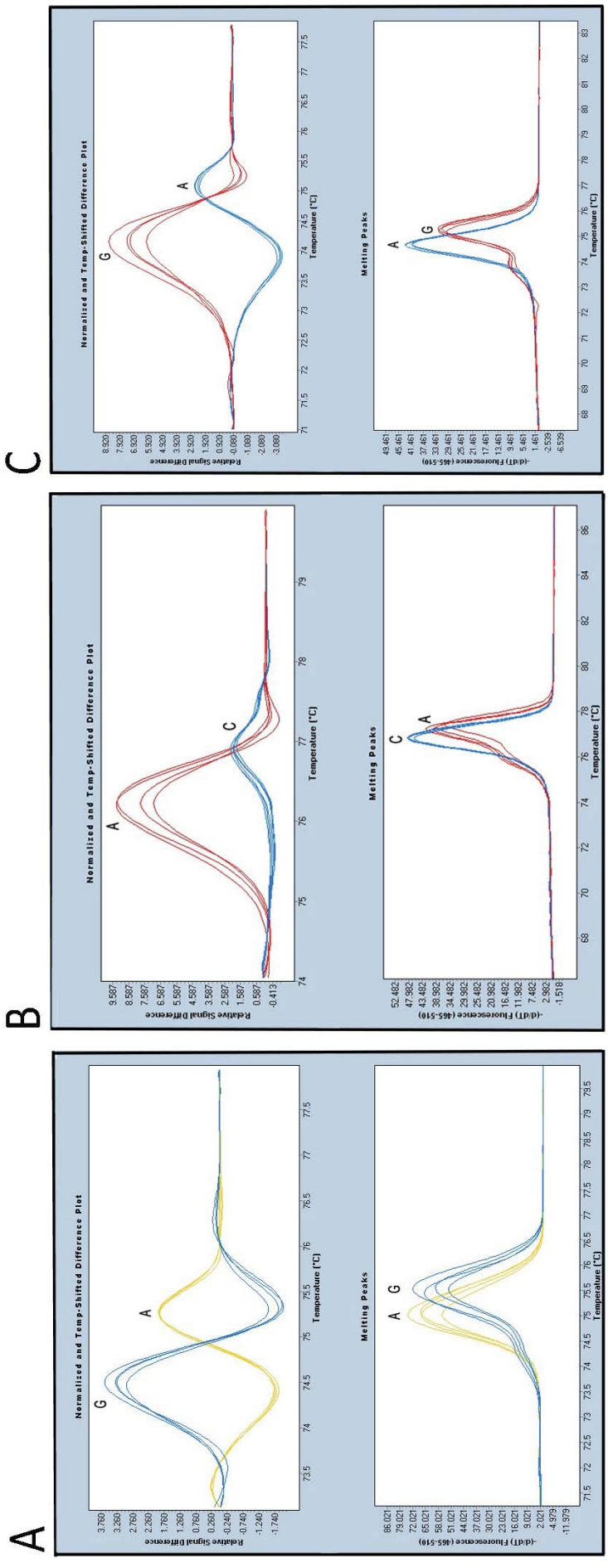

FIGURE 4.4a: High Resolution Melting analysis in the karaka chloroplast genome. (A) SNP1 (G→A); (B) SNP3 A→C; (C) SNP8 G→A. The normalised and temperature-shifted difference plot and the melting peaks were obtained using the LightCycler480software (Roche).

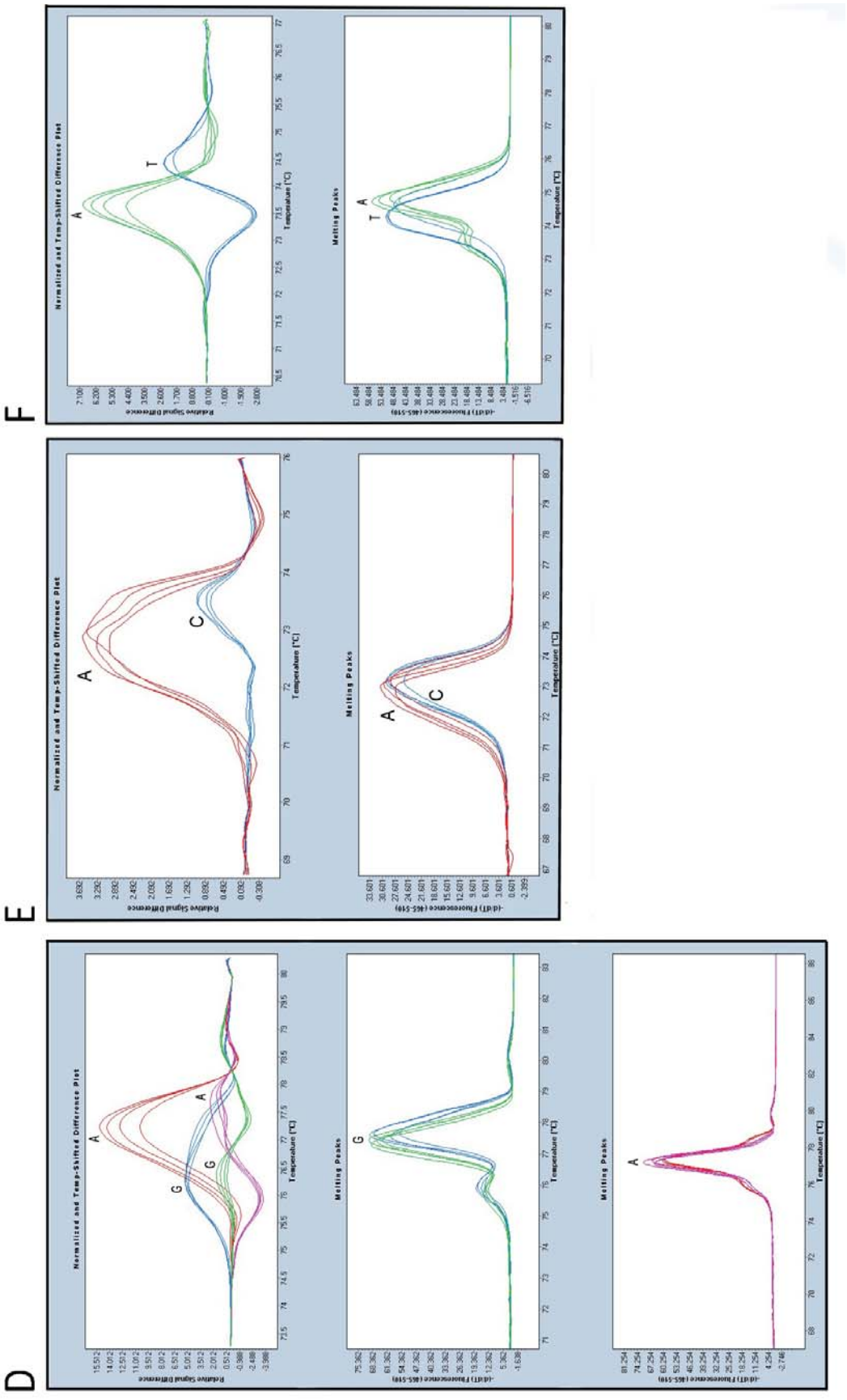

FIGURE 4.4b: High Resolution Melting analysis in the karaka chloroplast genome. (D) SNP016 (A→G); (E) SNP041 (A→C); (F) SNP049 (A→T). The normalised and temperature-shifted difference plot and the melting peaks were obtained using the LightCycler480software (Roche)

4.7.6 EXPLORING THE DISTRIBUTION OF KARAKA IN NEW ZEALAND

4.7.6.1 CHLOROTYPES AND THEIR RELATIONSHIPS

Table 4.3 shows a summary of polymorphisms for karaka at seven chloroplast loci. Table 6.1 in Appendix 6 contains the full data set. Figure 4.4 displays the relationship between accessions of karaka and the chlorotype found in other species of *Corynocarpus* studied by Wagstaff and Dawson (2000). *C. cribbianus*, *C. similis*, *C. dissimilis*, *C. rupestris* ssp. *rupestris* and *C. rupestris* ssp. *arborescens* were all identical at the six loci tested and together make up the species at the ancestral node in the network.

Karaka appears to have separated into two major groups: the first consisting of chlorotype 6 (yellow) and the second of chlorotype 1 (red), which in turn has given rise to chlorotype 2 (black), 3 (blue), 4 (purple) and 5 (orange).

Chlorotype	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
A	A	-	A	G	T	A	A
1	G	-	A	G	T	A	T
2	G	-	A	G	T	DEL	T
3	G	A	A	G	T	C	T
4	G	C	A	G	T	A	T
5	G	-	C	G	T	C	T
6	A	-	A	A	G	A	A

Table 4.3: Summary of chloroplast polymorphisms distinguishing chlorotypes. A = Ancestral type (*Corynocarpus cribbianus*, *C. dissimilis*, *C. similis*, *C. rupestris* ssp. *rupestris*, *C. rupestris* ssp. *arborescens*); DEL = deletion. Coloured squares correspond to points on the map on Fig. 4.5)

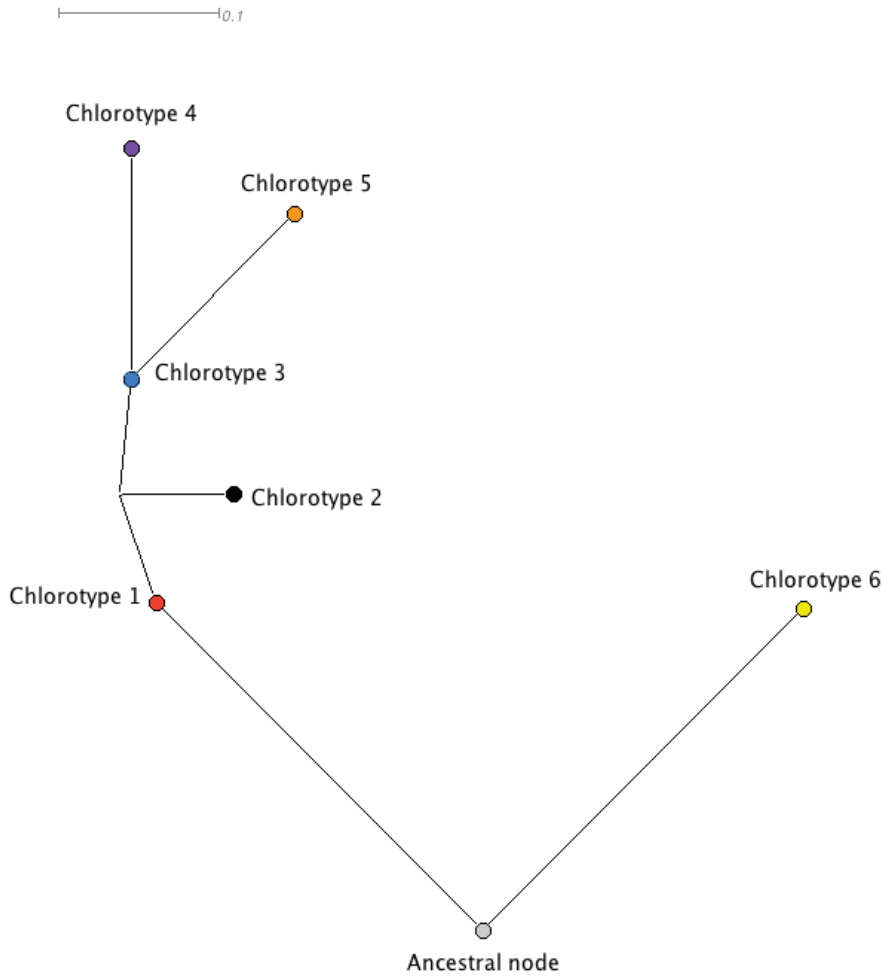

Figure 4.5: NEIGHBOURNET distance network, as implemented in the SplitsTree 4.0 package (Huson & Bryant, 2006), for the karaka chloroplast DNA dataset. Colours indicate chlorotypes whose distributions are shown in Figure 4.5

4.7.6.2 DISTRIBUTION OF HAPLOTYPES IN NZ

The six characterised SNPs and the additional SNP2, which was added in with a small number of accessions, were used for a phylogeographic study of karaka in New Zealand. The combination of these seven SNPs identified six chlorotypes across New Zealand (Figure 4.6). Chlorotype 1 (red) occurs on the Three Kings Islands, mid to northern Northland and one accession in Tauranga Harbour. Chlorotype 2 (black) is represented by a single accession on Waiheke Island. Chlorotype 3 (blue) appears to be restricted to northern Northland. Chlorotype 4 (purple) occurs from Kaitaia in Northland then south along the western side of the North Island to Taranaki. South of here, it grows along the west coast of the Wellington region. Other than these locations, it was sampled from

trees growing at Mount Maunganui in the Bay of Plenty, inland from Papatea Bay in northern Gisborne, and near Wairoa in Hawkes Bay. It is the only chlorotype represented on the Chatham Islands. Chlorotype 5 (orange) is found in Whangaruru and Matapouri (northeast of Whangarei), Kaipara, Tauranga Harbour and the northeastern Bay of Plenty. It is also the only chlorotype detected from the Kermadec Islands. Chlorotype 6 (yellow) has a very wide distribution across the North Island. It also occurs in several locations in the South Island (refer to Figure 1.9 in Chapter 1 for locations).

4.7.6.3 Comparison between chlorotype distribution and spatial, and climate data of karaka in New Zealand

The comparison between the karaka chlorotype distribution (Fig. 4.6) and the spatial and climatic work of Stowe (2003) (Figure 4.7) highlights that many of the trees sampled in this study occur in the vicinity of many of the putative cultural trees in Stowe's study. Figure 4.7 contains the plotted distribution of cultural and unknown trees. A comparison of Figures 4.6 and 4.7 suggests that there are specific haplotypes showing a geographic association with many putative cultural sites. These are chlorotype 1 (yellow) in the lower North Island and South Island; chlorotype 4 in the lower North Island, South Island and the Chatham Island, and to a lesser extent, chlorotype 1 (red) in Northland.

4.8 DISCUSSION

4.8.1 SNP discovery and verification

Reasons why putative SNPs were not confirmed following independent Sanger sequencing could be deemed probable Illumina sequencing errors or PCR artifacts. Other potential reasons include the SNP location, either they were located too close to an indel or occurring before long run of mononucleotides (SNPs 4, 5, 11-14). In several cases, the Illumina sequencing coverage was too low (<25 reads), this was the case for SNPs 018-022 and 028-036. SNPs 38-40 and SNP42 are situated in the *rpl32-trnL* intergenic region of the chloroplast genome, which is known to have high variability in other species (Shaw *et al.*, 2007). Although they were rare (MAF < 0.1) these SNPs were tested further.

4.8.2 EFFECTIVENESS OF HRM PROFILING

This study examined the reliability of HRM as a method for screening SNP variants in the chloroplast genome of karaka. In order to screen a series of SNPs in a large number of accessions, a cost-effective yet high-throughput method of genotyping was required. The HRM platform proved to be an efficient and reasonably accurate method for genotyping SNPs from chloroplast regions in a large number of karaka accessions. The utilisation of HRM as a screening technique for chloroplast SNPs has been shown to be successful for the majority of the markers characterised for this species. However, it is not without its limitations.

A disadvantage of HRM technique is that interpretation of melt curves can potentially be difficult where chloroplast mutational dynamics are complex. This has been suggested to be the case for some fast-evolving chloroplast genome regions. Ahmed *et al.* (2012) have suggested that there is a genome-wide association between repeats, indels and substitutions, and for some of the fastest evolving regions, characterisation for PCR and even sequencing can be problematic (Ahmed *et al.*, 2013). For such cases, HRM profiles are also expected to be complex and more difficult to interpret. This was apparent for SNP8, for which there were only data for 56% of accessions and for SNP3, for which there was only 75% concordance between HRM and Sanger sequencing.

Our initial investigations using HRM to screen chloroplast SNPs revealed that the melt curves of some amplicons were difficult to distinguish from one another due to the small differences in melting temperature (T*m*˚). Single base changes can be difficult to distinguish, the largest temperature change being from a G⇔C and the smallest from A⇔T. This is because G-C base pairings have three hydrogen bonds between them while A-T base pairs have only two.

Chloroplast sequence appears as a homoduplex when using HRM and can differ only in the temperature shift, not the shape of the curve. To remedy this, total DNA of all 60 samples were 'spiked-in' 50:50 with one accession (RA123) to allow the formation of heteroduplexes, which usually exhibit a very distinctive shape and would therefore be easier to distinguish between. This was the method used for SNPs 3, 8, 16 and 49.

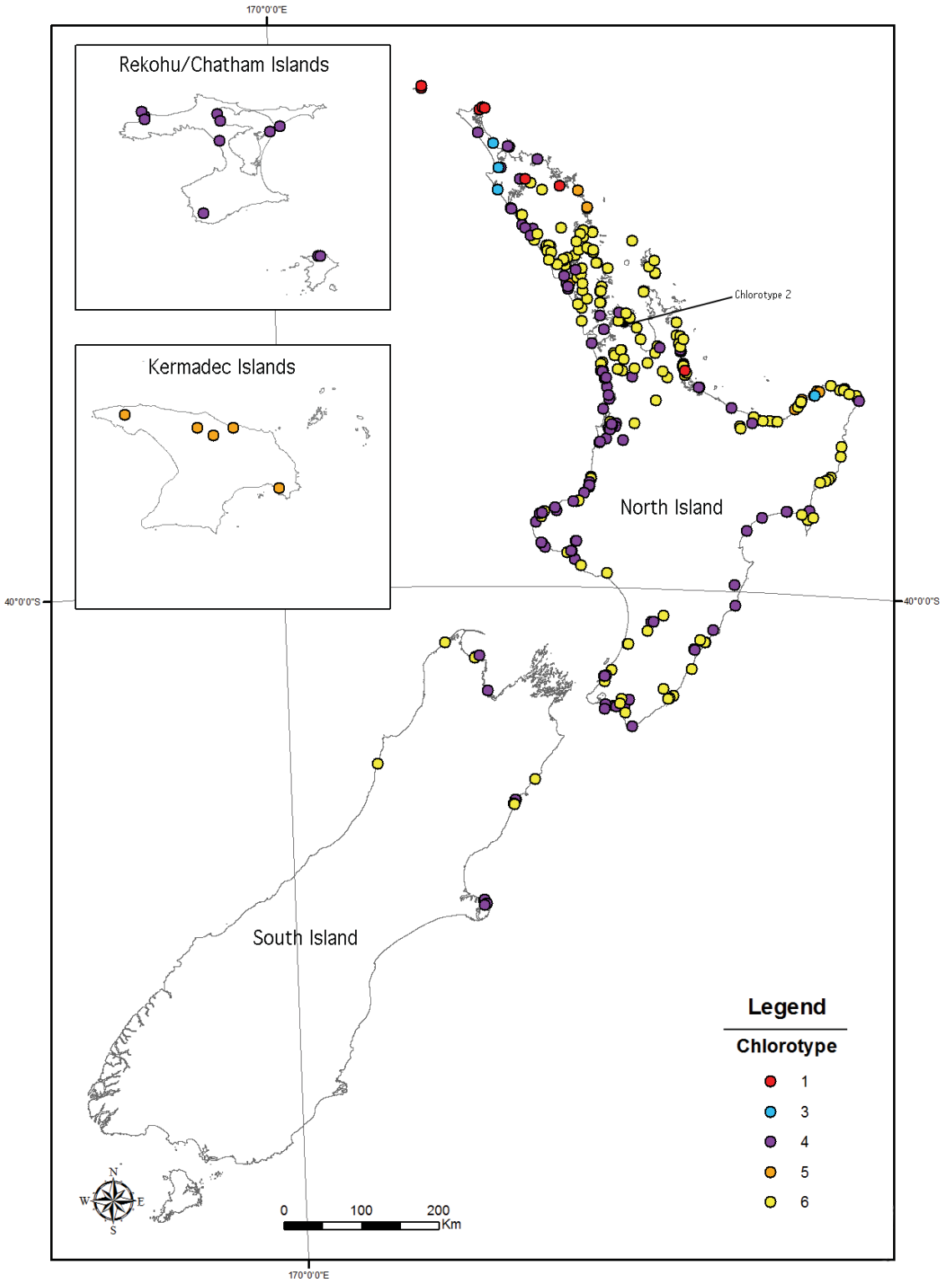

FIGURE 4.6: Distribution and genetic variation in karaka (*Corynocarpus laevigatus*). The circles indicate sampled individuals and the colours represent one of 6 chlorotypes (red: chlorotype 1; black: chlorotype 2; blue: chlorotype 3; purple: chlorotype 4; orange: chlorotype 5; and yellow: chlorotype 6, as described in the text. Kermadec Islands to the north-east contain just one chlorotype (5, orange) and Chatham Islands to the east of the South Island also contain just one chlorotype (4, purple). Figure courtesy of Matt Irwin, Massey University.

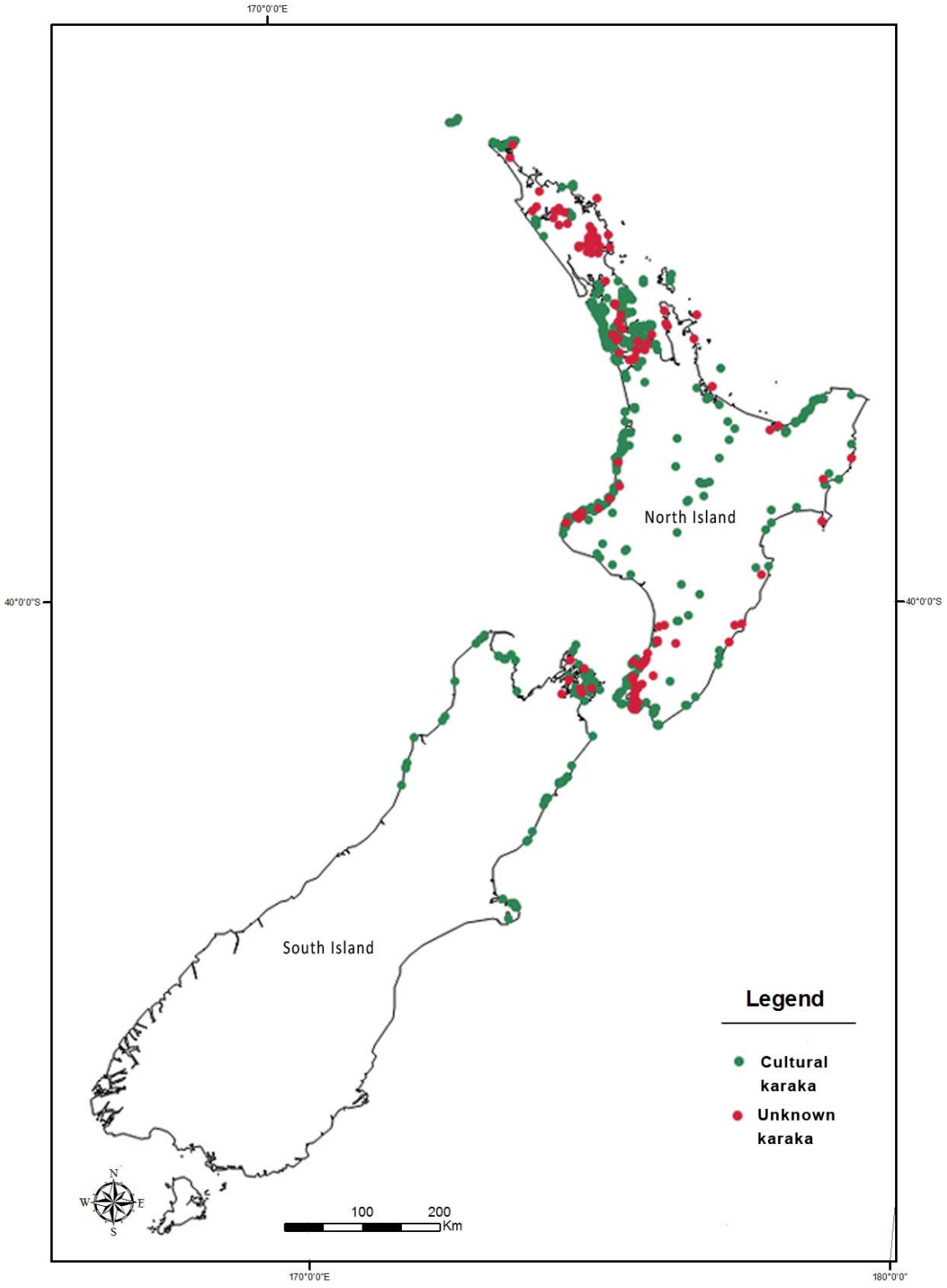

FIGURE 4.7: Distribution of cultural and non-cultural karaka (*Corynocarpus laevigatus*) as per Stowe (2003). Figure courtesy of Rachael Ouwejan.

AT-rich sequences in both coding and non-coding regions are a feature of plastid genomes (Howe *et al.*, 2003). The AT-rich sequence poses significant challenges to finding suitable priming sites compatible with HRM, limiting this as a method of screening for SNPs in the chloroplast genome, particularly in low diversity species with few SNPs available. For HRM analysis, it is typically preferable to design primers to amplify regions <150 bp. Wittwer (2003) found HRM analysis could distinguish melting curves up to 304bp, however, this ability decreased as amplicon size increased as HRM analysis is more sensitive when there is less flanking DNA (Liew *et al.*, 2004; Reed & Wittwer, 2004). Primers were designed to amplify sequences of length 75-150 bp. This limited the search for suitable primers within an already GC poor sequence and several of the designed primers had high primer-dimer scores. Of nineteen primer pairs, thirteen were discarded after the initial testing stage.

Success of our HRM reactions relied upon the quality of the extracted genomic DNA. Dang *et al* (2012) found the sensitivity of their analysis may have been affected by the DNA template quality which had also been extracted using a modified CTAB protocol. Whilst CTAB is a very simple and effective method of DNA extraction, the resulting DNA can be of varying quality between accessions, and with plants such as karaka, the presence of secondary compounds can decrease the quality of DNA. The potentially uneven and low quality of our DNA template may thus have had an impact on subsequent melting analysis, generating system errors between melting temperature (Tm) readings of HRM assays. A more robust DNA extraction protocol could have been developed, however, at this stage of the project, it was more financially viable to correct HRM errors using Sanger sequencing, rather than re-extracting sample DNA. However, our results show that HRM, followed by Sanger sequencing, can be an effective two-step strategy for the detection of SNP mutations in the chloroplast genome of karaka.

4.8.3 Evolution and distribution of karaka chlorotypes in New Zealand

Karaka appears to exhibit very little chloroplast variation across its distributional range. It is unlikely that universal chloroplast markers alone would have provided the level of variation detected using high-throughput sequencing. However, genetic variation is sufficient to draw some preliminary conclusions. The extant distribution of karaka comprises at least six chlorotypes, five of which are closely related. The ancestral

haplotype, inferred as such because of its presence in out-groups, has not been found in New Zealand. These observations raise a number of interesting questions, including whether the ancestral type has gone extinct in New Zealand, or rather it was confined to a more northerly landmass, or whether there have been multiple dispersal events into New Zealand, one lineage giving rise to five chlorotypes, and then a second founding event giving rise to another chlorotype, which is now widely-distributed in New Zealand.

This study suggests the biogeographical history of karaka is complex and is consistent with human-assisted dispersal. However, the extent to which this has happened was not resolved with the current level of resolution in our genetic data. The data presented in this thesis alone are insufficient to distinguish between recent natural dispersal of karaka and translocations from northern refugia. These results point to the possibility that karaka was restricted to Northland, which served as a refugium during the Last Glacial Maximum (LGM) in New Zealand, from ca. 29 ka[9] to ca. 19 ka. (Newnham *et al.*, 2007). The presence of chlorotype 5 (orange) in the Bay of Plenty may also suggest karaka could have been restricted to this region too during the LGM. These results are in concordance with Garnier (1958), who suggested the southern limit of many of the species restricted to the northern North Island is approximately 38°S; this boundary is where the warmer climate of the northern region meets the cooler climate of the southern region.

The geographic distribution of the haplotypes suggested that the dispersal of haplotypes is very restricted - consistent with restricted seed dispersal (natural and or human mediated). The chlorotype distribution pattern may also occur if cpDNA was transmitted through karaka pollen and it was poorly dispersed. There is not enough genetic resolution to distinguish between these possibilities.

All six chlorotypes occur in karaka populations from the region north of this boundary, whereas in the southern North Island and northern South Island, only two of the chlorotypes (4 and 6) are represented. However the distribution and observed levels of chlorotype diversity suggest directions for future analyses. Chlorotype 4 and chlorotype 6 in particular appear to be candidates for testing hypotheses on translocations further due to their association with putative cultural trees. However, higher resolution markers

[9] ka - thousand years ago

are needed to test between hypotheses of natural and human assisted dispersal. At the moment it is not possible to distinguish between natural and human dispersal, but candidate chlorotypes for human dispersal are suggested because of their association with many of the cultural trees described in Stowe (2003). Analyses with higher resolution markers, in particular for chlorotypes 4 and 6, are needed to test for the occurrence of more significant non-random patterns that might indicate translocation.

4.9 ALTERNATIVE APPROACHES

Given that translocation is likely to have occurred within the last 1000 years, and given observed levels of chlorotype diversity, translocated karaka is likely to be identical between points of translocation for chloroplast markers; e.g. the Chatham Island accessions exhibit the purple chlorotype associated with several locations across New Zealand. Further analysis of uncharacterized chloroplast regions is required. Currently, 64.73% (69817 kb/107854 kb) of the karaka chloroplast genome (not including inverted repeats) has been studied; of this, 16500 bp are predicted hotspot regions (Figure 4.6) and 14500 bp predicted regions remain unstudied. Nuclear microsatellites could be utilised to match Chatham Island accessions to a specific mainland locality or source of origin. Microsatellites have had a long history of use in ecology and evolutionary biology, but require time and money to develop for each species. In addition, the locus number can be limiting, and levels of polymorphism must remain below the threshold at which problems arise due to homoplasy (Grover *et al.*, 2012).

Additional sequence data, obtained by increasing the number and type of molecular marker, combined with a geological or molecular calibration method will be useful for illuminating not only the timing of the movement of accessions, but also the direction. Did karaka enter New Zealand through the Three Kings Islands as ITS data suggests? Was karaka already growing on the Chatham Islands before human occupation?

We have interesting candidate plants relevant to translocation history; however, further genetic characterisation is required to develop a clearer picture of the dispersal and translocation history of karaka in New Zealand. Our method suggests while HRM analysis of short amplicons containing SNPs was not without its limitations, it enabled a

large number of accessions to be broadly partitioned into chlorotypes that could be verified by Sanger sequencing (Table A5.1, Appendix 5).

An alternative approach could have made use of sequencing tags added to the PCR product before Illumina sequencing. This would have been preferable to Sanger sequencing all individuals to determine which accessions have a given mutation. This approach may have made the detection of rarer SNPs easier too. Despite long-range PCR products being pooled in equimolar amounts, one individual may have been represented in greater quantity than others, making it appear as though the SNP is rare, when in fact the individual(s) carrying the mutation are simply under-represented in the reads. In future work, with the recent developments in indexing, more use could be made of these for SNP validation.

Preliminary results using the Repliphi™ PHI29 DNA polymerase reagent kit (Epicentre Technologies Corp., Chicago), which utilises AT-rich hexamers to preferentially amplify organellar DNA (Howe *et al.*, 2003) suggests this is a promising direction for future work. Repliphi™ can be used to amplify single copy regions of the chloroplast genome of other accessions in chlorotype 4 (purple) followed by Illumina sequencing. Twelve SNPs could not be amplified using HRM (Table A4.1, Appendix 4) for several reasons, including PCR failure and the formation of primer dimers. These SNPs have the potential to be informative and increase resolution within chlorotype 4. This could help determine the source location of trees growing on the Chatham Islands.

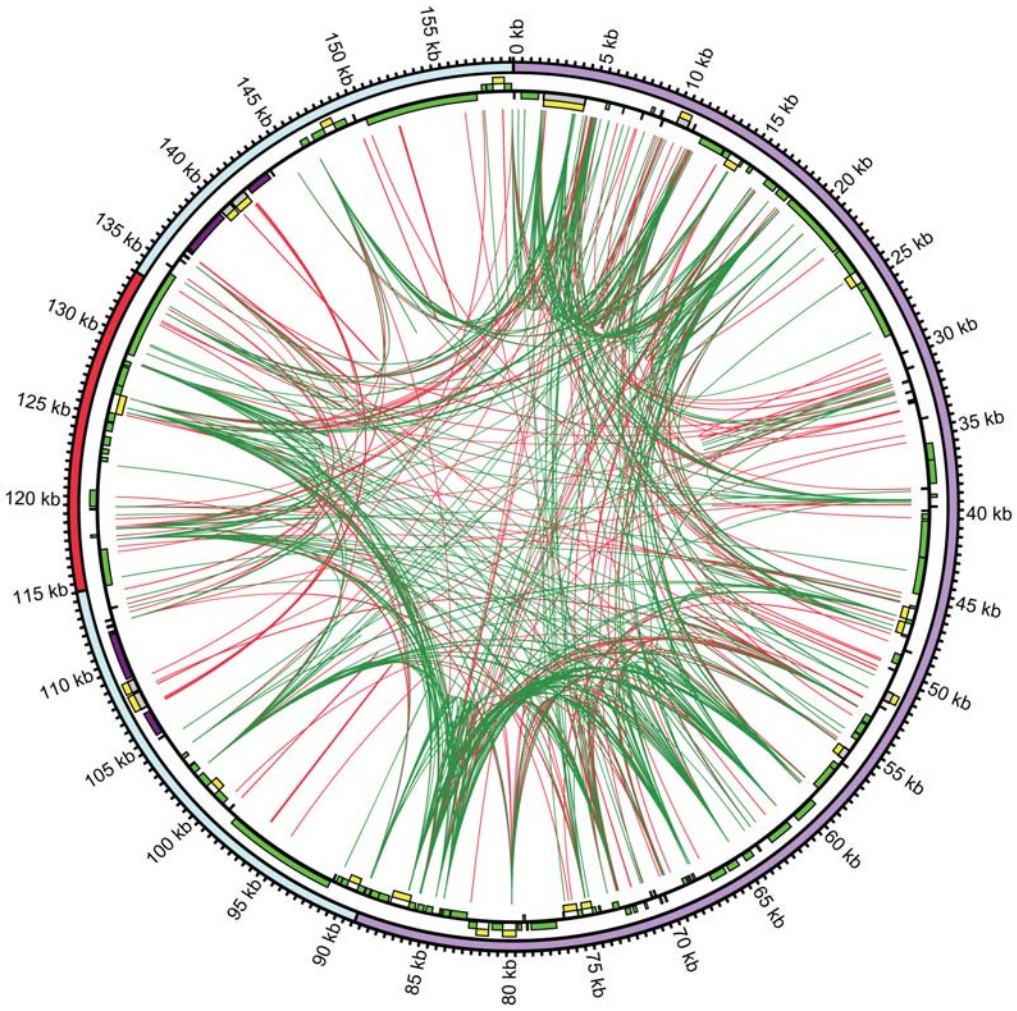

FIGURE 4.8: Circos plot of karaka (*Corynocarpus laevigatus*) showing the relationship between short repeats within the chloroplast genome and distribution of indels. Large single copy is shown in purple, small single copy in red and inverted repeats in pale blue; genome annotation on the positive and negative strand (genes in green; tRNAs in yellow and rRNAs in purple). The figure centre shows the results of Reputer mapping using the karaka chloroplast genome. Two ends of a red line mark the two locations of the forward (direct) repeats, while those of a green line mark the two locations of the reverse (inverted) repeats on the genome. In this part of the figure, the large inverted repeats are not plotted, as they would obscure a large part of the figure. Number of repeats shown in the diagram is 448 forward and 481 reverse, with a size range from 15 bp to 39 bp (average repeat size: 16.8 bp). Reputer plots calculated with repeats >15 and Hamming distance of 0. Image courtesy of Patrick Biggs, mEpiLab, Massey University, Palmerston North.

4.10 CONCLUSION

This genetic study of karaka, coupled with the spatial and climatic distribution data of Stowe (2003), sheds yet more light on the complex history of the karaka tree in New Zealand. It is not possible with the present genetic data of this resolution to precisely illuminate individual translocation events.

The results presented in this chapter show the potential of the chloroplast genome to study recent events in plant history, and the use of HRM to assay several hundred accessions for a suite of chloroplast SNPs. They show an interesting relationship between Kermadec Island and mainland karaka, and between Chatham Islands and mainland karaka. Kermadec Island karaka are represented by chlorotype 5, which is found in a few locations on the mainland. Chatham Islands karaka all demonstrate chlorotype 4, a chlorotype found in accessions growing throughout the mainland from Northland, through the Bay of Plenty, the East Cape (northwestern coast of the Gisborne region), to Taranaki and the Kapiti Coast (east coast of the Wellington region). These results suggest the karaka on the Chatham Islands (or their ancestors) were translocated from the mainland, though when, how and by whom is not determinable from our data. Karaka haplotypes on the Kermadecs matched to a handful of accessions on the Northland coast between Whangaruru North Head and Matapouri. This could be explained by a translocation *to* New Zealand, consistent with oral histories (Smith, 1891). However, a translocation from New Zealand to the Kermadecs and natural dispersal cannot be ruled out with the chloroplast data. To be able to pinpoint the location of the source for Kermadec and Chatham Islands karaka, more genetic work is required.

However, these results are promising in their ability to trace the translocation of one of New Zealand's most important ethnobotanical species. By developing a more detailed picture of the genetic variation of karaka, this work has the potential to be the foundation for a deeper study into the translocation of the species. This has implications for further understanding the level of domestication in karaka, which at present cannot be ascertained.

4.11 REFERENCES

Ahmed, I., Biggs, P., Matthews, P., Collins, L., Hendy, M., and Lockhart, P. 2012. Mutational dynamics of aroid chloroplast genomes. *Genome Biology and Evolution* **4:** 1316-1323.

Ahmed, I., Matthews, P. J., Biggs, P. J., Naeem, M., McLenachan, P. A., and Lockhart, P. J. 2013. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of Colocasia esculenta (L.) Schott (Araceae) and closely related taxa. *Molecular Ecology Resources***:** n/a-n/a.

Allaby, R. G., and Brown, T. A. 2003. AFLP data and the origins of domesticated crops. *Genome* **46:** 448-453.

Anderson, A., and McFadgen, B. 1990. Prehistoric two-way voyaging between New Zealand and East Polynesia: Mayor Island obsidian on Raoul Island and possible Raoul Island obsidian in New Zealand. *Archaeology in Oceania* **25:** 37-42.

Anderson, A. B. 1991. The chronology of colonization in New Zealand. *Antiquity* **65:** 767–795

Atherton, R. A., McComish, B. J., Shepherd, L. D., Berry, L. A., Albert, N. W., and Lockhart, P. J. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods* **6**.

Belich, J. 1996. *Making Peoples*. Allen Lane/Penguin: Auckland.

Best, E. 1902. Art. V.—Food Products of Tuhoeland: being notes on the food-supplies of a non-agricultural tribe of the natives of New Zealand; together with some account of various customs, superstitions, &c., pertaining to food. *Transactions and Proceedings of the Royal Society of New Zealand* **35:** 54.

———. 1972. *Tuhoe: the children of the mist*. 2nd edition. Memoirs of the Polynesian Society. A.H. and A.W.Reed for The Polynesian Society: New Plymouth.

Best, E. 1976. *Maori agriculture: the cultivated food plants of the natives of New Zealand with some account of native methods of agriculture its ritual and origin myths*. Government Printer: Wellington.

———. 1977. *Forest lore of the Māori*. E.C. Keating: Wellington.

Birky, C. 1995. Uniparental inheritance of mitochondrial and chloroplast genes — mechanisms and evolution. *Proceedings of the National Academy of Sciences of the USA* **92:** 11331–11338.

Buick, T. 1903. *Old Manawatu, or The Wild Days of the West*. Buick & Young: Palmerston North.

Chagné, D., Gasic, K., Crowhurst, R. N., Han, Y., Bassett, H. C., Bowatte, D. R., Lawrence, T. J., Rikkerink, E. H. A., Gardiner, S. E., and Korban, S. S. 2008. Development of a set of SNP markers present in expressed genes of the apple. *Genomics* **92:** 353-358.

Clarke, A. 2007. *The great sacred forest of Tane: Te wao tapu nui a Tane - a natural pre-history of Aotearoa New Zealand*. 1st edition. Reed Publishing: Auckland.

Coart, E., van Glabeke, S., De Loose, M., Larsen, A. S., and Roldan-Ruiz, I. 2006. Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Molecular Ecology* **15:** 2171-2182.

Cox, M. P., Peterson, D. A., and Biggs, P. J. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *Bmc Bioinformatics* **11:** -.

Cronn, R., Liston, A., Parks, M., Gernandt, D., Shen, R., and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* **36:** e122-e122.

Demesure, B., Sodzi, N., and Petit, R. J. 1995. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology* **4:** 129-134.

Diekmann, K., Hodkinson, T., Fricke, E., Barth, S., and Baxter, I. 2008. An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. *PLoS One* **3**.

Dodson, J. R. 1976. Modern pollen spectra from Chatham Island, New Zealand. *New Zealand Journal of Botany* **14:** 341-347.

Doebley, J. 1992. Molecular systematics and crop evolution. *In* Soltis, P. S., Soltis, D., Doyle, J.J. (Ed.), *Molecular systematics of plants*, pp. 202–222. Chapman & Hall: London.

Doyle, J. J., and Doyle, J. L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue *Phytochemistry Bulletin* **19:** 11–15.

Drummond, A., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T., and Wilson, A. 2009. Geneious v4.7**:** Available from http://www.geneious.com.

Feleke, Y., Pasquet, R., and Gepts, P. 2006. Development of PCR-based chloroplast DNA markers that characterize domesticated cowpea (*Vigna unguiculata* ssp. *unguiculata* var. *unguiculata*) and highlight its crop-weed complex. *Plant Systematics and Evolution* **262:** 75-87.

Fu, Y.-B., and Allaby, R. 2010. Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. *Genetic Resources and Crop Evolution* **57:** 667-677.

Garnier, B. J. 1958. *The climate of New Zealand*. Edward Arnold: London.

Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S., and Hellwig, F. H. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution* **20:** 1499-1505.

Grover, C. E., Salmon, A., and Wendel, J. F. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis1. *American Journal of Botany* **99:** 312-319.

Hasebe, M., Omori, T., Nakazawa, M., Sano, T., Kato, M., and Iwatsuki, K. 1994. rbcL gene-sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 5730-5734.

Holmgren, P., Holmgren, N., and Barnett, L. 1990. Index herbariorum. Part 1, The herbaria of the world. *Regnum Vegetabile* **120:** 1-693.

Holt, K. 2009. *The Quarternary History of Chatham Island, New Zealand. Ph.D. thesis*. Massey University, Palmerston North.

Houston, J. 1965. *Maori life in old Taranaki*. AH & AW Reed: Wellington.

Howe, C. J., Barbrook, A. C., Koumandou, V. L., Nisbet, R. E. R., Symington, H. A., and Wightman, T. F. 2003. Evolution of the chloroplast genome. *Journal Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **358:** 9.

Huson, D. H., and Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23:** 254-267.

Leach, H. M., and Stowe, C. J. 2005. Oceanic arboriculture at the margins: the case of the karaka (*Corynocarpus laevigatus*) in Aotearoa. *Journal of the Polynesian Society* **114:** 7-27.

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

Liew, M., Pryor, R., Palais, R., Meadows, C., Erali, M., Lyon, E., and Wittwer, C. 2004. Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clin Chem* **50:** 1156-1164.

Matisoo-Smith, E., and Robins, J. H. 2004. Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 9167-9172.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez, G. J., Buckler, E., and Doebley, J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 6080-6084.

McCauley, D. E. 1995. The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology & Evolution* **10:** 198-202.

Mildenhall, D. C. 1994. *Palynological reconnaissance of Early Cretaceous to Holocene sediments, Chatham Islands, New Zealand*. Institute of Geological & Nuclear Sciences Ltd: Lower Hutt, New Zealand.

Miller, A., and Gross, B. 2011. From forest to field: Perennial fruit crop domestication. *American Journal of Botany* **98** 1389-1414.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. 2010. Tablet-next generation sequence assembly visualization. *Bioinformatics* **26:** 401-402.

Mitira, T. J. H. M. 1972. *Takitimu*. A. H. & A. W. Reed: Wellington.

Molloy, B. P. J. 1990. The origin, relationships, and use of karaka or kopi (*Corynocarpus laevigatus*). *In* Kapoor, W. H. a. P. (Ed.), *Nga Mahi Maori o te Wao Nui a Tane: contributions to an International Workshop on Ethnobotany, Te Rehua Marae*, pp. 48–53. Botany Division, Department of Scientific and Industrial Research: Christchurch, New Zealand.

Newcomb, R. D., Crowhurst, R. N., Gleave, A. P., Rikkerink, E. H., Allan, A. C., Beuning, L. L., Bowen, J. H., Gera, E., Jamieson, K. R., Janssen, B. J., Laing, W. A., McArtney, S., Nain, B., Ross, G. S., Snowden, K. C., Souleyre, E. J., Walton, E. F., and Yauk, Y. K. 2006. Analyses of expressed sequence tags from apple. *Plant Physiology* **141:** 147-166.

Newnham, R. M., Lowe, D. J., Giles, T., and Alloway, B. V. 2007. Vegetation and climate of Auckland, New Zealand, since ca. 32 000 cal. yr ago: support for an extended LGM. *Journal of Quaternary Science* **22:** 517-534.

Petersen, J., Parker, I., and Potter, D. 2012. Origins and close relatives of a semi-domesticated neotropical fruit tree: *Chrysophyllum cainito* (Sapotaceae). *American Journal of Botany* **99:** 585-604

Reed, G. H., and Wittwer, C. T. 2004. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clinical Chemistry* **50:** 1748-1754.

Rozen, S., and Skaletsky, H. J. 2000. Primer3 on the WWW for general users and for biologist programmers. *In* Krawetz, S. and Misener, S. (Eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pp. pp 365-386. Humana Press: Totowa, NJ.

Sanger, F., Nicklen, S., and Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74:** 5463-5467.

Shand, A. 1896. The Moriori peoples of the Chatham Islands, their traditions and their history. *Journal of the Polynesian Society* **5:** 13–32.

Shaw, J., Lickey, E., Schilling, E., and Small, R. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94:** 275.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., and Liu, W. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92:** 142–166.

Smith, S. P. 1891. Notes on the geographical knowledge of the Māoris. *In* Hector, J. (Ed.), *Australasian Association for the Advancement of Science Conference*, Christchurch, New Zealand, Vol. 3, pp. 280-310. Wellington: George Didsbury.

———. 1893. 28 - Notes and queries. *Journal of the Polynesian Society* **2:** 125-127

———. 1900. The Aotea Canoe - The migration of Turi to Aotearoa (New Zealand). *Journal of the Polynesian Society* **9:** 211-233.

Stowe, C. J. 2003. *The ecology and ethnobotany of karaka (Corynocarpus laevigatus) [MSc. thesis]*. University of Otago, Dunedin.

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17:** 1105-1109.

Wagstaff, S., and Dawson, M. 2000. Classification, origin, and patterns of diversification of *Corynocarpus* (Corynocarpaceae) inferred from DNA sequences. *Systematic Botany***:** 134-149.

Wardle, P. 1963. Evolution and distribution of the New Zealand flora, as affected by Quaternary climates. *Evolution*.

Weising, K., Nybom, H., Wolff, K., and Kahl, G. 2005. Applications of DNA fingerprinting in plant science. *In DNA fingerprinting in plants: principles, methods, and applications*, p. 472. CRC Press: Boca Raton.

Whatahoro, H. 1915 Lore of the whare wānanga. Volume II. Te Kauwae-Raro, or 'Things Terrestrial' (translated by Smith, S.P.). *Journal of the Polynesian Society* **24:** 29-56.

Whistler, W. 1991. Polynesian plant introductions. *In* Cox, P. A. and Banack, S. A. (Eds), *Islands, plants, and Polynesians: an introduction to Polynesian ethnobotany*. Dioscorides Press: Portland, Oregon.

Whittall, J. B., Syring, J., Parks, M., Buenrostro, J., Dick, C., Liston, A., and Cronn, R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* **19:** 100-114.

Wills, D. 2006. Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.). *Journal of Heredity* **97:** 6.

Wilmshurst, J. M., Higham, T. F. G., Allen, H., and Johns, D. 2004. Early Maori settlement impacts in northern coastal Taranaki, New Zealand. *New Zealand Journal of Ecology* **28:** 167-179.

Wilmshurst, J. M., Anderson, A. B., Higham, T. F. G., and Worthy, T. H. 2008. Dating the late prehistoric dispersal of Polynesians to New Zealand using the commensal Pacific rat. *Proceedings of the National Academy of Sciences*.

Wilmshurst, J. M., Hunt, T. L., Lipo, C. P., and Anderson, A. J. 2011. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proceedings of the National Academy of Sciences* **108:** 1815-1820.

Wittwer, C. T., Reed, G. H., Gundry, C. N., Vandersteen, J. G., and Pryor, R. J. 2003. High-resolution genotyping by amplicon melting analysis using LCGreen. *Clinical Chemistry* **49:** 853-860.

Wu, S. B., Wirthensohn, M., Hunt, P., Gibson, J., and Sedgley, M. 2008. High resolution melting analysis of almond SNPs derived from ESTs. *Theoretical and Applied Genetics* **118:** 1-14.

Yan, G., Lv, X., Lv, P., Xu, K., Gao, G., Chen, B., and Wu, X. 2012. Application of high-resolution melting for variant scanning in chloroplast gene *atpB* and *atpB-rbcL* intergenic spacer region of Crucifer species. *African Journal of Biotechnology* **11:** 7016-7027.

Zhang, Y., Fang, Z., Wang, Q., Yang, L., Zhuang, M., and Sun, P. 2012. Chloroplast subspecies-specific SNP detection and its maternal inheritance in *Brassica oleracea* L. by using a dCAPS marker *Journal of Heredity* **103:** 6.

Zheng, Y.-H., Alverson, A., Wang, Q.-F., and Palmer, J. 2013. Chloroplast phylogeny of *Cucurbita*: Evolution of the domesticated and wild species. *Journal of Systematics and Evolution*.

5

THESIS SUMMATION AND FUTURE DIRECTIONS

The objectives of this study were fivefold:

1. What are the evolutionary relationships of the species within the genus *Corynocarpus*?
2. Can whole genome sequencing of chloroplast genomes of a non-model species provide a sufficient number of molecular markers for a phylogeographic study?
3. Is there a cost-effective method of genotyping multiple markers in a large number of accessions?
4. Does the karaka chloroplast genome variation provide sufficient phylogenetic resolution to elucidate the history of its translocation and therefore Maori settlement in New Zealand/Aotearoa?
5. Is there enough resolution in the data to determine whether karaka was brought into cultivation once or multiple independent times?

5.1 THESIS SUMMARY

5.1.1 GENERAL SUMMARY

The approach taken for this work has been to examine single nucleotide polymorphisms (SNP) variation in the chloroplast genome of karaka to elucidate the genetic relationship

between trees growing in the natural range of karaka and putative translocated trees and the extent to which karaka was domesticated, if at all.

Assessing genetic diversity in plants has become more sophisticated with the advent of high-throughput sequencing techniques. Although microsatellites are often the favoured option for these studies due to their multi-allelic states, development and genotyping of large numbers of accessions can be expensive. Chloroplast DNA sequence variation is a major source of data for inferring plant phylogenies (Shaw *et al.*, 2005). The chloroplast genome has been used to search for markers for the study of domestication in apple, (Coart *et al.*, 2006), cowpea (Feleke *et al.*, 2006), *Brassica oleracea* (Wills, 2006), sunflower (Zhang *et al.*, 2012), *Cucurbita* (Zheng *et al.*, 2013) and *Linum* (Fu & Allaby, 2010) amongst many others. However, evolution in the chloroplast genome can be so slow, which leads to very little per-nucleotide variation (Zurawski & Clegg, 1987), (Palmer, 1985) and thus the variation may be too low to investigate intraspecific variation.

Our initial investigations of chloroplast sequence diversity, using six loci amplified with universal primers, suggested diversity across the entire chloroplast would be high enough to develop a suite of chloroplast SNP markers. However, our genome-wide assessment of SNP variation in the karaka chloroplast revealed very low levels of genetic variation and structure.

5.1.2 CHLOROPLAST ISOLATION

To develop chloroplast SNP markers for this study, a number of protocols to isolate chloroplast DNA from nuclear DNA were investigated. One of the primary issues was obtaining sufficient chloroplasts in the isolate, the second was carry through of nuclear DNA during the chloroplast DNA isolation process. Whilst it is possible to use a DNAse enzyme to remove nuclear DNA, our chloroplast yield was too small to subject the sample to further possible destruction through enzymatic activity on the chloroplast molecules. Because the chloroplast yield was not sufficient to use with the Illumina GAII instrument, i.e., less than 5ug, samples were subjected to rolling circle amplification (RCA) to increase the amount of DNA. RCA using the RepliG kit (Qiagen) appears to

preferentially amplify nuclear DNA, which is due to the sequence composition of the primers in the kit reagents. However, RepliG amplified DNA contained fewer contaminants than non amplified DNA, making it more suitable for Illumina sequencing.

Whole genome sequencing (WGS) of the enriched chloroplast DNA showed that just 21% of the isolated DNA mapped to the chloroplast genome, however, depth of coverage was sufficient enough that this was not an issue. The WGS project produced a complete chloroplast genome for two geographically isolated accessions (RA83, Rekohu/Chatham Islands and 1162, Kermadec Islands) and highlighted two single nucleotide polymorphisms, one in the *ndhA* intron (SNP2) and another in the *psbB* gene (SNP3) between the two. This approach to sequencing the chloroplast genome of karaka provided a fast and efficient protocol for obtaining whole chloroplast genome sequences for seed plants. This protocol has since been used to sequence chloroplast genomes in *Trithuria inconspicua* (Goremykin *et al.*, 2012) and *Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis* (Zhong *et al.*, 2011) (these articles form part of the appendices for this thesis) and ongoing work on *Pachycladon* species. The method has applications for sequencing the chloroplast genomes of other angiosperms and gymnosperms in New Zealand and beyond. The article detailing this method, presented as chapter two, was published in BMC Plant Methods (IF 2.83) and has been cited 20 times (Web of Science, accessed 15th July 2014).

5.1.3 HRM screening

HRM has the capacity as a method for screening accessions for SNPs and mutations can be detected without direct sequencing (Dang *et al.*, 2012; Yan *et al.*, 2012). HRM screening of close to 350 karaka accessions to assist in the determination of chlorotypes was a fast and efficient method. However, it was not without its limitations.

Success of our HRM reactions relied upon the quality of the extracted genomic DNA. Whilst CTAB is a very simple and effective method of DNA extraction, the resulting DNA can be of varying quality between accessions, and with plants such as karaka, several secondary compounds can add to the range of quality of DNA. The potentially

uneven and low quality of our DNA template may thus have had an impact on subsequent melting analysis, generating system errors between melting temperature (T*m*˚) readings of HRM assays. For this study, Sanger sequencing was used to obtain data where HRM data was missing or incorrect).

When all marker data were combined, just six chloroplast haplotypes could be defined. Whilst this allowed for some of our research questions to be answered, the resolution was not high enough to elucidate the exact location of the source of Chatham Islands karaka nor determine the natural range of the tree. Our data suggests the karaka on Chatham Islands could come from a number of locations across New Zealand, including.

The sequence variability of the karaka chloroplast genome was investigated as a potential source for seed dispersal markers. Sufficient resolution in the data enabled an evaluation of the phylogeographic distribution of karaka to provide insight into the extent of human-mediated dispersal of the tree in New Zealand.

The results of the analysis of species-specific markers show the potential of the chloroplast genome to study recent events in plant history. They show an interesting relationship between Kermadec Island karaka and mainland karaka, and between Chatham Islands karaka and mainland. To be able to pinpoint the location of the source for Chatham Islands karaka, more genetic work is required. However, these results are promising in their ability to trace the translocation of one of New Zealand's most important ethnobotanical species.

By developing a more detailed picture of the genetic variation of karaka, this work has the potential to be the foundation for a deeper study into the translocation of the species. This has implications for further understanding the level of domestication in karaka, which at present cannot be ascertained.

5.2 FUTURE DIRECTIONS

The work presented in this thesis raises some interesting questions which have the potential to broaden the study of the karaka tree in New Zealand:

1. Is it possible to find more genetic variation in karaka by looking at different marker systems?
2. Was karaka brought into cultivation and if so, was it once or multiple independent times?
3. Do the locations of the initial domestication sites, inferred from patterns of karaka chloroplast, or nuclear, sequence variation, correspond to known early settlement sites?
4. Do routes of karaka translocation correlate with oral traditions of linkages between iwi?
5. Did Māori select for desirable characteristics in cultivated karaka? And if so, to what extent?

5.2.1 DEVELOPMENT OF MICROSATELLITE MARKERS

As well as providing chloroplast sequence data, Illumina GAII sequencing of the chloroplast genome also generated nuclear sequence data (discussed in Chapter Three). Although it was considered a contaminant in the chloroplast data, it has the potential to provide additional useful information. These nuclear ribosomal DNA (rDNA) reads were used to search for microsatellite markers for further studies. Of 13 microsatellites discovered, six amplified DNA from karaka. These six will be tested further to begin to develop a suite of microsatellite markers. Second-generation sequencing technology could be used to develop a further set of microsatellite markers and use these to test hypotheses relating to determining the natural range of karaka in New Zealand, following a similar methodology to Avery *et al.* (2013) when determining whether Bermuda songbirds achieved their current distributions via direct or indirect human actions.

Ribosomal DNA (rDNA) is multicopy, and as such requires much lower coverage than single copy nuclear genes (Kane *et al.*, 2012). The whole rDNA repeat evolves in a concerted way and polymorphisms in concerted sequences are found within, rather than between species (Hillis & Davis, 1988). This makes rDNA a suitable source for further molecular marker development with the potential to be highly informative at the population-level (Kane *et al.*, 2012)

5.2.2 Double-Digest Restriction Associated DNA sequencing (ddRADSeq)

Restriction Associated DNA sequencing (RADSeq) is a form of reduced-representation sequencing providing an efficient genotyping method for a large number of accessions. The approach allows oversequencing of those nucleotides adjacent to restriction sites and the detection of SNPs (Baird *et al.*, 2008). Choice of restriction enzyme determines the size of the fragments and therefore the number of potential markers, and multiple enzymes can be employed (Baird *et al.*, 2008). The RADSeq approach is a relatively simple concept: restriction enzymes are use to shear the genome into random fragments, P1 adapters are then ligated to the sticky overhanging ends which contains forward amplification and sequencing primers sites specific to the Illumina platform along with a 4-5 bp nucleotide barcode which is used to identify individuals within the pooled sample (Baird *et al.*, 2008).

Once adapters have been ligated to the fragments and the fragments pooled, randomly sheared and size selected the DNA is then ligated to the second (P2) adapter on the Illumina flowcell (Baird *et al.*, 2008) (Figure 5.1). Further development of this reduced-representation method has led to the use of two restriction enzymes resulting in a five-fold reduction in the cost of library production (Peterson *et al.*, 2012). Peterson, *et al* (2012) report the cost of library production as US$5 per sample, some US$20 cheaper than traditional RADSeq, further evidence of the cost-effectiveness of the method, especially for smaller research groups.

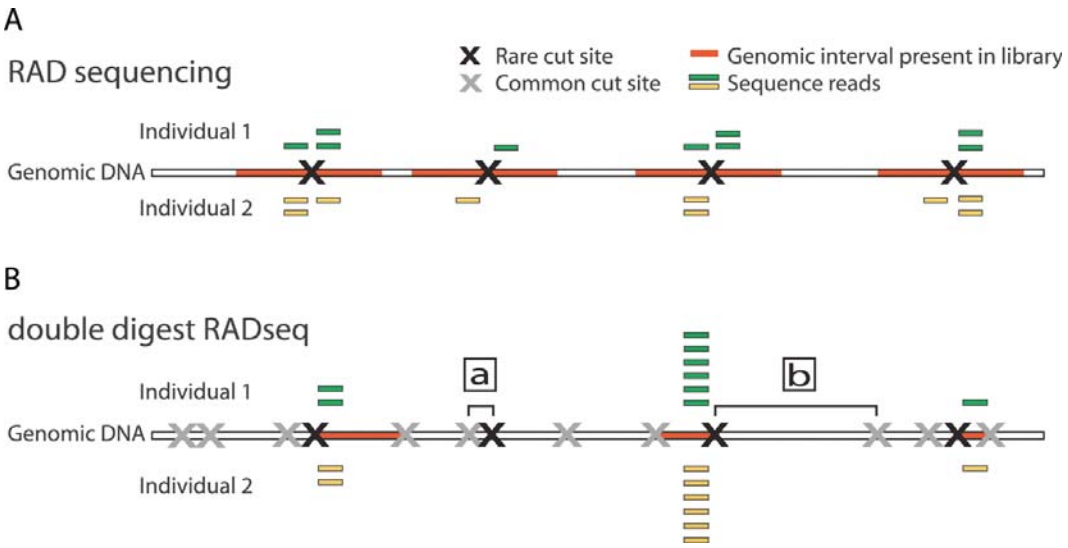

FIGURE 5.1: Double digest RAD sequencing improves efficiency and robustness while minimizing cost. (A) Traditional Restriction-Site Associated DNA sequencing (RADSeq) uses a single restriction enzyme (RE) digest coupled with secondary random fragmentation and broad size selection to generate reduced representation libraries consisting of all genomic regions adjacent to the RE cut site (red segments). (B) Double digest RAD sequencing (ddRADseq), by contrast, uses a two enzyme double digest followed by precise size selection that excludes regions flanked by either [a] very close or [b] very distant RE recognition sites, recovering a library consisting of only fragments close to the target size (red segments). Representation in this library is expected to be inversely proportional to deviation from the size-selection target, thus read counts across regions are expected to be correlated between individuals (yellow and green bars). Figure reproduced from Peterson *et al.* (2012).

5.2.3 CIRCOS PLOTS AND HOTSPOT REGIONS

In Chapter 4 it was suggested that more could be made of chloroplast genome variation. A cost effective way forward would be to specifically characterise chloroplast hotspot regions. Following Ahmed *et al.* (2012), these can be predicted by the distribution of repeats sequences. Primers have already been developed for these regions, which will be used to genotype and attempt to bring more resolution to purple and yellow haplotypes that might then allow testing the specific oral histories mentioned in Chapter 4.

5.2.4 AMPLIFYING CHLOROPLAST GENOMES USING REPLIPHI™ PHI29 DNA POLYMERASE

Using similar methodology to RepliG (Qiagen), Repliphi™ PHI29 DNA polymerase reagent kit (Epicentre Technologies Corp., Chicago) utilises hexanucleotides, which preferentially amplify AT-rich sequences. Karaka chloroplast DNA for a Chatham Island

accession has already been amplified using the Replify kit and sequenced using the Illumina MySeq (Illumina, San Diego, USA). Six accessions sharing the same haplotype as Rekohu/Chatham Island karaka will be sequenced using this method to pinpoint the exact location of the source for trees on this island group.

5.2.5 Exome capture

Exome capture is another form of high throughput reduced representation sequencing. Exome capture is targeted sequencing that sequences just the exome, the mRNA-coding portion of the whole genome. Next generation sequencing (NGS) has the capacity to sequence many of regions of interest, some of which, in plants, is highly repetitive. Therefore, much of the resulting NGS data, while interesting, is, ultimately, of no use or irrelevant (Grover *et al.*, 2012). Exome capture of barley (*Hordeum vulgare*) had the ability to reduce the nuclear genomic complexity more than 50-fold, which, in turn, dramatically reduced the sequencing and analysis workload for this species (Mascher *et al.*, 2013). Reducing the sequencing space using this strategy has several benefits: because sequences are reduced, sample multiplexing becomes possible, which further reduces the costs; the complexity of analysis is reduced by targeting only the portion of the genome that is necessary for the study; and finally, the sequencing depth afforded by targeted NGS increases the likelihod of identifying both the orthologs, and its paralogs, in population and infraspecific genomics assays (Grover *et al.*, 2012).

5.2.6 Oral histories

Māori and Moriori korero have played a role in recording the history of karaka in New Zealand. Māori oral histories make mention of the arrival of karaka in New Zealand with different waka. A full and in depth study of oral histories has the potential to compliment this genetic study further elucidating the cultural history of karaka in New Zealand.

The additional techniques presented as future work will add important temporal, spatial, genetic and cultural dimensions to a future study of karaka in New Zealand.

5.3 REFERENCES

Ahmed, I., Biggs, P., Matthews, P., Collins, L., Hendy, M., and Lockhart, P. 2012. Mutational dynamics of aroid chloroplast genomes. *Genome Biology and Evolution* **4:** 1316-1323.

Avery, J. D., Fonseca, D. M., Campagne, P., and Lockwood, J. L. 2013. Cryptic introductions and the interpretation of island biodiversity. *Molecular Ecology* **22:** 12.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **3**.

Coart, E., van Glabeke, S., De Loose, M., Larsen, A. S., and Roldan-Ruiz, I. 2006. Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Molecular Ecology* **15:** 2171-2182.

Dang, X.-D., Kelleher, C., Howard-Williams, E., and Meade, C. 2012. Rapid identification of chloroplast haplotypes using High Resolution Melting analysis. *Molecular Ecology Resources* **12:** 894–908.

Feleke, Y., Pasquet, R., and Gepts, P. 2006. Development of PCR-based chloroplast DNA markers that characterize domesticated cowpea (*Vigna unguiculata* ssp. *unguiculata* var. *unguiculata*) and highlight its crop-weed complex. *Plant Systematics and Evolution* **262:** 75-87.

Fu, Y.-B., and Allaby, R. 2010. Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. *Genetic Resources and Crop Evolution* **57:** 667-677.

Goremykin, V., Nikiforova, S., Biggs, P., Zhong, B., DeLange, P., Martin, W., Woetzel, S., Atherton, R., McLenachan, T., and Lockhart, P. 2012. The evolutionary root of flowering plants. *Systematic Biology* **62:** 50–61.

Grover, C. E., Salmon, A., and Wendel, J. F. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis1. *American Journal of Botany* **99:** 312-319.

Hillis, D. M., and Davis, S. K. 1988. Ribosomal DNA: intraspecific polymorphism, concerted evolution, and phylogeny reconstruction. *Systematic Zoology* **37:** 63-66.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., Engels, J. M., and Cronk, Q. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am J Bot* **99:** 10.

Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A., Clissold, L., Sampath, D., Ayling, S., Steuernagel, B., Pfeifer, M., D'Ascenzo, M., Akhunov, E. D., Hedley, P. E., Gonzales, A. M., Morrell, P. L., Kilian, B., Blattner, F. R., Scholz, U., Mayer, K. F., Flavell, A. J., Muehlbauer, G. J., Waugh, R.,

Jeddeloh, J. A., and Stein, N. 2013. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *Plant J* **76:** 494-505.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* **19:** 325-354.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. 2012. Double Digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., and Liu, W. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92:** 142–166.

Wills, D. 2006. Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.). *Journal of Heredity* **97:** 6.

Yan, G., Lv, X., Lv, P., Xu, K., Gao, G., Chen, B., and Wu, X. 2012. Application of high-resolution melting for variant scanning in chloroplast gene *atpB* and *atpB-rbcL* intergenic spacer region of Crucifer species. *African Journal of Biotechnology* **11:** 7016-7027.

Zhang, Y., Fang, Z., Wang, Q., Yang, L., Zhuang, M., and Sun, P. 2012. Chloroplast subspecies-specific SNP detection and its maternal inheritance in *Brassica oleracea* L. by using a dCAPS marker *Journal of Heredity* **103:** 6.

Zheng, Y.-H., Alverson, A., Wang, Q.-F., and Palmer, J. 2013. Chloroplast phylogeny of *Cucurbita*: Evolution of the domesticated and wild species. *Journal of Systematics and Evolution*.

Zhong, B., Deusch, O., Goremykin, V., Penny, D., Biggs, P., Atherton, R., Nikiforova, S., and Lockhart, P. 2011. Systematic error in seed plant phylogenomics. *Genome Biology and Evolution* **3**.

Zurawski, G., and Clegg, M. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. *Annual Review of Plant Physiology* **38:** 391-418.

Permission to work on New Zealand taonga[1]

1. **Resource consent** - Before beginning research on karaka it was appropriate to begin a cross-cultural dialogue with tangata whenua (native inhabitants) of New Zealand. As part of the consultation process it was important to recognise the kaitiaki (guardian) status of the different iwi (tribes) and hapū (subtribes) around the country. Our consultation with Māori iwi and hapū was more than a process to comply with moral obligations or simply asking for permission to conduct research, it also provided an open forum in which to raise and discuss any issues regarding possible benefits or risks the proposed research may have on Māori culture and well-being. Additionally, it brought a whole new dimension to our research objectives, as we were able to utilise and record valuable Mātauranga Māori (Māori traditional knowledge). The research does not affect Māori culture and traditions or the relationship tangata whenua have with their ancestral lands, water, sites, wahi tapu (sacred places), valued flora and fauna, mahinga kai (food gardens) or other natural resources or taonga. During these meetings, both researcher and iwi representatives had an opportunity to discuss the pros and cons of such research and each party suggested the conditions of consent that could be applied once consent was granted.

2. **Consultation process** – In the first instance it was necessary to determine whether the proposed research had the potential to directly affect iwi or hapū. Often it is the tangata whenua who know whether this is an issue or not, but because it involved taonga, a consultation process was necessary (Wilcox *et al.*, 2008). We contacted the regional Department of Conservation (DOC) office for advice and they gave us contact details for iwi and hapū in whose rohe (tribal boundary) populations of karaka occurred. On our behalf DOC sent out details of the project to all iwi and hapū requesting contact be made should there be any reason to refuse to grant the permit. From this initial mail-out we made contact with several iwi and hapū across the country to discuss our project in greater detail. As a result, I attended thirteen meetings with iwi and hapū to discuss our proposed project.

[1] Taonga – cultural treasures

Motivation for sampling strategy

Sample sites were identified to represent the key characteristics of the populations being studied. Effective sampling allows the determination of the levels and distribution of genetic variation in the natural distribution of karaka. Whilst an idea of the genetic variation in a natural population is helpful before sampling it is often the case that the population structure for the target population is unknown in advance of sampling (Gapare *et al.*, 2008).

1. Target population - To select initial sampling sites we used the appendix from the MSc thesis by Chris Stowe (2003), Auckland Museum records, Landcare Research, Te Papa Tongarewa Museum of New Zealand, Victoria University Herbarium and information gathered from iwi meetings. Samples were collected from Department of Conservation (DOC) scenic reserves with permission from DOC (permit numbers WA-23814-FLO, NO-23360-FLO and BOP-23814-FLO). Collections took place in QE2 covenants with permission from QE2 National Trust and the relevant landowners. In some cases samples were collected on private land with permission from the landowners and from native plant garden such as Otari Native Botanic Garden/Wilton's Bush (permit number 145) in Wellington. We chose populations across the country that were easily accessible and a good representation of the distribution.

2. Accessible population - For collecting in scenic reserves, those individuals within easy reach of a known thoroughfare and those whose branches were low enough were sampled. A maximum of eight randomly chosen samples were taken per population in a reserve to adequately sample variation present in the population.

3. Eligibility criteria - Samples were taken from trees of mature size showing no signs of disease or pest attack. Leaf material had to be young, fresh and supple and there needed to be enough material for both DNA extraction and a herbarium sample.

Samples consisted of small 5-8cm^2 pieces of young, fresh, healthy leaf material placed immediately in clean labelled plastic bags with a herbarium specimen approximately 25cm in length and representative of the sampled tree. At the end of a collection day the small leaf pieces are placed in silica gel and labelled with a collection number. Herbarium specimens were pressed in newspaper in a plant press with a label. The newspaper was periodically changed as necessary and replaced with dry newspaper. Once the plant press and herbarium specimens arrived at Massey University they were oven dried for three days at 40^{o}C to dry up any residual dampness.

REFERENCES

Gapare, W. J., Yanchuk, A. D., and Aitken, S. N. 2008. Optimal sampling strategies for capture of genetic diversity differ between core and peripheral populations of *Picea sitchensis* (Bong.) Carr. *Conservation Genetics* **9:** 411-418.

Stowe, C. J. 2003. *The ecology and ethnobotany of karaka (Corynocarpus laevigatus) [MSc. thesis].* University of Otago, Dunedin.

Wilcox, P. L., Charity, J. A., Roberts, M. R., Tauwhare, S., Tipene-Matua, B., Kereama-Royal, I., Hunter, R., Kani, H. M., and Moke-Delaneyz, P. 2008. A values-based process for cross-cultural dialogue between scientists and Māori. *Journal of the Royal Society of New Zealand* **38:** 215-227

Table A2.1: Geographical location of karaka (*Corynocarpus laevigatus*) accessions sampled across New Zealand

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		1005		Boat Cove, Kermadec Islands	S	29	16	42.03	E	177	53	43.00	-	5
		1006		Low Flat, Kermdec Islands	S	29	15	13.07	E	177	55	33.45	-	5
		1007		Maipito Rd, Swamp forest Chatham Is	S	44	5	42.79	E	176	36	38.07	-	4
		1015		Caravan Bush, Pitt Island	S	44	14	24.29	E	176	13	25.87	-	4
		1026		Blackhead, Stingray Bay. Next to Info signs	S	40	9	29.17	E	176	50	28.61	-	4
		1028		Blackhead, 1km inland on road	S	40	9	25.51	E	176	50	34.33	-	6
		1033		Blackhead, 1km inland on road	S	40	9	25.51	E	176	50	34.33	-	6
		1087		Goose Bay- Omihi	S	42	29	4.79	E	173	31	39.52	5	6
		1107		McKee Memorial Reserve, Nelson.	S	41	12	49.59	E	173	5	3.78	-	4
		1109		McKee Memorial Reserve, Nelson.	S	41	12	42.88	E	173	5	4.35	10	4
		1119		Abel Tasman Track, Wainui Bay	S	40	48	16.05	E	172	57	12.37	10	4
		1125		Abel Tasman Memorial, west end Tata beach	S	40	49	15.20	E	172	54	11.49	2	6
✓	✓	1127		Abel Tasman Memorial, west end Tata beach	S	40	49	14.94	E	172	54	10.55	2	6
	✓	1130		Hanson Winter Scenic Reserve	S	40	49	54.38	E	172	53	20.43	5	6
		1136		Puponga, Farewell Spit	S	40	30	43.77	E	172	44	17.97	15	6
	✓	1140		Patarau River Mouth	S	40	38	41.89	E	172	25	50.49	30	6
		1141		Patarau River Mouth	S	40	38	41.89	E	172	25	50.49	-	6
✓	✓	1162		Otari-Wilton Bush, ex. Kermadec Islands	S	29	14	37.83	E	177	58	2.79	-	5
		1172		Elsthorpe Scenic Reserve, Hawkes Bay	S	39	55	5.94	E	176	49	9.71	50	4
✓	✓	1185		Fantail Bay, Coromandel Peninsula	S	36	31	31.34	E	175	19	44.51	-	6
✓	✓	1296		Bream Head, Whangarei	S	35	50	45.74	E	174	34	48.92	20	6
		1299		Bream Head, Whangarei	S	35	50	58.34	E	174	34	46.42	-	6
		1311		Maitai Bay	S	34	51	17.60	E	173	25	36.23	-	4
		1314		Te Arai	S	34	41	29.36	E	172	56	20.57	-	4
		1319		Whangape Ridge	S	35	21	50.71	E	173	13	8.16	-	4

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		1321		Whangape Ridge	S	35	21	54.51	E	173	13	23.63	-	3
✓	✓	1345		Near junction of Awanui River and Waihou Stream, near Kaitaia	S	35	6	27.62	E	173	13	35.12	-	3
		1444		Moanaroa QE2 covenant	S	40	35	59.11	E	176	24	38.13	-	6
		1445		Moanaroa QE2 covenant	S	40	35	58.04	E	176	24	36.89	-	6
		1446		Near Mara	S	40	41	27.83	E	176	15	11.34	-	4
		1448		Near Mara	S	40	41	13.62	E	176	14	47.04	-	6
		1477		Rimatakas, near Pseudopanax ferox site	S	41	22	10.19	E	175	2	35.06	-	4
		1648		Tongaporutu river hillside	S	38	49	4.86	E	174	35	26.10	-	6
		1687		Between Nuhaka and Wairoa	S	39	2	23.59	E	177	33	53.68	-	4
		1692		Mahia Peninsular Scenic Reserve	S	39	7	31.99	E	177	52	43.94	-	6
		1981		Limestone Creek coastal forest	S	42	3	11.20	E	171	21	57.23	-	6
		1982		Limestone Creek coastal forest	S	42	3	11.20	E	171	21	57.23	-	6
		2285		Waipatiki Scenic Reserve	S	39	16	59.19	E	176	58	0.06	14	4
		2289		Waipatiki Scenic Reserve	S	39	16	41.64	E	176	57	57.84	180	6
		2290		Mohaka Township bridge	S	39	7	5.66	E	177	11	3.53	20	4
		4650		south of Kaikoura, Goose Bay, Omihi Scenic Reserve	S	42	29	9.27	E	173	31	31.50	20	4
		4651		North of Kaikoura, Ohau Stream track.	S	42	14	40.80	E	173	49	48.09	20	6
		4787		Opua, Harrisons Scenic Reserve	S	35	19	-	E	174	6	-	-	1
		4799		North-east of Whangarei, Matapouri, Whale Bay	S	35	34	-	E	174	30	-	60	5
		4800		North-east of Whangarei, Matapouri, Whale Bay	S	35	34	-	E	174	30	-	60	6
		4882		North Taranaki bight, Oakura, coastal walkway from Oakura camping ground to Ahuahu Road.	S	39	7	-	E	173	56	-	5	6
		4914		Waihi Beach, Orokawa Scenic Reserve.	S	37	23	-	E	175	56	-	60	4
		4933		Northern Coromandel Peninsula, near Port Jackson, Fantail Bay.	S	36	32	-	E	175	20	-	20	6
		5002		Southern Wairarapa, near Pirinoa, junction of Pirinoa and Whakatomotomo Roads.	S	41	22	-	E	175	12	-	40	4
		5042		Karikari Peninsula, northern end of Tokerau Beach.	S	34	52	-	E	173	23	-	10	4
		5078		Western Puketi Forest, Omahuta, alongside Kauri Sanctuary Road.	S	35	14	-	E	173	37	-	260	1

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		5085		Southern Puketi Forest, track from northern end of Forest Road, alongside Waipapa River.	S	35	17	-	E	173	41	-	40	6
		5091		Near Waiwera, Wenderholm Regional Park	S	36	32	12.00	E	174	42	42.00	40	6
✓	✓	5314		South of the Hokianga Harbour, near the Waimamaku River mouth, Waimamaku Beach Road.	S	35	35.25	-	E	173	25	-	5	4
		5707		Between Wimbledon and Herbertville, alongside Herbertville Road.	S	40	27	27.00	E	176	31	19.00	40	4
		6007		Southern Wairarapa, Tuhitarata, Tuhitarata Bush Scenic Reserve	S	41	17	41.66	E	175	16	18.60	-	4
		6241		Between Pongaroa and Akitio, Te Tumu covenant	S	40	34	33.13	E	176	19	41.97	-	6
		6277		Near Paraparaumu, scarp east of Raumati South, Mataihuka, walk from Valley Road.	S	40	56	57.27	E	174	59	29.22	60	6
		RA02		Mount William Scenic Reserve, Kauri Ridge Track	S	37	13	1.00	E	175	1	40.00	119	6
		RA04		Okura Estuary Scenic Reserve	S	36	40	26.87	E	174	42	46.85	90	6
		RA10		Okura Estuary Scenic Reserve	S	36	40	27.52	E	174	42	28.50	5	6
		RA11		Pakiri Reserve	S	36	12	24.82	E	174	39	7.49	11	6
		RA14		Pakiri Reserve	S	36	12	44.82	E	174	38	46.23	31	6
		RA16		Coastal reserve near to Goat Island Marine Reserve	S	36	16	23.08	E	174	48	12.58	58	6
		RA17		Coastal reserve near to Goat Island Marine Reserve	S	36	16	23.61	E	174	48	13.67	59	6
		RA20		Mahurangi Scenic Reserve	S	36	29	14.47	E	174	43	41.58	sea level	6
		RA21		Mahurangi Scenic Reserve	S	36	29	17.22	E	174	43	39.72	sea level	6
		RA22		Mahurangi Scenic Reserve	S	36	29	1.45	E	174	43	37.72	2	6
		RA23		Te Henga Scenic Reserve, Bethell's Beach	S	36	53	32.93	E	174	26	37.32	37	6
		RA25		Mount Auckland Walkway	S	36	27	2.46	E	174	26	53.98	190	6
		RA26		Mount Auckland Walkway	S	36	27	2.23	E	174	26	54.34	186	6
		RA29		Mount Auckland Walkway	S	36	26	56.06	E	174	26	59.19	220	6
✓	✓	RA31		Tauwhare Pa Scenic Reserve, Ohope, Whakatane	S	37	59	3.14	E	177	4	13.01	-	6
		RA33		Castlepoint Scenic Reserve	S	40	55	3.55	E	176	13	16.02	115	6
	✓	RA38		Castlepoint Scenic Reserve	S	40	55	5.36	E	176	13	15.13	142	6
		RA41		Okains Bay, private land owned by Murray Thacker	S	43	41	54.77	E	173	4	18.63	18	4

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA42		Okains Bay, private land owned by Murray Thacker	S	43	41	54.97	E	173	4	18.00	-	4
		RA44		Okains Bay	S	43	41	53.03	E	173	4	16.84	25	4
		RA49		Long Lookout Point, near Raupō Bay, Banks Peninsula	S	43	39	30.15	E	173	1	30.66	71	4
	✓	RA50		Long Lookout Point, near Raupō Bay, Banks Peninsula	S	43	39	30.09	E	173	1	29.72	88	4
		RA58		North side of Long Lookout Point, near Raupō Bay, Banks Peninsula	S	43	39	11.51	E	173	2	6.57	-	4
	✓	RA64		Whangaruru North Head	S	35	21	45.04	E	174	21	36.73	25	5
		RA65		Tauranga Bay	S	35	0	25.02	E	173	46	55.75	15	4
	✓	RA72		Otito Scenic Reserve	S	35	33	44.78	E	174	29	57.82	15	5
		RA79		Mangawhai Heads Reserve	S	36	3	28.49	E	174	35	26.35	14	6
		RA82		Seedling from Murray Thacker Okains Bay	S	43	42	37.96"	E	173	2	29.02	-	4
✓	✓	RA83		Seedling from Chathams (Harold Pierce Reserve) collected by PJ de Lange	S	43	55	51.64	E	176	30	52.01	-	4
	✓	RA84		Fanal Island, Hauraki Gulf	S	35	56	26.12	E	175	8	49.98	-	6
	✓	RA86		Papaitonga Scenic Reserve	S	40	38	53.47	E	175	13	50.81	7	4
		RA90		Papaitonga Scenic Reserve	S	40	38	54.02	E	175	13	58.92	19	6
		RA98		Pukerua Scenic Reserve, near Wellington	S	41	1	48.37	E	174	52	38.27	14	4
	✓	RA99		Pukerua Scenic Reserve, near Wellington	S	41	1	48.03	E	174	52	57.36	2	4
		RA100		Oaru Bay, near Kaikoura	S	42	31	46.47	E	173	30	5.72	5	4
		RA101		Oaru Bay, near Kaikoura	S	42	31	46.15	E	173	30	5.45	1	4
✓	✓	RA103		Oaru Bay, near Kaikoura	S	42	32	6.31	E	173	30	12.31	3	6
		RA111		Awapuni Racecourse, Palmerston North	S	40	22	59.22	E	175	34	24.73	8	4
		RA113		Awapuni Racecourse, Palmerston North	S	40	22	57.96	E	175	34	24.73	8	4
	✓	RA117	AK305672	Kermadec Islands Nature Reserve, Raoul Island, Ravine 8, "Hebe Site"	S	29	15	-	E	177	56	-	-	5
	✓	RA119	AK305674	Kermadec Islands Nature Reserve, Raoul Island, Fishing Rock Road, Crater Wall near Ammunition Store Cliffs	S	29	15	-	E	177	55	-	-	5
✓	✓	RA123		Hot Water Beach	S	36	53	24.90	E	175	49	19.67	-	6
✓	✓	RA124		Flatpoint, private land (Flatpoint Station)	S	41	14	24.47	E	175	57	33.83	3	6
		RA131		Flatpoint, private land (Flatpoint Station)	S	41	15	42.38	E	175	53	44.04	10	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
	✓	RA133		Flatpoint, private land (Flatpoint Station)	S	41	9	30.32	E	175	47	27.34	320	6
		RA134		Flatpoint, private land (Flatpoint Station)	S	41	9	30.32	E	175	47	27.34	320	6
		RA136	AK307589	East of Tokatoka, Maungaraho Rock Scenic Reserve	S	36	1	23.00	E	173	58	32.00	120	6
		RA137	AK307591	State Highway 2, c.1 km west of Maungaturoto	S	36	7	11.00	E	174	20	17.00	40	6
		RA138	AK307611	Waimamaku River, Kaikai Beach	S	35	35	25.00	E	173	24	50.00	10	4
	✓	RA139	AK307613	Maropiu, Te Kawa Stream	S	35	48	45.00	E	173	43	40.00	10	4
		RA141	AK304285	Ihumatao, Oruarangi Creek	S	36	59	9.00	E	174	46	7.00	4	4
		RA145	AK304296	Waitemata Harbour, Kendall Bay, Kauri Point	S	36	49	38.00	E	174	42	35.00	10	4
		RA146	AK307182	Coromandel Peninsula, State Higway 25, Tairua, River Estuary, Green Pt (Pa Site)	S	37	2	8.00	E	175	49	57.00	10	6
		RA147	AK307183	Coromandel Peninsula, Pauanui Walkway to Mt Pauanui	S	37	2	23.00	E	175	52	44.00	200	6
		RA148	AK307177	Tairua, Paku Hill	S	37	8	0.00	E	175	52	2.00	179	6
		RA149	AK304485	Te Akau, Te Akau South Road	S	37	44	38.00	E	174	52	10.00	60	4
		RA150	AK304496	Te Akau Coast Road, above Waimai Stream	S	37	38	47.00	E	174	49	48.00	20	4
		RA151	AK304502	Waikato River, just north of Tuakau Road Bridge	S	37	15	10.00	E	174	55	20.00	5	6
		RA152	AK304503	Buckland Road	S	37	15	10.00	E	174	55	20.00	20	6
		RA153		West Island Manawatawhi - Tasman Stream	S	34	10	19.96	E	172	8	33.47	-	1
		RA154		West Island Manawatawhi - Tasman Stream	S	34	10	19.96	E	172	8	33.47	-	1
✓	✓	RA155		West Island Manawatawhi - Tasman Stream	S	34	10	19.96	E	172	8	33.47	-	1
		RA158	AK304206	Bombay Hills, State Highway 1	S	37	13	18.00	E	175	1	0.00	140	6
		RA160	AK304208	Huntly Basin, Lake Waikare, eastern Shoreline	S	37	25	41.00	E	175	13	19.00	20	6
		RA161	AK304213	Huntly Basin, State Highway 1, Meremere near Waikato River	S	37	19	23.00	E	175	3	55.00	1	6
		RA162	AK304214	Huntly Basin, Lake Whangape, northern shoreline	S	37	24	31.00	E	175	2	52.00	20	6
		RA163	AK304215	Huntly Basin, Lake Kimihia	S	37	31	28.00	E	175	11	28.00	20	4
✓	✓	RA165	AK304224	Te Aroha, Tui Stream	S	37	31	25.00	E	175	42	58.00	140	6
		RA166	AK304225	Hauraki Plains, Tirohia, Waihou River	S	37	27	15.00	E	175	38	28.00	5	6
		RA167	AK304226	Totara, Totara Pa	S	37	10	3.00	E	175	33	26.00	10	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA168	AK304226	Foothills near Kaihere	S	37	21	56.00	E	175	25	8.00	40	6
		RA169	AK304230	Hauraki Plains, Bush Road, near Kopuarahi	S	37	14	55.00	E	175	31	7.00	3	6
		RA171	AK304234	Road	S	37	5	12.00	E	175	18	4.00	1	6
		RA172	AK304235	Orere Point, Orere Stream	S	36	57	39.00	E	175	14	52.00	2	6
		RA173	AK309379	State Highway 16, west of Kiwitahi Road	S	36	44	25.88	E	174	26	31.25	22	6
		RA174	AK309392	Kaukapakapa River, Rapsons Road Bridge	S	36	37	55.82	E	174	31	12.81	10	6
		RA177	AK309383	State Highway 16, near Kakanui	S	36	32	8.72	E	174	27	19.94	20	6
		RA179	AK304301	Maraetai, Magazine Bay	S	36	53	7.00	E	175	3	28.00	10	4
		RA180	AK304302	Hauraki Gulf, Motukaraka Island	S	36	52	41.46	E	174	58	43.78	10	6
		RA181	AK304329	Te Anga, Marokopa River, Ngahuinga Bluffs Scenic Reserve, Rock Shelter	S	38	15	29.00	E	174	50	6.00	40	4
		RA182	AK304330	Te Kauri Scenic Reserve, Devilin Track Rock Shelter	S	38	4	19.00	E	174	58	41.00	100	4
	✓	RA183	AK304331	Kihi Road, Pukenui	S	38	6	39.00	E	174	58	12.00	220	4
		RA184	AK304332	Kawhia Harbour Road, Hautapu	S	38	6	22.00	E	174	55	15.00	100	4
		RA185	AK304333	Waiharakeke, Grey Road, Bluffs above the Awaawaroa Stream	S	38	8	0.00	E	174	52	1.00	80	4
		RA186	AK304334	Taumatatotara Range, Whenuaapo Road, Whenuaapo Peak	S	38	8	54.00	E	174	53	12.00	240	4
		RA187	AK304335	Waitomo, Ruakuri Caves and Bush Scenic Reserve, main tourist entrance to Ruakuri	S	38	16	3.00	E	175	4	47.00	80	4
		RA189	AK304337	Kawhia Harbour, Ngatokakairiri Island (Pa)	S	38	3	26.00	E	174	52	51.00	10	6
		RA190	AK304338	Kawhia Harbour, Awaroa Scenic Reserve (Coastal), Mouth of the Awaroa River, Kotongareia Point	S	38	5	1.61	E	175	54	13.36	10	4
		RA195	AK304344	Ruapuke, Whaanga Road, Waitake Stream	S	37	54	18.34	E	174	54	13.36	60	4
		RA197	AK304486	Te Akau Wharf Road, near Te Akau Wharf	S	37	47	33.00	E	174	52	0.00	10	4
	✓	RA201		Caravan Bush, Pitt Island, Chatham Islands	S	44	14	22.21	E	176	12	53.23	-	4
	✓	RA204		Nikau Bush, Chatham Islands	S	43	51	38.06	E	176	32	4.00	-	4
		RA209	AK304493	Glen Murray - Rangiriri Road, "Lowry Kauri Forest" above Tikotiko Road	S	37	26	37.00	E	174	58	58.00	100	6
		RA210	AK304492	Port Waikato - Waikaretu Road	S	37	26	52.00	E	174	44	44.00	140	4
✓	✓	RA211	AK304494	Port Waikato Road, Waikato River, Okahu	S	37	22	2.00	E	174	45	42.00	2	6
		RA212	AK304504	Coromandel Peninsula, Kopu-Hikuai Road, Kirikiri Stream	S	37	10	46.00	E	175	35	9.00	40	4

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA214	AK308766	Whangamata, Hauturu (Clark's Island)	S	37	12	55.00	E	175	53	26.00	10	6
		RA215	AK308782	Whangamata, Whenuakura Island	S	37	13	14.00	E	175	53	47.00	20	6
		RA216	AK308798	Whangamata, Harbour Mouth (North Side), Te Karaka Point	S	37	12	10.00	E	175	53	3.00	10	4
		RA217	AK308170	Coromandel Peninsula, Omarupotiki Point	S	37	10	10.00	E	175	52	59.00	10	6
✓	✓	RA218	AK308166	Coromandel Peninsula, Rabbit Island (off Pauanui Coastline)	S	37	4	19.00	E	175	55	34.00	40	6
		RA219	AK308801	Pakaroa Range, Te Miro, Gray Road, near Ruru Hill	S	37	47	49.00	E	175	33	11.00	240	6
		RA220	AK308802	Mt Kakepuku, Mt Kakepuku Scenic Reserve	S	38	4	23.00	E	175	14	50.00	240	6
		RA224	AK308904	Bream Bay, Ruakaka	S	35	54	33.00	E	174	27	9.00	3	6
		RA225	AK308906	Bream Bay, Waipu Cove	S	36	1	51.00	E	174	30	30.00	4	6
		RA227	AK308912	Bream Bay, Langs Beach, Cove Road (above McKenzie Cove)	S	36	2	55.00	E	174	32	19.00	10	6
		RA228	AK308914	Mangawhai Heads Road, Mangawhai River near Mangawhai Heads	S	36	4	55.00	E	174	35	53.00	20	6
		RA229	AK308916	Whangarei Harbour, Marsden, One Tree Point, Pyle Road West	S	35	49	35.00	E	174	26	50.00	4	6
		RA231	AK308922	Bream Bay, State Highway One, Ruakaka River, Flyger Road	S	35	52	12.00	E	174	24	14.00	10	6
		RA232	AK308943	Ripiro Beach, Kaiiwi Track above beach, c.4.5 km south of Aranga Beach Settlement	S	35	48	48.00	E	173	36	50.00	20	4
	✓	RA233	AK308944	Aranga Beach Settlement, Maunganui Bluff Scenic Reserve, Maunganui Bluff	S	35	46	7.00	E	173	34	25.00	80	4
		RA234	AK308954	Ripiro Beach, Moremonui Gully, "The Monument"	S	35	53	48.00	E	173	41	40.00	20	4
		RA235	AK308955	State Highway 12, c.1 km north of Mititai	S	36	0	29.00	E	173	55	39.00	2	6
		RA236	AK308833	Waiotama, Wheki Valley, Wheki Stream	S	35	48	5.00	E	174	7	37.00	20	6
	✓	RA237	AK308962	Kaipara, State Highway 16, just north of Woodhill, Kaipara River	S	36	44	28.00	E	174	26	40.00	20	4
		RA239	AK308970	South Head Road, Lake Ototoa Scenic Reserve, Lake Ototoa	S	36	31	3.00	E	174	14	32.00	80	4
		RA240	AK308972	South Head Road, Lake Ototoa Scenic Reserve	S	36	29	35.00	E	174	14	42.00	120	4
		RA241	AK308976	Kaipara, South Head, Te Rokotai, upper Kawau Creek	S	36	29	6.00	E	174	14	53.00	100	6
	✓	RA242	AK308979	Kaipara, South Head, Mosquito Bay (Kawau Creek mouth), near Te Kawa Point	S	36	27	19.00	E	174	15	24.00	10	5
		RA243	AK308981	Kaipara, South Head, Lagoon Road, Waionui	S	36	27	4.00	E	174	12	47.00	120	4
		RA244	AK308982	Kaipara, Woodhill Forest, north of Rimmer Road	S	36	41	22.00	E	174	23	19.00	60	6
		RA246	AK306565	Okahukura Peninsula, near Run Road - Burma Road, Hiki Stream Scenic	S	36	22	54.00	E	174	22	12.00	20	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
				Reserve										
		RA261	AK306568	Okahukura Peninsula, Tauhoa - Port Albert Road, Whanaki River	S	36	20	45.51	E	174	25	41.12	1	6
		RA263	AK309398	Okahukura Peninsula, Oruawharo River, near Port Albert	S	36	16	25.58	E	174	26	7.71	1	6
		RA266	AK309408	Puketotara Peninsula, Oneriri	S	36	17	24.88	E	174	21	0.37	20	4
	✓	RA269	AK309413	Pahi, Pahi Road	S	36	9	1.28	E	174	13	22.24	20	6
		RA272	AK309431	Kaipara Harbour, Te Kiakia Bay, Kaiwhitu Island	S	36	14	27.18	E	174	11	20.08	20	6
		RA273	AK309434	Tinopai Road, Hukatere Scenic Reserve	S	36	11	16.79	E	174	9	52.62	125	6
		RA274	AK309436	Tinopai Road, upper Te Taumataka Creek	S	36	10	11.73	E	174	10	7.47	20	6
		RA275	AK309437	Wairoa River (east bank), State Highway 12, Donovan's Bluff	S	36	4	15.97	E	173	58	25.85	10	6
		RA277	AK309440	Montgomery Scenic Reserve	S	36	0	39.00	E	173	57	46.61	20	6
		RA278	AK309441	Pouto Peninsula, Pouto	S	36	22	5.04	E	174	10	44.18	20	4
		RA280	AK309451	Pouto Peninsula, Kellys Bay Road, Tangitiki Bay	S	36	14	5.80	E	174	4	35.23	2	6
		RA281	AK309455	Pouto Peninsula, Guys Road	S	36	4	10.00	E	173	56	11.00	20	6
		RA282	AK309456	Pouto Peninsula, Te Kopuru	S	36	1	32.54	E	173	54	59.68	20	6
		RA283	AK309458	Kaipara, State HIghway 12, Ruawai Flats, Naumai	S	36	5	23.64	E	173	59	10.84	3	6
		RA285	AK309462	Waipu Gorge, Waipu Gorge Scenic Reserve, Ahuroa River	S	36	3	35.63	E	174	23	23.63	40	6
		RA287	AK310021	Waiheke Island, Matiatia Bay, north of wharf	S	36	46	48.14	E	174	59	32.31	1	4
		RA296	AK310076	Waiheke Island, North end of Hooks Bay	S	36	44	34.60	E	175	10	25.00	2	2
		RA300	AK310089	Waiheke Island, Orapiu Bay	S	36	50	38.00	E	175	8	49.40	25	6
		RA302	AK310102	Waiheke Island, Onetangi	S	36	47	30.20	E	175	4	53.40	19	6
		RA304	AK310542	Waihi Beach, Rapatiotio Point	S	37	23	37.50	E	175	56	20.90	9	6
✓	✓	RA305	AK309834	Te Paki, Tomokanga, "Tomokanga Stream"	S	34	25	43.00	E	172	57	41.00	-	1
		RA306		Waitomo to Marakopa Road near Te Anga CDT Club sheds and fishing shack number 11052	S	38	17	36.74	E	174	44	54.17	6	4
		RA307		Waitomo to Marakopa Road near Te Anga CDT Club sheds and fishing shack number 11052	S	38	18	2.56	E	174	44	7.15	33	4
		RA311		Waitomo to Marakopa Road	S	38	18	0.32	E	174	44	0.71	10	3
		RA312		Private land, Marakopa Road	S	38	17	52.59	E	174	44	13.61	10	4

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
	✓	RA314		Private native reserve, Marakopa Road	S	38	17	43.70	E	174	45	1.22	8	4
		RA315		Private native reserve, Marakopa Road	S	38	17	44.41	E	174	45	1.11	12	4
		RA319		Waverley to Wanganui Road (State Highway 3)	S	39	49	36.26	E	174	52	54.98	105	4
✓	✓	RA320		Waverley to Wanganui Road (State Highway 3)	S	39	49	35.72	E	174	52	56.14	92	6
		RA322		Hapupu Reserve, Chatham Islands	S	43	48	7.20	E	176	21	10.50	13	4
	✓	RA324		Hapupu Reserve, Chatham Islands	S	43	48	3.06	E	176	21	13.26	22	4
	✓	RA325		Henga Reserve near Chathma Lodge, Chatham Islands	S	43	51	6.12	E	176	33	16.56	58	4
	✓	RA328		Blind Jim's Beach, on the Western Shore of Te Whanga Lagoon, Chatham Islands	S	43	47	0.24	E	176	33	12.18	18	4
	✓	RA331		North-east corner of Taia Farm, Chatham Islands	S	43.0	49.0	47.50	E	176.0	22.0	42.45	-	4
	✓	RA333		Private land at Waihi, Chatham Islands	S	43	46	4.44	E	176	48	29.46	8	4
	✓	RA335		Private land at Waihi, Chatham Islands	S	43	46	40.56	E	176	48	48.54	16	4
	✓	RA337		Mount Chudleigh, Chatham Islands	S	43	43	47.58	E	176	34	8.58	52	4
	✓	RA338		Nikau Bush, Chatham Islands	S	43	45	54.78	E	176	33	56.22	30	4
✓	✓	RA340		Ashhurst Domain	S	40	18	10.25	E	175	45	25.79	63	6
		RA341		Ashhurst Domain	S	40	18	6.43	E	175	45	30.72	68	6
		RA342		Ashhurst Domain	S	40	18	5.18	E	175	45	31.47	70	6
	✓	RA344		Horseshoe Bend Scenic Reserve, Tokomaru	S	40	29	26.98	E	175	31	35.61	30	6
		RA345		Horseshoe Bend Scenic Reserve, Tokomaru	S	40	29	24.67	E	175	31	33.92	23	4
		RA346		Te Karaka Grove (Te koha o te whenua), Plant Growth Unit, Fitzherbert Science Centre	S	40	22	39.17	E	175	36	47.71	23	6
		RA347		Waikaretu Valley Road	S	37	33	8.84	E	174	47	52.39	35	4
		RA349		Native bush belonging to Ann and Philip Woodward, Nikau Caves Café on Waikaretu Valley Road	S	37	32	45.36	E	174	48	33.18	85	6
		RA350		Waikaretu Valley Road 500m north from swing bridge on Swingbridge Walkway	S	37	32	51.69	E	174	48	40.68	57	4
		RA351		Limestone Downs on track towards Massey bush experimentla site	S	37	28	5.46	E	174	46	14.44	116	4
		RA356		Waikaretu Valley Road, Limestone Downs waterfall, 1km from Limestone Downs sign towards Port Waikato	S	37	28	20.45	E	174	45	34.76	78	4
		RA360		Waikaretu Valley Road, growing on plain at valley bottoom along	S	37	27	36.96	E	174	44	4.88	13	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
				Waikawai Stream towards the ocean										
		RA361		Port Waikato Road - Tuakau Bridge Road near Port Waikato. Just after bend before large pine plantation	S	37	23	10.80	E	174	44	19.80	10	6
		RA366		Awhitu Peninsula	S	37	8	53.32	E	174	35	48.24	80	4
		RA367		Awhitu Peninsula	S	37	8	43.59	E	174	35	48.24	80	4
		RA368		Awhitu Peninsula	S	37	8	43.59	E	174	35	48.04	80	4
		RA371		Just north of Pukerua on the edge of State Highway 1	S	41	1	4.68	E	174	54	51.04	4	6
		RA372		Just north of Pukerua on the edge of State Highway 1	S	41	1	4.99	E	174	54	50.60	5	4
		RA374		Between Pauatahanui and Plimmerton on Gray's Rd	S	41	5	11.20	E	174	53	25.53	19	6
		RA375		Between Pauatahanui and Plimmerton on Gray's Rd	S	41	5	11.03	E	174	53	26.01	10	6
		RA376		Aorangi Forest Park, Old Te kōpi ranger station	S	41	26	46.26	E	175	13	7.44	15	6
		RA377		Aorangi Forest Park, Old Te kōpi ranger station	S	41	26	46.14	E	175	13	6.90	15	4
		RA378		Cape Palliser, on Ngāti Hinewaka land	S	41	36	12.24	E	175	19	31.32	13	4
		RA381		Cape Palliser, on Ngāti Hinewaka land	S	41	36	9.36	E	175	19	33.78	12	4
		RA384		Bottom of Wharekauhau Road	S	41	22	38.64	E	175	4	39.84	6	4
		RA386		South of Battery Hill pa site on Western Lake Road	S	41	20	38.94	E	175	8	9.06	9	6
		RA388		Wilderness Busk AKA Karaka Bush, south of Wairongo mai River, southern Wairarapa	S	41	17	22.74	E	175	8	55.20	12	6
		RA392		Wilderness Busk AKA Karaka Bush, south of Wairongo mai River, southern Wairarapa	S	41	17	13.68	E	175	9	6.18	4	4
		RA398		Waipoua Bridge, Waipoua River Road, Waipoua Forest	S	35	39	15.12	E	173	33	51.90	92	6
		RA400	AK310545	Coromandel Peninsula, north of Waihi Beach, Te Puru Creek	S	37	22	38.50	E	175	56	14.70	12	6
		RA401	AK310546	Coromandel Peninsula, north of Waihi Beach, Orokawa Bay Track, above Oukori Stream	S	37	23	21.30	E	175	56	22.20	42	6
		RA402	AK310549	Coromandel Peninsula, north of Waihi Beach, Orokawa Bay	S	37	23	11.00	E	175	56	20.00	9	6
		RA403	AK310563	Coromandel Peninsula, Coromandel Peninsula, Homunga Bay, Fraser Creek	S	37	21	42.20	E	175	56	11.60	10	4
		RA404	AK310566	Tauranga Harbour, Athenree, Koutunui Road, Waiau Estuary	S	37	26	24.30	E	175	57	41.20	15	1
		RA405	AK310575	Tauranga Harbour, Ongare Point	S	37	29	45.70	E	175	57	49.80	3	6
		RA406	AK310578	Tauranga Harbour, Bowentown, Anzac Bay, near Papatu Point Pa	S	37	27	58.40	E	175	59	12.00	15	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
	✓	RA407	AK310579	Tauranga Harbour, Bowentown, Bowentown Heads Road	S	37	27	41.30	E	175	58	53.30	13	5
		RA408	AK310580	Waihi Beach, Wilson Road	S	37	24	38.10	E	175	56	22.30	14	4
		RA412	AK311210	Mt Maunganui, Mauao (Mt Maunganui)	S	37	38	3.80	E	176	10	25.20	10	4
		RA414	AK311216	Mt Maunganui, Moturiki Island	S	37	37	49.60	E	176	11	6.70	10	4
✓	✓	RA416	AK311227	Bay of Plenty, Pikowai	S	37	51	29.20	E	176	40	5.20	20	4
		RA417	AK311229	Bay of Plenty, Ohope, Ohope Sandspit	S	37	59	25.40	E	177	8	1.00	1	6
		RA419	AK311233	Bay of Plenty, Waiotahi Beach	S	37	59	31.20	E	177	14	1.00	20	6
	✓	RA420	AK311235	Bay of Plenty, Opotiki, Tablelands, Opotiki Trig	S	37	59	36.60	E	177	18	18.20	40	5
	✓	RA421	AK311237	Bay of Plenty, Hikuwai Beach, Tirohanga	S	37	59	22.80	E	177	20	53.50	8	6
		RA425	AK311307	State Highway 35 (East Cape Road), Whitianga Bay, Okawhiti Stream mouth	S	37	50	23.80	E	177	35	54.90	1	5
		RA426	AK311308	Omaio Bay, Otuwhare, Paerata Stream	S	37	48	40.10	E	177	38	33.30	1	6
		RA428	AK311311	State Highway 35 (East Cape Road), Awanui, above Te Muka Urupa	S	37	47	28.90	E	177	39	27.70	19	6
		RA429	AK311312	State Highway 35 (East Cape Road), Hariki Beach	S	37	45	37.30	E	177	41	6.10	18	6
		RA430	AK311313	State Highway 35 (East Cape Road), Puremutahuri Stream	S	37	45	1.30	E	177	40	57.90	12	6
	✓	RA431	AK311314	State Highway 35 (East Cape Road), Te Kaha, Wharekura Point	S	37	43	35.40	E	177	41	33.30	19	5
		RA434	AK311382	State Highway 35 (East Cape Road), Papatea Bay, west bank of the Raukokore River	S	37	40	26.60	E	177	52	15.90	20	6
	✓	RA435	AK311384	Papatea Bay, East Bank of Raukokore River	S	37	40	24.20	E	177	52	31.50	13	3
	✓	RA436	AK311387	Waihau Bay, Waihau	S	37	37	6.50	E	177	54	43.40	1	5
	✓	RA437	AK311392	Te Rangiharu Bay, Oruaiti Beach, Wairuru Stream	S	37	37	9.30	E	177	56	46.70	3	5
	✓	RA438	AK311393	Potikirua Road, Upokongaruru	S	37	32	54.50	E	178	6	40.60	10	6
		RA440	AK311407	State Highway 35 (Te Araroa Road), Hoia	S	37	35	29.80	E	178	14	32.10	17	6
		RA442	AK311258	Hicks Bay, near Kapokapo Bay	S	37	34	6.30	E	178	18	38.70	10	6
		RA443	AK311421	Hicks Bay, Onepoto Bay	S	37	35	35.90	E	178	17	56.10	10	6
		RA445	AK311431	East Cape, Waikuta Stream	S	37	42	13.50	E	178	32	12.80	35	4
		RA448	AK311436	Te Araroa - East Cape Road, Waipapa Stream	S	37	39	27.50	E	178	29	36.00	19	6
		RA449	AK311437	Te Araroa - East Cape Road, unnamed stream north of Nohomanga Stream	S	37	38	51.20	E	178	29	2.50	19	6

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA450	AK311445	Te Araroa - East Cape Road, Te Mangaroa	S	37	37	45.30	E	178	23	3.40	12	6
		RA454	AK311460	Anaura Bay, Anaura Road	S	38	15	1.90	E	178	18	37.50	53	6
		RA455	AK311462	Tolaga Bay (Uawa), cliffs at northern end of beach	S	38	21	43.40	E	178	18	18.70	24	6
		RA456	AK311464	Pouawa	S	38	36	20.90	E	178	11	12.20	14	6
		RA457	AK311465	Tatapouri	S	38	38	33.40	E	178	8	44.70	14	6
		RA458	AK311466	Makorori Beach	S	38	39	27.60	E	178	6	37.10	60	6
		RA460	AK311470	Whakatane River, Rewatu Road	S	38	1	21.00	E	176	59	7.40	19	4
		RA461	AK311476	Whakatane, Pohaturua Rock	S	37	57	5.20	E	176	59	48.60	10	6
		RA462	AK311477	Tahuna Road, Old Pa Site (South End)	S	38	3	46.70	E	176	47	59.80	37	6
		RA463	AK311483	Rangitaiki River, near Te Mahoe, Pa Site	S	38	5	41.00	E	176	48	48.50	56	6
✓	✓	RA464	AK311485	Gisborne, Turanganui River, near Cook Monument	S	38	40	31.40	E	178	1	33.60	17	6
		RA467	AK311956	Aotea (Great Barrier Island), Tryphena Harbour, Shoal Bay Road	S	36	18	57.40	E	175	29	29.70	20	6
		RA468	AK311959	Aotea (Great Barrier Island), Tryphena Harbour, Shoal Bay Road	S	36	18	35.10	E	175	29	35.90	20	6
		RA469	AK311960	Aotea (Great Barrier Island), Whangaparapara Harbour, Whangaparapara Wharf	S	36	14	40.00	E	175	23	52.70	10	6
✓	✓	RA470	AK311984	Aotea (Great Barrier Island), Harataonga Bay, Harataonga Stream	S	36	10	14.30	E	175	28	47.90	10	6
		RA473	AK313174	Near Lake Omapere, Tarahi Hill, Remuera Settlement Road, Pungatere Stream	S	35	21	40.40	E	173	51	25.02	240	6
		RA474		Catchpool Valley	S	41	21	12.76	E	174	54	27.77	-	4
✓	✓	RA475		Between Wainuiomata River and Orongorongo River	S	41	24	43.33	E	174	53	27.20	-	4
		RA476	AK316305	Waipoua Forest, State Highway 12, Waipoua River	S	35	39	9.13	E	173	34	12.15	92	6
		RA477	AK316308	North of Waimamaku River, Pukorokoro Stream	S	35	34	47.76	E	173	24	8.81	17	4
		RA478	AK316326	West of Dargaville, Baylys Beach	S	35	56	57.37	E	173	44	45.29	40	6
		RA479	AK316370	Kaihu Valley, Rotu River	S	35	52	41.55	E	173	47	45.88	13	6
		RA480	AK318723	Waipu River Catchment, North River, by Grant Road	S	35	57	46.47	E	174	20	50.67	98	6
		RA481	AK319043	State Highway 10, near Kerikeri, Puketona Scenic Reserve, upper Waitangi River	S	36	10	34.47	E	173	57	44.66	100	6
		RA482	AK319066	Te Paki, North Cape, North Cape Scientific Reserve, Ngawhenua Stream	S	34	24	5.44	E	173	0	6.63	24	1
		RA486		Te Paki, North Cape, North Cape Scenic Reserve, "Wasp Sting Bush"	S	34	24	25.00	E	173	2	9.00	111	1

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA501		Matthews Park Kaitaia. Kevin says it's a really old tree and former Māori settlement site	S	35	6	7.83	E	173	15	43.51	18	4
		RA504		At the base of Mount Camel above the urupa	S	35	6	7.83	E	173	15	43.51	-	3
	✓	RA511		Mangataipa Road, Northand	S	35	14	32.58	E	173	31	54.60	36	4
		RA512		Waiwera	S	36	32	19.18	E	174	42	18.47	58	6
		RA513		Waiwera	S	36	32	6.70	E	174	42	36.60	80	6
		RA514		Mahia Peninsula Scenic Reserve, Mahia Peninsula, Hawkes Bay	S	39	7	32.00	E	177	52	43.71	-	6
		RA516		Private property, Happy Jacks Road, Mahanga, Hawkes Bay	S	39	0	39.73	E	177	53	17.77	7	4
		RA517		Whangawehi Coronation Reserve, Whangawehi landing/river mouth, Mahia Peninsula	S	39	5	38.30	E	177	57	2.24	15	6
		RA518		Above Black's Beach, between Mahia and Wairoa	S	39	3	37.62	E	177	46	51.94	34	6
		RA519		Near Whakaki Lagoon, Whakaki, near Wairoa	S	39	2	12.79	E	177	33	13.78	12	4
		RA523		QEII covenant (Miriam's Gully) north of Patea near Karkaramea	S	39	40	18.95	E	174	24	9.21	74	4
		RA525		Hawera Intermediate School, State Highway 3, Hawera	S	39	35	37.07	E	174	16	43.82	108	6
		RA526		Hawera Intermediate School, State Highway 3, Hawera	S	39	35	39.65	E	174	16	46.04	115	6
		RA527		Oeo Stream, South Road (SH45) Kaupokonui, north of Hawera	S	39	31	42.36	E	173	57	5.22	31	4
		RA528		Puneho Stream, just south of Watino Road on South Road (SH45) 1km from Papaka Te Rangi historic site	S	39	29	45.89	E	173	54	23.11	26	4
		RA530		Taungatara Stream, just south of Mataikahawai Road on South Road (SH45)	S	39	28	52.61	E	173	53	29.05	25	4
	✓	RA533		Warea River, South Road (SH45)	S	39	14	19.25	E	173	48	23.46	47	4
		RA534		Kaihihi Stream, South Road (SH45)	S	39	11	13.54	E	173	52	17.94	66	4
		RA535		Maitahi Scientific Reserve	S	39	8	31.54	E	173	52	15.90	3	4
		RA538		Junction of Leith and Perth Roads, Taranaki (private farmland)	S	39	10	28.78	E	173	53	16.74	83	6
		RA539		Tataraimaka Historic Reserve, Lower Pitone Road	S	39	8	1.54	E	173	53	25.56	20	4
		RA540		Tataraimaka Historic Reserve, Lower Pitone Road	S	39	8	0.76	E	173	53	26.28	18	4
		RA542		Mangore Power Station Reserve, Hydro Road	S	39	6	17.85	E	174	6	57.24	149	6
		RA543		Meeting of the Waters National Park	S	39	6	12.81	E	174	6	54.36	75	4
		RA544		Te Henui Walkway, New Plymouth	S	39	4	3.39	E	174	5	41.16	10	4

Long-range	*HRM test*	*Accession no.*	*Herbarium no. where available*	*Locality Details*	*Latitude*				*Longitude*				*Alt*	*Chlorotype*
		RA547		Below Te Urenui Pa Historic Reserve, Avenue Road north of Urenui, State Highway 3	S	38	59	48.68	E	174	24	23.58	1	4
		RA548		Tongaporutu Conservation Area, Tongaporutu River, 15km south of Mokau	S	38	49	2.47	E	174	35	50.87	16	4
		RA549		On banks of Mohakatino River, 3km south of Mokau.	S	38	43	49.02	E	174	36	55.43	9	4
		RA551		South of Mokau on State Highway 3 opposite entrance to Taniui Wetere Domain	S	38	42	30.24	E	174	37	16.49	1	4
		RA552		South of Mokau on State Highway 3	S	38	43	36.72	E	174	37	3.05	17	6
		RA553		Te Kawau Pa Historic Reserve, off State Highway 3 south of Mokau	S	38	46	9.73	E	174	36	3.47	10	4
		RA554		Above Rapanui Stream on State Highway 3	S	38	47	57.73	E	174	35	34.19	19	4
		RA557		White Cliff Track, Pukearuhe Road off State Highway 3	S	38	53	41.17	E	174	30	56.46	14	4
		RA558		Pukearuhe Road, off State Highway 3	S	38	54	48.37	E	174	29	30.72	23	4
		RA559		South of White Cliffs Brewery, just past Waitoetoe Road, State Highway 3	S	38	58	58.70	E	174	26	10.98	23	6
		RA560		Onaero River Scenic Reserve	S	38	59	45.26	E	174	21	49.80	16	4
		RA562		Sangster Road, Lake Rotokare	S	39	27	17.26	E	174	24	17.24	168	6
		RA563		Lake Rotokare Scenic Reserve, 500m down walking track on left	S	39	27	14.08	E	174	24	44.24	183	4
		RA564		Tangahoe River off Davidson Road	S	39	34	28.31	E	174	20	29.12	39	4
		RA565		Tangahoe River on Ohangi Road	S	39	34	30.35	E	174	20	46.40	33	6
		RA566		SH south of Patea	S	39	44	39.65	E	174	29	29.79	59	6
		RA569		Barley Flat Road, Te Wharau	S	41	9	30.57	E	175	47	27.55	-	6
		RA571		Glenburn Station	S	41	16	18.52	E	175	52	34.45	-	6

The first two columns show the accessions used for long-range PCR prior to Illumina sequencing and which were used to ts the HRM method, respectively.

Table A3.1: Table of sequencing and HRM primers

Primer type	Primer	Primer code	Sequence 5' - 3'	Primer length (bp)	Size bp
Nuclear primers					
ITS	Forward	ITS28CC*	CGCCGTTACTAGGGGAATCCTTGTAAG	27	~760
	Reverse	ITS5$^{\Delta}$	GGAAGTAAAAGTCGTAACAAGG	21	
Waxy	Forward	Waxy 7F‡	GYYTTSTGCATCCACAACATTGC	23	
	Reverse	Waxy 13R‡	GGAGTGGCRACGTTTTCCTT	20	
Universal cp primers					
rpl32-trnL	Forward	trnL(UAG)†	CTGCTTCCTAAGAGCAGCGT	20	1692
	Reverse	rpL32-F†	CAGTTCCAAAAAAACGTACTTC	22	
trnQ-5'-rpS16†	Forward	trnQ(UUG)†	GCGTGGCCAAGYGGTAAGGC	20	1220
	Reverse	rps16x1†	GTTGCTTTYTACCACATCGTTT	22	
3'trnV-ndhA†	Forward	trnV(UAC)x2†	GTCTACGGTTCGARTCCGTA	20	787
	Reverse	ndhC†	TATTATTAGAAATGYCCARAAAATATCATATTC	33	
psbD-trnT$^{GCU-R\dagger}$	Forward	psbD†	CTCCGTARCCAGTCATCCATA	21	2607
	Reverse	trnT(GGU)$^{-R\dagger}$	CCCTTTTAACTCAGTGGTAG	20	
trnfM-trnS$^{\#}$	Forward	trnS$^{UGA\#}$	GAGAGAGAGGGATTCGAACC	20	1619
	Reverse	trnfM$^{CAU\#}$	CATAACCTTGAGGTCACGGG	20	
rbcL$^{\#}$	Forward	aF (rbcLAsF1)$^{+}$	ATGTCACCACAAACAGAGACTAAAGC	26	1324
	Reverse	cR (rbcLAsR1)$^{+}$	GCAGCAGCTAGTTCCGGGCTCCA	23	
trnL-trnF	Forward	F (trnLf, TABF)$^{\Delta}$	ATTTGAACTGGTGACACGAG	20	
	Reverse	C (trnLc, TABc)$^{\Delta}$	CGAAATCGGTAGACGCTACG	20	
Long-range primers					
1	Forward	CorLae psbA – trnS-F	AGCAATACCAACCCTCGTGAGAGAACAA	28	8916
	Reverse	CorLae psbA – trnS-R	CCCTCTCTTTCCGTTTCTGTCGATGACT	28	
2	Forward	CorLae psbK - atpF-F	AGCTTTTGTTTGGCAAGCTGCTGTAAGT	28	6136
	Reverse	CorLae psbK - atpF-R	TTTTTGGAAAGGGAGTGTGTGCGAGTTG	28	
3	Forward	CorLae atpF – rpoC1-F	ACTGATCTGCTTCCATTTCGACTTTCCG	28	11559
	Reverse	CorLae atpF – rpoC1-R	TGCCCAGTAACCCATGTGTGGTATTTGA	28	
4	Forward	CorLae rpoC1 - rpoB-F	CGATCTTTTAGGTCCCCTCTTCACCTCG	28	3563
	Reverse	CorLae rpoC1 - rpoB-R	GAAAAGCAAGGATATGGGCTCGTGTGAG	28	
5	Forward	CorLae rpoB – trnT-F	AGGGCCCAAATAACTCGATTTTCTCCCA	28	6667
	Reverse	CorLae rpoB – trnT-R	CCTTACCATGGCGTTACTCTACCACTGA	28	
6	Forward	CorLae trnE – psaB-F	GAGATGTCCTGAACCACTAGACGATGGG	28	9107
	Reverse	CorLae trnE – psaB-R	CGGGTTGGTTACACCTACAACCGAAATG	28	
7	Forward	Cor Lae psaB - ycf3-F	CCCAAAGTATGGAACCCCAGAAAAAGGC	28	6242
	Reverse	Cor Lae psaB - ycf3-R	ACTTCAGGGGAAAAAGAGGCATTCACCT	28	
8	Forward	CorLae ycf3 - ndhJ-F	ATGAACTGAGTGGGGCTAGTGTTTTTGC	28	6227
	Reverse	CorLae ycf3 - ndhJ-R	TGTTTTCTGGGTTTGGAAAAGTGCGGAT	28	
9	Forward	Cor Lae trnF - atpB-F	GCTCAGTTGGTAGAGCAGAGGACTGAAA	28	6035

Primer type	Primer	Primer code	Sequence 5' - 3'	Primer length (bp)	Size bp
	Reverse	Cor Lae trnF - atpB-R	AGAGGAATGGAAGTGATTGACACGGGAG	28	
10	Forward	CorLae ycf4 - petG -F	ACAAGCAATTTCTGCTGGGCCTTTATCC	28	6192
	Reverse	CorLae ycf4 - petG-R	AGGTCCAACTGATCACCACGTCTGTATT	28	
11	Forward	Cor Lae psbE - clpP-F	GTGCTGACGAATAACCAACCTGCAATGA	28	7716
	Reverse	Cor Lae psbE - clpP-R	GTCATATGGGATTTCCCCGTTCTCTCCC	28	
12	Forward	CorLae clpP - rps8 -F	ACCGTACATGCACCTTTTGATGCATACG	28	9006
	Reverse	CorLae clpP - rps8 -R	CTCGACTAGAAGGAATCGGCGGAGAAAT	28	
13	Forward	CorLae rpl6 - rps9 -F	ACGGAGGCCCTTATTTTCATATTTCGCA	28	4655
	Reverse	CorLae rpl6 - rps9 -R	GGTCCCGAGCATCTACCATTATACCCAC	28	
14	Forward	Cor Lae ndhF - ndhD-F	GCGTTAATTCAGCTAATCCTCTTATACCCCC	31	7269
	Reverse	Cor Lae ndhF - ndhD -R	CGCGGGTTCCTTTCTTTTCTTTTTCCCT	28	
15	Forward	CorLae ndhD - ndhH-F	GTACTGTAATGTGCCTCTCCATGGGTGT	28	6758
	Reverse	CorLae ndhD - ndhH-R	ATCAATGCACGGTGTTCTTCGACTCATC	28	
16	Forward	CorLae ndhH - ycf1 -F	GAAGCTCGCAGCATTGGTCCTGATAAAC	28	6053
	Reverse	CorLae ndhH - ycf1-R	TCCTCCGGATGGCAATCAAGAAAATTCG	28	
Primers for SNP validation					
SNP 1	Forward	CorLaeSNP001F	TTCTGCGTAGTTTATTTGACTTAAGAGG	28	380
	Reverse	CorLaeSNP001R	TACGTCTCATATATTTTCAATGATGCAT	28	
SNP 2	Forward	CorLaeSNP002F	CTCAACTATATCAACTGTACTGGAACTG	28	511
	Reverse	CorLaeSNP002R	TTATTCTTAGTCTTTCTTATCTCCATCC	28	
SNP 3	Forward	CorLaeSNP003F	CCCCAACGTTTATACAAAGGCTTACGTA	28	500
	Reverse	CorLaeSNP003R	GATTCTGCCCTTCTAAAAGGAACATCAG	28	
SNPs 4-7	Forward	CorLaeSNP004-007F	TGAACGACTCGACTCTGCAT	20	865
	Reverse	CorLaeSNP004-007R	AGGCCGTGGAAATAAAAAGG	20	
SNP 8	Forward	CorLaeSNP004-008F	CCATGGATAAAGGTAGAAAGGTGT	24	159
	Reverse	CorLaeSNP004-008R	GGTTTTTGTTTCACCGAGCTA	21	
SNPs 9-14	Forward	CorLaeSNP009-014F	GGAATGAAAAGCGTCCATTG	20	850
	Reverse	CorLaeSNP009-014R	CGAATAAACCCCAGTTCCAA	20	
SNPs 23-27	Forward	CorLaeSNP023-027F	ACGTTCTCCTGTGCTTCCAG	20	352
	Reverse	CorLaeSNP023-027R	TTCTAGCCCGAAGCCAATAA	20	
SNPs 28-36	Forward	CorLaeSNP028-036F	TTACCTCTCCTTTTCCTCATTAAA	24	619
	Reverse	CorLaeSNP028-036R	TGGAAAACAAGACAGGGAT	20	
SNP 38	Forward	CorLaeSNP038F	CCTGTATCAGAATGTAAGACAATGC	25	299
	Reverse	CorLaeSNP038R	GGTACCAAATCGTAGTCAAGTTTT	24	
SNPs 39-42	Forward	CorLaeSNP038-042F	CCGAATTCAATTGGTTTGCT	20	590
	Reverse	$trnL^{UAG}$	CTGCTTCCTAAGAGCAGCGT	20	
SNPs 43-45	Forward	CorLaeSNP043-045F	AAAGGGAATGGGCGTGATA	20	591
	Reverse	CorLaeSNP043-045R	GCCTGGACCGATACATGATT	20	
SNPs 46	Forward	CorLaeSNP046F	ATGTATCGGTCCAGGCAAAT	20	664
	Reverse	CorLaeSNP046R	CGTTGTTGATTGGAAATTGG	20	
SNPs 48-49	Forward	CorLaeSNP048-049F	TGCTACGGGAAATATAGGAAAAA	23	696

Primer type	Primer	Primer code	Sequence 5' - 3'	Primer length (bp)	Size bp
	Reverse	CorLaeSNP048-049R	CCCTCGAGGTCGTAGAGAGT	20	
SNPs 50-51	Forward	CorLaeSNP050-051F	TTGGCCATAGGACCTTGACT	20	783
	Reverse	CorLaeSNP050-051R	TGAATTGGGCAAATTATTCAT	21	
Primers for HRM analysis					
HRMA-SNP1	Forward	CorLaeHRM1F	TCACCTCTGGCTCAATTCTT	20	149
	Reverse	CorLaeHRM1R	CCAATGTCTGTGTCTGTACGAA	22	
HRMA-SNP3	Forward	CorLaeHRM3F	GAAGAGTTGGTGCTGGGCTA	20	111
	Reverse	CorLaeHRM3R	CACCTTTTGCGGGATTATTG	20	
HRMA-SNP16	Forward	CorLaeHRM16F	AACGGGTCTTCCATCTTGC	19	145
	Reverse	CorLaeHRM16R	TCCCGAAATGATTCGTGT	18	
HRMA-SNP38+41	Forward	CorLaeHRM38F	CCGAATTCAATTGGTTTGCT	20	126
	Reverse	CorLaeHRM38R	ACCTAAGCACTACACTCAAAAA	22	
HRMA-SNP42	Forward	CorLaeHRM42F	TCCATGGATTAAAGCCAGAAC	21	84
	Reverse	CorLaeHRM42R	AACAATGGGATTTTTCGTCA	20	
HRMA-SNP49	Forward	CorLaeHRM49F	CGATTTGGTACCTTATTTGCATT	23	110
	Reverse	CorLaeHRM49R	TCCCCTCGAGGTCGTAGA	18	
HRMA-SNP50	Forward	CorLaeHRM50F	TTTCGCCATTTTTGTGGT	18	69
	Reverse	CorLaeHRM50R	AAGAGGTCATTATCAATACGATTT	24	

Symbol key:
* (Wagstaff & Garnock-Jones, 1998)
◊ (White *et al.*, 1990)
‡ (Olmstead, R *et al*. unpubl.)
† (Shaw *et al.*, 2007)
(Hasebe *et al.*, 1994)
+ (Demesure *et al.*, 1995)
Δ (Taberlet *et al.*, 1991)

REFERENCES

Demesure, B., Sodzi, N., and Petit, R. J. 1995. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology* **4:** 129-134.

Hasebe, M., Omori, T., Nakazawa, M., Sano, T., Kato, M., and Iwatsuki, K. 1994. rbcL gene-sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 5730-5734.

Shaw, J., Lickey, E., Schilling, E., and Small, R. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94:** 275.

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17:** 1105-1109.

Wagstaff, S. J., and Garnock-Jones, P. J. 1998. Evolution and biogeography of the Hebe complex (Scrophulariaceae) inferred from ITS sequences. *New Zealand Journal of Botany* **36:** 425-437.

White, T., Bruns, T., Lee, S., and Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M., Gelfand, D., Shinsky, J., and White, T. (Eds), *PCR Protocols: A Guide to Methods and Applications*, pp. 315-322. Academic Press.

Table A4.1: SNP marker development in the chloroplast genome of karaka

SNP	Location	cp region	SNP type		MAF	Comments	Sanger validated	HRM compatible	Used in final suite
1	7418	rps16 - trnQ intergenic region	G		n/a	discovered in whole genome project	Y	Y	Y
2	125487	ndhA gene	C		n/a	discovered in whole genome project	Y	N	Y
3	77944	psbB gene	A		n/a	discovered in whole genome project	Y	Y	Y
4	9033	psbK - psbI intergenic region	T	C	>0.1	after indel and long run of As and Ts	N	N	N
5	9035	psbK - psbI intergenic region	C	A	>0.1	after indel and long run of As and Ts	N	N	N
6	9502	trnS - trnG intergenic region	A	T	>0.5	Conserved with good coverage	Y	N	N
7	9503	trnS - trnG intergenic region	T	A	>0.5	Conserved with good coverage	Y	N	N
8	10509	trnG intron	C	A	>0.5	Conserved with good coverage	Y	Y	Y
9	11321	trnR gene - atpA intergenic region	T	A	>0.5	Conserved with good coverage	Y	N	N
10	11331	trnR gene - atpA intergenic region	T	C	>0.5	Conserved with good coverage	Y	N	N
11	11590	trnR gene - atpA intergenic region	A	T	<0.1	close to indel	N	N	N
12	11595	trnR gene - atpA intergenic region	G	A	<0.1	close to indel	N	N	N
13	11596	trnR gene - atpA intergenic region	A	G	<0.1	close to indel	N	N	N
14	11608	trnR gene - atpA intergenic region	T	A	>0.5	after indel and long run of As	N	N	N

SNP	Location	cp region	SNP type		MAF	Comments	Sanger validated	HRM compatible	Comments
15	25000	rpoC1 intron	T	G	>0.5	Conserved with good coverage	Y	Y	N
16	26452	rpoB gene	T	G	>0.5	Conserved with good coverage	Y	Y	Y
17	45581	psaA - ycf3 intergenic region	T	G	>0.5	Conserved with good coverage	Y	Y	N
18	47802	ycf3 exon	T	C	>0.1	coverage too low	N	N	N
19	48293	ycf3 - trnS intergenic region	A	C	>0.1	coverage too low	N	N	N
20	48790	trnS - rps4 intergenic region	T	C	<0.1	coverage too low	N	N	N
21	50878	trnT - trnL intergenic region	A	T	>0.1	coverage too low and too close to ends of reads	N	N	N
22	63001	accD - psaI intergenic region	T	A	<0.1	coverage too low	N	N	N
23	69010	psbE - petL intergenic region	G	A	>0.5	Conserved with good coverage	N	N	N
24	69013	psbE - petL intergenic region	G	A	>0.5	Conserved with good coverage	N	N	N
25	69048	psbE - petL intergenic region	A	G	>0.5	Conserved with good coverage	N	N	N
26	69070	psbE - petL intergenic region	G	T	>0.5	Conserved with good coverage	N	N	N
27	69075	psbE - petL intergenic region	G	C	>0.5	Conserved with good coverage	N	N	N
28	70443	petL - petG intergenic region	A	G	>0.5	coverage too low	N	N	N
29	70678	petG - trnW intergenic region	T	C	<0.1	coverage too low	N	N	N
30	70737	in trnW gene	G	A	>0.5	coverage too low	N	N	N
31	70902	trnW - trnP intergenic region	A	G	>0.5	coverage too low	N	N	N

SNP	Location	cp region	SNP type		MAF	Comments	Sanger validated	HRM compatible	Comments
32	70905	trnW - trnP intergenic region	T	G	>0.5	coverage too low	N	N	N
33	70920	in trnP gene	G	A	<0.1	coverage too low	N	N	N
34	70963	in trnP gene	C	G	<0.1	coverage too low	N	N	N
35	71004	trnP - psaJ intergenic region	A	G	<0.1	coverage too low	N	N	N
36	71016	trnP - psaJ intergenic region	A	G	<0.1	coverage too low	N	N	N
37	116208	in ndhF gene	C	T	>0.1	Conserved with good coverage	Y	Y	N
38	118699	rpl32-trnL gene	C	G	<0.1	Conserved with good coverage	Y	N	N
39	118701	rpl32-trnL gene	A	G	<0.1	Conserved with good coverage	Y	N	N
40	118702	rpl32-trnL gene	T	A	<0.1	Conserved with good coverage	Y	N	N
41	118754	rpl32-trnL gene	C	A	<0.1	Conserved with good coverage	Y	Y	Y
42	118942	rpl32-trnL gene	C	T	<0.1	Conserved with good coverage	N	N	N
43	123167	ndhE - ndhG intergenic region	G	T	>0.5	Conserved with good coverage	N	N	N
44	123607	ndhG gene	A	G	<0.1	coverage too low	N	N	N
45	123609	ndhG gene	A	G	<0.1	coverage too low	N	N	N
46	124129	indhI gene	C	T	>0.5	Conserved with good coverage	Y	Y	N
47	125079	ndhA exon 2	A	T	>0.5	Conserved with good coverage	Y	Y	N
48	128994	ycf1 gene	T	A	>0.1	Conserved with good coverage	Y	Y	N
49	129260	ycf1 gene	T	A	>0.5	Conserved with good coverage	Y	Y	Y
50	130659	ycf1 gene	A	G	>0.5	Conserved with good coverage	N	N	N
51	131245	ycf1 gene	A	C	>0.5	Conserved with good coverage	N	N	N

Location and region of the chloroplast genome of karaka scanned for SNP variants. The frequency at which the SNP was detected (MAF=minor allelic frequency); comments on the coverage for that base when sequenced using Illumina GAII instrument and the stage at which the marker was considered no longer viable for this study. The final suite of markers contained seven SNP loci.

Table A5.1: Concordance of SNP calling for 60 test accessions using HRM analysis and Sanger sequencing

	High resolution melt analysis						Sanger sequencing						% Concordance for chlorotype assignment
	SNP 1	SNP 3	SNP 8	SNP 16	SNP 41	SNP 49	SNP 1	SNP 3	SNP 8	SNP 16	SNP 41	SNP 49	
96102*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96103*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96104*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96105*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96106*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96138*	n/a	n/a	n/a	n/a	n/a	n/a	A	A	G	T	A	A	n/a
96160*	n/a	n/a	n/a	n/a	n/a	n/a	G	C	G	T	C	T	n/a
RA38	A	A	A	G	C	A	A	A	A	G	A	A	100
RA50	G	A	G	T	C	T	G	A	G	T	C	T	100
RA64	G	C	-	T	C	T	G	C	G	T	C	T	100
RA72	-	C	G	T	C	T	G	C	G	T	C	T	100
RA83	G	A	-	T	A	T	G	A	G	T	C	T	66.67
RA84	A	A	A	T	C	A	A	A	A	G	A	A	66.67
RA86	G	A	G	T	C	T	G	A	G	T	C	T	100
RA99	G	A	-	T	C	A	G	A	G	T	C	T	80
RA103	A	C	A	G	A	A	A	A	A	G	A	A	83.33
RA117	A	C	G	T	A	T	G	C	G	T	C	T	83.33
RA119	A	C	G	T	C	T	G	C	G	T	C	T	83.33
RA123	A	C	-	G	C	A	A	A	A	G	A	A	80
RA124	A	C	A	T	C	A	A	A	A	G	A	A	66.67
RA133	A	A	A	-	C	A	A	A	A	G	A	A	80
RA139	G	A	G	T	C	T	G	A	G	T	C	T	100
RA155	G	A	-	T	C	T	G	A	G	T	A	T	80
RA165	A	C	A	G	A	A	A	A	A	G	A	A	83.33
RA183	G	A	-	T	C	A	G	A	G	T	C	T	80
RA201	G	A	G	-	C	T	G	A	G	T	C	T	100
RA204	A	A	-	T	C	T	G	A	G	T	C	T	66.67
RA211	A	C	A	G	A	A	A	A	A	G	A	A	83.33
RA218	A	C	A	G	A	A	A	A	A	G	A	A	83.33
RA233	G	A	-	T	C	T	G	A	G	T	C	T	100
RA237	-	A	G	T	C	T	G	A	G	T	C	T	100
RA242	-	C	G	T	C	T	G	C	G	T	C	T	100
RA269	A	A	A	G	A	A	A	A	A	G	A	A	100
RA278	G	A	G	T	C	T	G	A	G	T	C	T	100
RA296	G	A	-	T	-	T	G	A	G	T	del	T	100
RA320	G	A	G	T	A	T	G	A	G	T	A	T	100
RA314	G	A	G	-	C	T	G	A	G	T	C	T	100
RA320	A	C	A	G	-	A	A	A	A	G	A	A	80
RA324	G	A	-	T	C	T	G	A	G	T	C	T	100
RA325	G	A	-	T	C	T	G	A	G	T	C	T	100
RA328	G	A	-	T	C	T	G	A	G	T	C	T	100
RA331	G	A	-	T	C	T	G	A	G	T	C	T	100
RA332	G	A	-	T	C	T	G	A	G	T	C	T	100

	High resolution melt analysis						Sanger sequencing						% Concordance for chlorotype assignment
	SNP 1	SNP 3	SNP 8	SNP 16	SNP 41	SNP 49	SNP 1	SNP 3	SNP 8	SNP 16	SNP 41	SNP 49	
RA333	G	A	-	T	C	A	G	A	G	T	C	T	80
RA335	G	A	-	T	C	T	G	A	G	T	C	T	100
RA337	G	A	A	T	C	A	G	A	G	T	C	T	66.67
RA338	A	A	-	T	C	T	G	A	G	T	C	T	80
RA340	A	C	A	G	C	A	A	A	A	G	A	A	66.67
RA344	A	A	A	G	A	A	A	A	A	G	A	A	100
RA407	G	A	-	T	C	T	G	C	G	T	C	T	100
RA412	-	A	-	T	C	T	G	A	G	T	C	T	100
RA420	-	C	-	T	C	T	G	C	G	T	C	T	100
RA421	-	A	A	G	A	A	A	A	A	G	A	A	100
RA431	A	C	G	T	C	-	G	C	G	T	C	T	80
RA435	A	A	G	T	C	T	G	A	G	T	C	T	80
RA436	-	C	A	T	C	T	G	C	G	T	C	T	80
RA437	G	C	A	T	C	T	G	C	G	T	C	T	83.33
RA438	A	A	A	G	C	A	A	A	A	G	A	A	83.33
RA464	A	C	A	G	A	A	A	A	A	G	A	A	83.33
RA475	G	C	-	T	C	T	G	C	G	T	C	T	100
RA511	A	A	A	T	C	T	G	A	G	T	C	T	66.67
RA533	G	A	-	T	C	T	G	A	G	T	C	T	100
1127	A	C	A	G	C	A	A	A	A	G	A	A	66.67
1130	A	C	A	G	C	A	A	A	A	G	A	A	66.67
1140	A	A	A	G	A	A	A	A	A	G	A	A	100
1296	A	C	A	G	C	A	A	A	A	G	A	A	66.67
1345	G	C	A	T	C	T	G	A	G	T	C	T	66.67

Accessions with an asterisk are the *Corynocarpus* species used by Wagstaff (2000), which were not subjected to HRM analysis: 96102 (); 96103 (); 96104 (); 96105 (); 96106 (); 96138 (); 96160 (*C. laevigatus*, Kermadec Islands). For chlorotype assignment, HRM showed 100% concordance with Sanger sequencing for 29 of the 60 accessions (48.33%); 83.33% concordance for 10 of the 60 accessions (16.67%); 80% concordance for 9 of the 60 accessions (15%); and 66.67% concordance with 11 of the 60 accessions (18.33%). Concordance was determined based upon whether HRM results allowed the partitioning of accessions into one of the six chlorotypes even if HRM SNP base-calls were missing or incorrect. Del = deletion

Table A6.1: Full data set for 360 accessions of karaka and 1-2 accession of each of the Pacific *Corynocarpus*

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
96102	*C. rupestris ssp arboreus*	A		A	G	T	A	A
96103	*C. rupestris ssp rupestris*	A		A	G	T	A	A
96104	*C. rupestris ssp rupestris*	A		A	G	T	A	A
96105	*C. dissimilis*	A		A	G	T	A	A
96106	*C. cribbianus*	A		A	G	T	A	A
96138	*C. similis*	A		A	G	T	A	A
96160	*C. laevigatus*	G		C	G	T	C	T
1007	*C. laevigatus*	G	C	A	G	T	C	T
1015	*C. laevigatus*	G	C	A	G	T	C	T
1026	*C. laevigatus*	G		A	G	T	C	T
1028	*C. laevigatus*	A		A	A	G	A	A
1033	*C. laevigatus*	A		A	A	G	A	A
1087	*C. laevigatus*	A		A	A	G	A	A
1107	*C. laevigatus*	G		A	G	T	C	T
1109	*C. laevigatus*	G		A	G	T	C	T
1119	*C. laevigatus*	A		A	A	G	A	A
1125	*C. laevigatus*	A		A	A	G	A	A
1127	*C. laevigatus*	A		A	A	G	A	A
1130	*C. laevigatus*	A		A	A	G	A	A
1136	*C. laevigatus*	A		A	A	G	A	A
1140	*C. laevigatus*	A		A	A	G	A	A
1141	*C. laevigatus*	A		A	A	G	A	A
1162	*C. laevigatus*	G		C	G	T	C	T
1172	*C. laevigatus*	G		A	G	T	C	T
1185	*C. laevigatus*	A		A	A	G	A	A
1296	*C. laevigatus*	A		A	A	G	A	A
1299	*C. laevigatus*	A		A	A	G	A	A
1311	*C. laevigatus*	G		A	G	T	A	T
1314	*C. laevigatus*	G		A	G	T	A	T
1319	*C. laevigatus*	G		A	G	T	A	T
1321	*C. laevigatus*	G	A	A	G	T	C	T
1345	*C. laevigatus*	G	A	A	G	T	C	T
1347	*C. laevigatus*	G		A	G	T	C	T
1444	*C. laevigatus*	A		A	A	G	A	A
1445	*C. laevigatus*	A	A	A	A	G	A	A
1446	*C. laevigatus*	G		A	G	T	C	T
1448	*C. laevigatus*	A		A	A	G	A	A
1477	*C. laevigatus*	G		A	G	T	C	T
1648	*C. laevigatus*	A		A	A	G	A	A
1687	*C. laevigatus*	G	C	A	G	T	C	T
1692	*C. laevigatus*	A	A	A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
1981	*C. laevigatus*	A		A	A	G	A	A
1982	*C. laevigatus*	A		A	A	G	A	A
2285	*C. laevigatus*	G	C	A	G	T	C	T
2289	*C. laevigatus*	A		A	A	G	A	A
2290	*C. laevigatus*	G	C	A	G	T	C	T
4650	*C. laevigatus*	G		A	G	T	C	T
4651	*C. laevigatus*	A		A	A	G	A	A
4787	*C. laevigatus*	G		A	G	T	A	T
4799	*C. laevigatus*	G		C	G	T	C	T
4800	*C. laevigatus*	A		A	A	G	A	A
4882	*C. laevigatus*	A		A	A	G	A	A
4914	*C. laevigatus*	G		A	G	T	C	T
4933	*C. laevigatus*	A		A	A	G	A	A
5002	*C. laevigatus*	G		A	G	T	C	T
5042	*C. laevigatus*	G		A	G	T	C	T
5078	*C. laevigatus*	G		A	G	T	A	T
5085	*C. laevigatus*	G		A	G	T	C	T
5091	*C. laevigatus*	A		A	A	G	A	A
5314	*C. laevigatus*	G		A	G	T	C	T
5707	*C. laevigatus*	G		A	G	T	C	T
6007	*C. laevigatus*	G		A	G	T	C	T
6241	*C. laevigatus*	A		A	A	G	A	A
6277	*C. laevigatus*	A		A	A	G	A	A
RA02	*C. laevigatus*	A		A	A	G	A	A
RA04	*C. laevigatus*	A		A	A	G	A	A
RA10	*C. laevigatus*	A		A	A	G	A	A
RA11	*C. laevigatus*	A		A	A	G	A	A
RA14	*C. laevigatus*	A		A	A	G	A	A
RA16	*C. laevigatus*	A		A	A	G	A	A
RA17	*C. laevigatus*	A		A	A	G	A	A
RA20	*C. laevigatus*	A		A	A	G	A	A
RA21	*C. laevigatus*	A		A	A	G	A	A
RA22	*C. laevigatus*	A		A	A	G	A	A
RA23	*C. laevigatus*	A		A	A	G	A	A
RA25	*C. laevigatus*	A		A	A	G	A	A
RA26	*C. laevigatus*	A		A	A	G	A	A
RA29	*C. laevigatus*	A		A	A	G	A	A
RA31	*C. laevigatus*	A		A	A	G	A	A
RA33	*C. laevigatus*	A		A	A	G	A	A
RA38	*C. laevigatus*	A		A	A	G	A	A
RA41	*C. laevigatus*	G		A	G	T	C	T
RA42	*C. laevigatus*	G		A	G	T	C	T
RA44	*C. laevigatus*	G		A	G	T	C	T
RA49	*C. laevigatus*	G		A	G	T	C	T

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA50	*C. laevigatus*	G	C	A	G	T	C	T
RA58*	*C. laevigatus*	G		A	G	T	C	T
RA64	*C. laevigatus*	G	A	C	G	T	C	T
RA65	*C. laevigatus*	G		A	G	T	C	T
RA72	*C. laevigatus*	G	A	C	G	T	C	T
RA79	*C. laevigatus*	A		A	A	G	A	A
RA82*	*C. laevigatus*	G		A	G	T	C	T
RA83	*C. laevigatus*	G		A	G	T	C	T
RA84	*C. laevigatus*	A	A	A	A	G	A	A
RA86	*C. laevigatus*	G	C	A	G	T	C	T
RA90	*C. laevigatus*	A		A	A	G	A	A
RA98	*C. laevigatus*	G		A	G	T	C	T
RA99	*C. laevigatus*	G		A	G	T	C	T
RA100	*C. laevigatus*	G		A	G	T	C	T
RA101	*C. laevigatus*	G		A	G	T	C	T
RA103	*C. laevigatus*	A		A	A	G	A	A
RA111	*C. laevigatus*	G		A	G	T	C	T
RA113	*C. laevigatus*	G		A	G	T	C	T
RA117	*C. laevigatus*	G	A	C	G	T	C	T
RA119	*C. laevigatus*	G	A	C	G	T	C	T
RA123	*C. laevigatus*	A		A	A	G	A	A
RA124	*C. laevigatus*	A		A	A	G	A	A
RA131	*C. laevigatus*	A		A	A	G	A	A
RA133	*C. laevigatus*	A	A	A	A	G	A	A
RA134	*C. laevigatus*	A		A	A	G	A	A
RA136	*C. laevigatus*	A		A	A	G	A	A
RA137	*C. laevigatus*	A		A	A	G	A	A
RA138	*C. laevigatus*	G		A	G	T	C	T
RA139	*C. laevigatus*	G		A	G	T	C	T
RA141	*C. laevigatus*	G		A	G	T	C	T
RA145	*C. laevigatus*	G		A	G	T	C	T
RA146	*C. laevigatus*	A		A	A	G	A	A
RA147	*C. laevigatus*	A	A	A	A	G	A	A
RA148	*C. laevigatus*	A		A	A	G	A	A
RA149	*C. laevigatus*	G		A	G	T	C	T
RA150	*C. laevigatus*	G		A	G	T	C	T
RA151	*C. laevigatus*	A		A	A	G	A	A
RA152	*C. laevigatus*	A		A	A	G	A	A
RA153	*C. laevigatus*	G	A	A	G	T	A	T
RA154	*C. laevigatus*	G		A	G	T	A	T
RA155	*C. laevigatus*	G		A	G	T	A	T
RA158	*C. laevigatus*	A		A	A	G	A	A
RA160	*C. laevigatus*	A		A	A	G	A	A
RA161	*C. laevigatus*	A		A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA162	*C. laevigatus*	A		A	A	G	A	A
RA163	*C. laevigatus*	G		A	G	T	C	T
RA165	*C. laevigatus*	A		A	A	G	A	A
RA166	*C. laevigatus*	A		A	A	G	A	A
RA167	*C. laevigatus*	A		A	A	G	A	A
RA168	*C. laevigatus*	A		A	A	G	A	A
RA169	*C. laevigatus*	A	A	A	A	G	A	A
RA171	*C. laevigatus*	A		A	A	G	A	A
RA172	*C. laevigatus*	A		A	A	G	A	A
RA173	*C. laevigatus*	A		A	A	G	A	A
RA174	*C. laevigatus*	A		A	A	G	A	A
RA177	*C. laevigatus*	A		A	A	G	A	A
RA179	*C. laevigatus*	G	C	A	G	T	C	T
RA180	*C. laevigatus*	A		A	A	G	A	A
RA181	*C. laevigatus*	G		A	G	T	C	T
RA182	*C. laevigatus*	G		A	G	T	C	T
RA183	*C. laevigatus*	G		A	G	T	C	T
RA184	*C. laevigatus*	G		A	G	T	C	T
RA185	*C. laevigatus*	G		A	G	T	C	T
RA186	*C. laevigatus*	G		A	G	T	C	T
RA187	*C. laevigatus*	G		A	G	T	C	T
RA189	*C. laevigatus*	A		A	A	G	A	A
RA190	*C. laevigatus*	G		A	G	T	C	T
RA195	*C. laevigatus*	G		A	G	T	C	T
RA197	*C. laevigatus*	G	C	A	G	T	C	T
RA201	*C. laevigatus*	G		A	G	T	C	T
RA204	*C. laevigatus*	G	C	A	G	T	C	T
RA209	*C. laevigatus*	A		A	A	G	A	A
RA210	*C. laevigatus*	G		A	G	T	C	T
RA211	*C. laevigatus*	A		A	A	G	A	A
RA212	*C. laevigatus*	G		A	G	T	C	T
RA214	*C. laevigatus*	A		A	A	G	A	A
RA215	*C. laevigatus*	A		A	A	G	A	A
RA216	*C. laevigatus*	G		A	G	T	C	T
RA217	*C. laevigatus*	A		A	A	G	A	A
RA218	*C. laevigatus*	A		A	A	G	A	A
RA219	*C. laevigatus*	A		A	A	G	A	A
RA220	*C. laevigatus*	A		A	A	G	A	A
RA223	*C. laevigatus*	A		A	A	G	A	A
RA224	*C. laevigatus*	A		A	A	G	A	A
RA225	*C. laevigatus*	A		A	A	G	A	A
RA227	*C. laevigatus*	A		A	A	G	A	A
RA228	*C. laevigatus*	A		A	A	G	A	A
RA229	*C. laevigatus*	A		A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA23	*C. laevigatus*	A		A	A	G	A	A
RA231	*C. laevigatus*	A		A	A	G	A	A
RA232	*C. laevigatus*	G		A	G	T	C	T
RA233	*C. laevigatus*	G	C	A	G	T	C	T
RA234	*C. laevigatus*	A		A	A	G	A	A
RA235	*C. laevigatus*	A		A	A	G	A	A
RA236	*C. laevigatus*	A		A	A	G	A	A
RA237	*C. laevigatus*	G		A	G	T	C	T
RA239	*C. laevigatus*	G	C	A	G	T	C	T
RA240	*C. laevigatus*	G		A	G	T	C	T
RA241	*C. laevigatus*	A		A	A	G	A	A
RA242	*C. laevigatus*	G	A	C	G	T	C	T
RA243	*C. laevigatus*	G		A	G	T	C	T
RA244	*C. laevigatus*	A		A	A	G	A	A
RA246	*C. laevigatus*	A		A	A	G	A	A
RA261	*C. laevigatus*	A		A	A	G	A	A
RA263	*C. laevigatus*	A		A	A	G	A	A
RA266	*C. laevigatus*	G		A	G	T	C	T
RA269	*C. laevigatus*	A		A	A	G	A	A
RA272	*C. laevigatus*	A		A	A	G	A	A
RA273	*C. laevigatus*	A		A	A	G	A	A
RA274	*C. laevigatus*	A		A	A	G	A	A
RA275	*C. laevigatus*	A		A	G	G	A	A
RA277	*C. laevigatus*	A		A	A	G	A	A
RA278	*C. laevigatus*	G		A	G	T	C	T
RA280	*C. laevigatus*	A		A	A	G	A	A
RA281	*C. laevigatus*	A		A	A	G	A	A
RA282	*C. laevigatus*	A		A	G	G	A	A
RA283	*C. laevigatus*	A		A	A	G	A	A
RA285	*C. laevigatus*	A		A	A	G	A	A
RA287	*C. laevigatus*	G		A	G	T	C	T
RA296	*C. laevigatus*	G		A	G	T	-	T
RA300	*C. laevigatus*	A		A	A	G	A	A
RA302	*C. laevigatus*	A		A	A	G	A	A
RA304	*C. laevigatus*	A		A	A	G	A	A
RA305	*C. laevigatus*	G		A	G	T	A	T
RA306	*C. laevigatus*	G		A	A	T	C	T
RA307	*C. laevigatus*	G		A	G	T	C	T
RA311	*C. laevigatus*	G		A	G	T	C	T
RA312	*C. laevigatus*	G		A	G	T	C	T
RA314	*C. laevigatus*	G		A	G	T	C	T
RA315	*C. laevigatus*	G		A	G	T	C	T
RA319	*C. laevigatus*	G		A	G	T	C	T
RA320	*C. laevigatus*	A		A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA324	*C. laevigatus*	G		A	G	T	C	T
RA325	*C. laevigatus*	G	C	A	G	T	C	T
RA328	*C. laevigatus*	G	C	A	G	T	C	T
RA331	*C. laevigatus*	G	C	A	G	T	C	T
RA332	*C. laevigatus*	G	C	A	G	T	C	T
RA333	*C. laevigatus*	G	C	A	G	T	C	T
RA335	*C. laevigatus*	G	C	A	G	T	C	T
RA337	*C. laevigatus*	G	C	A	G	T	C	T
RA338	*C. laevigatus*	G	C	A	G	T	C	T
RA340	*C. laevigatus*	A		A	A	G	A	A
RA341	*C. laevigatus*	A		A	A	G	A	A
RA342	*C. laevigatus*	A		A	A	G	A	A
RA344	*C. laevigatus*	A		A	A	G	A	A
RA345	*C. laevigatus*	G		A	G	T	C	T
RA346	*C. laevigatus*	A		A	A	G	A	A
RA347	*C. laevigatus*	G		A	G	T	C	T
RA349	*C. laevigatus*	A		A	A	G	A	A
RA350	*C. laevigatus*	G	C	A	G	T	C	T
RA351	*C. laevigatus*	G	C	A	G	T	C	T
RA356	*C. laevigatus*	G		A	G	T	C	T
RA360	*C. laevigatus*	A		A	A	G	A	A
RA361	*C. laevigatus*	A		A	A	G	A	A
RA366	*C. laevigatus*	G		A	G	T	C	T
RA367	*C. laevigatus*	G		A	G	T	C	T
RA368	*C. laevigatus*	G		A	G	T	C	T
RA371	*C. laevigatus*	A		A	A	G	A	A
RA372	*C. laevigatus*	G		A	G	T	C	T
RA374	*C. laevigatus*	A		A	A	G	A	A
RA375	*C. laevigatus*	A		A	A	G	A	A
RA376	*C. laevigatus*	A		A	A	G	A	A
RA377	*C. laevigatus*	G		A	G	T	C	T
RA378	*C. laevigatus*	G		A	G	T	C	T
RA381	*C. laevigatus*	G	C	A	G	T	C	T
RA384	*C. laevigatus*	G		A	G	T	C	T
RA386	*C. laevigatus*	A		A	A	G	A	A
RA388	*C. laevigatus*	A	A	A	A	G	A	A
RA392	*C. laevigatus*	G		A	G	T	C	T
RA398	*C. laevigatus*	A		A	A	G	A	A
RA400	*C. laevigatus*	A		A	A	G	A	A
RA401	*C. laevigatus*	A		A	A	G	A	A
RA402	*C. laevigatus*	A		A	A	G	A	A
RA403	*C. laevigatus*	G		A	G	T	C	T
RA404	*C. laevigatus*	G		A	G	T	A	T
RA405	*C. laevigatus*	A		A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA406	*C. laevigatus*	G		A	G	T	C	T
RA407	*C. laevigatus*	G	C	C	G	T	C	T
RA408	*C. laevigatus*	G	C	A	G	T	C	T
RA412	*C. laevigatus*	G		A	G	T	C	T
RA414	*C. laevigatus*	G	C	A	G	T	C	T
RA416	*C. laevigatus*	G		A	G	T	C	T
RA417	*C. laevigatus*	A		A	A	G	A	A
RA419	*C. laevigatus*	A		A	A	G	A	A
RA420	*C. laevigatus*	G	A	C	G	T	C	T
RA421	*C. laevigatus*	A		A	A	G	A	A
RA425	*C. laevigatus*	G	A	C	G	T	C	T
RA426	*C. laevigatus*	A		A	A	G	A	A
RA428	*C. laevigatus*	A		A	A	G	A	A
RA429	*C. laevigatus*	A		A	A	G	A	A
RA430	*C. laevigatus*	A		A	A	G	A	A
RA431	*C. laevigatus*	G	A	C	G	T	C	T
RA434	*C. laevigatus*	A		A	A	G	A	A
RA435	*C. laevigatus*	G	A	A	G	T	C	T
RA436	*C. laevigatus*	G	A	C	G	T	C	T
RA437	*C. laevigatus*	G		C	G	T	C	T
RA438	*C. laevigatus*	A		A	A	G	A	A
RA440	*C. laevigatus*	A		A	A	G	A	A
RA442	*C. laevigatus*	A		A	A	G	A	A
RA443	*C. laevigatus*	A		A	A	G	A	A
RA445	*C. laevigatus*	G		A	G	T	C	T
RA448	*C. laevigatus*	A		A	A	G	A	A
RA449	*C. laevigatus*	A		A	A	G	A	A
RA450	*C. laevigatus*	A		A	A	G	A	A
RA454	*C. laevigatus*	A		A	A	G	A	A
RA455	*C. laevigatus*	A		A	A	G	C	A
RA456	*C. laevigatus*	A		A	A	G	A	A
RA457	*C. laevigatus*	A		A	A	G	A	A
RA458	*C. laevigatus*	A		A	A	G	A	A
RA460	*C. laevigatus*	G		A	G	T	C	T
RA461	*C. laevigatus*	A		A	A	G	A	A
RA462	*C. laevigatus*	A		A	A	G	A	A
RA463	*C. laevigatus*	A		A	A	G	A	A
RA464	*C. laevigatus*	A		A	A	G	A	A
RA467	*C. laevigatus*	A		A	A	G	A	A
RA468	*C. laevigatus*	A		A	A	G	A	A
RA469	*C. laevigatus*	A		A	A	G	A	A
RA470	*C. laevigatus*	A		A	A	G	A	A
RA473	*C. laevigatus*	A		A	A	G	A	A
RA474	*C. laevigatus*	G	C	A	G	T	A	T

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA475	*C. laevigatus*	G	C	C	G	T	C	T
RA476	*C. laevigatus*	A		A	A	G	A	A
RA477	*C. laevigatus*	G		A	G	T	C	T
RA478	*C. laevigatus*	A		A	A	G	A	A
RA479	*C. laevigatus*	A		A	A	G	A	A
RA480	*C. laevigatus*	A		A	A	G	A	A
RA481	*C. laevigatus*	G		A	G	T	C	T
RA482	*C. laevigatus*	G		A	G	T	A	T
RA486	*C. laevigatus*	G		A	G	T	C	T
RA501	*C. laevigatus*	G		A	G	T	C	T
RA504	*C. laevigatus*	G		A	G	T	C	T
RA511	*C. laevigatus*	G		A	G	T	C	T
RA512	*C. laevigatus*	A		A	A	G	A	A
RA513	*C. laevigatus*	A		A	A	G	A	A
RA514	*C. laevigatus*	A		A	A	G	A	A
RA516	*C. laevigatus*	G	C	A	G	T	C	T
RA517	*C. laevigatus*	A	A	A	A	G	A	A
RA518	*C. laevigatus*	A	A	A	A	G	A	A
RA519	*C. laevigatus*	G	C	A	G	T	C	T
RA523	*C. laevigatus*	G		A	G	T	C	T
RA525	*C. laevigatus*	A		A	A	G	A	A
RA526	*C. laevigatus*	A		A	A	G	A	A
RA527	*C. laevigatus*	G		A	G	T	C	T
RA528	*C. laevigatus*	G		A	G	T	C	T
RA530	*C. laevigatus*	G		A	G	T	C	T
RA533	*C. laevigatus*	G		A	G	T	C	T
RA534	*C. laevigatus*	G	C	A	G	T	C	T
RA535	*C. laevigatus*	G		A	G	T	C	T
RA538	*C. laevigatus*	A		A	A	G	A	A
RA539	*C. laevigatus*	G		A	G	T	C	T
RA540	*C. laevigatus*	G		A	G	T	C	T
RA542	*C. laevigatus*	A	A	A	A	G	A	A
RA543	*C. laevigatus*	G		A	G	T	C	T
RA544	*C. laevigatus*	G		A	G	T	C	T
RA547	*C. laevigatus*	G		A	G	T	C	T
RA548	*C. laevigatus*	G		A	G	T	C	T
RA549	*C. laevigatus*	G		A	G	T	C	T
RA551	*C. laevigatus*	G		A	G	T	C	T
RA552	*C. laevigatus*	A		A	A	G	A	A
RA553	*C. laevigatus*	G		A	G	T	C	T
RA554	*C. laevigatus*	G		A	G	T	C	T
RA557	*C. laevigatus*	G		A	G	T	C	T
RA558	*C. laevigatus*	G	C	A	G	T	C	T
RA559	*C. laevigatus*	A		A	A	G	A	A

Accession no.	Species	SNP1	SNP2	SNP3	SNP8	SNP16	SNP41	SNP49
RA560	*C. laevigatus*	G		A	G	T	C	T
RA562	*C. laevigatus*	A		A	A	G	A	A
RA563	*C. laevigatus*	G		A	G	T	C	T
RA564	*C. laevigatus*	G		A	G	T	C	T
RA565	*C. laevigatus*	G		A	G	T	C	T
RA566	*C. laevigatus*	A		A	A	G	A	A
RA569	*C. laevigatus*	A		A	A	G	A	A
RA571	*C. laevigatus*	A		A	A	G	A	A

DNA for accessions marked with an asterisk (*) was provided courtesy Steve Wagstaff

GENOME BIOLOGY AND EVOLUTION

SMBE

GBE

Systematic Error in Seed Plant Phylogenomics

Bojian Zhong[1,2,*], Oliver Deusch[1], Vadim V. Goremykin[3], David Penny[1], Patrick J. Biggs[4], Robin A. Atherton[1], Svetlana V. Nikiforova[3], and Peter James Lockhart[1,5]

[1]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand

[2]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

[3]Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy

[4]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

[5]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

*Corresponding author: E-mail: bjzhong@gmail.com.

Accepted: 6 October 2011

Abstract

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here, we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers), we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy, and the fit of conifer chloroplast genome sequences to a general time reversible + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2,250 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favor a close evolutionary relationship between the Gnetales and Pinaceae—the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

Key words: compositional heterogeneity, heterotachy, Gnetales, systematic error.

Introduction

Gnetales are a morphologically and ecologically diverse group of Gymnosperms, united as a monophyletic group based on special features of their cytology. Initially, they were thought to be the nearest relatives of flowering plants (angiosperms) based on the morphological similarities (the "Anthophyte" hypothesis) (Crane 1985). However, all recent molecular work has separated Gnetales away from the angiosperms and instead placed them with or within conifers. Some analyses have placed them as sister group to conifers (the "Gnetifer" hypothesis, Chaw et al. 1997), others close to Pinaceae (the "Gnepine" hypothesis, Bowe et al. 2000; Chaw et al. 2000; Finet et al. 2010; Zhong et al. 2010), and others within conifers but close to Cupressophyta (non-Pinaceae conifers; the "Gnecup" hypothesis, Nickrent et al. 2000; Doyle 2006). These alternative hypotheses are illustrated in figure 1*A*.

It has been reported that Gnetales have a faster substitution rate of sequence evolution than other gymnosperms, which could potentially cause a "long-branch attraction" (LBA) artifact in phylogenetic reconstruction (Zhong et al. 2010). The effects of LBA are well understood, even though the significance of contributing causes is often difficult to determine. These can include faster substitution rates in nonadjacent phylogenetic lineages (Felsenstein 1978), poor taxon sampling due to extinction or limited availability of some taxa (Hendy and Penny 1989), and properties of sequences not well described by stationary time reversible models. The latter include base compositional heterogeneity (Foster 2004; Jermiin et al. 2004) and lineage-specific changes in evolutionary constraint that can alter the proportion of variable sites in homologs (Lockhart and Steel 2005).

To improve taxonomic sampling of the Cupressophyta, we determined sequences for 52 genes from the chloroplast

GENOME BIOLOGY AND EVOLUTION
SMBE

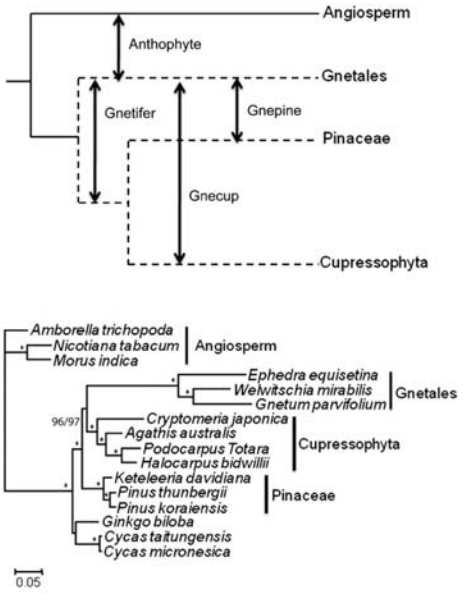

FIG. 1.—(*A*) Four major hypotheses for phylogenetic relationships involving Gnetales. (*B*) Optimal PhyML tree (GTR + G substitution model) reconstructed from all codon positions. The same topology is obtained using 1st + 2nd position sites. Bootstrap support for Gnecup hypothesis is 96% for all sites and 97% for 1st + 2nd position sites.

DNA (cpDNA) genomes of *Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis* using Illumina GAII sequencing. In phylogenetic analyses of concatenated seed plant chloroplast genome sequences, we demonstrate that sites exhibiting greatest character state variation are not well described by a time reversible substitution model. We show that this data property significantly impacts on the reconstruction accuracy of seed plant phylogeny.

Materials and Methods

Sample Collection and DNA Sequences

Tissue for Cupressophyta (*H. kirkii*, *P. totara*, and *A. australis*) was obtained with permission from the living collection at Massey University, Palmerston North. Chloroplasts were isolated and enriched DNA sequenced using the protocols described in Atherton et al. (2010). Short reads were filtered for the longest contigous subsequences below 0.05 error probability using DynamicTrim (Cox et al. 2010). Filtered reads were assembled with Velvet (Zerbino and Birney 2008) and a k-mer range from 23 to 63. Contigs were further assembled using the Geneious assembler (Drummond et al. 2011). Initial annotations for protein-coding genes were carried out using DOGMA (Wyman et al. 2004). Annotations were manually refined by comparison with genes of more closely related species.

We retrieved 13 cp genomes from the NCBI database, including the three genera of Gnetales, one Cupressophyta conifer (*Cryptomeria japonica*), three representatives of Pinaceae conifers (*Pinus thunbergii*, *Pinus koraiensis*, and *Keteleeria davidiana*), and three species from the Cycads/Ginkgo group, with three angiosperms representing the outgroup. GenBank accession numbers for gene sequences used and determined in the present study are listed in supplementary table S1 (Supplementary Material online). Fifty-two protein-coding genes were first aligned as proteins using MUSCLE (Edgar, 2004). Gaps were excluded from these alignments so that only blocks of ungapped residues bounded by similar or identical amino acids were used in phylogenetic analyses. Se-Al v2.0all (Rambaut 2002) was used to edit the underlying DNA sequences into the amino acid alignments. These alignments were then concatenated using Geneious v5.4 (Drummond et al. 2011). This approach produced an alignment of 33289 ungapped positions (not divisible by three as some gaps occur in Genbank sequences).

Sorting Sites Based on Character State Variation

The positions in our concatenated alignments were sorted based on their character state variation. As we demonstrate, this facilitated the study of systematic error in these data. Several methods have been suggested for ordering sites (e.g., discussed in Hansmann and Martin 2000; Goremykin et al. 2010). We used the method of observed variability (OV) sorting as described in Goremykin et al. (2010), which previously has been found to be efficient in concentrating saturated positions toward the most varied end of the sorted alignment. The alignment was ordered from the most highly varied sites to the most conserved sites, and a series of alignments was generated by successively shortening the OV-sorted alignment in steps of 250 sites. For each shortening step, two data partitions were obtained: 1) the shortened alignment containing the most conserved sites (partition "A") and 2) an alignment containing the more varied sites (partition "B"). After model fitting for each partition data, the maximum likelihood (ML) distance and uncorrected *p* distance were calculated using PAUP* (Swofford 2002). Two Pearson correlation analyses of pairwise distances were conducted at each shortening step: 1) correlation of the ML and uncorrected *p* distances for partition B and 2) correlation of the ML distances for partition A and B. The stopping point for site removal was determined as the point at which the two correlations showed a significant improvement (Goremykin et al. 2010).

GENOME BIOLOGY AND EVOLUTION
SMBE

Data Model Fit

We used MISFITS (Nguyen et al. 2011) to determine the occurrence of site patterns in our sorted alignment that were unexpected under a general time reversible (GTR) + G model using three alternative Gnetales phylogenetic trees incorporated as part of the evolutionary model. That is, given a GTR + G substitution model and weighted tree, the expected pattern likelihood vector was computed. For each entry in the vector, a simultaneous $\alpha = 95\%$ Gold confidence region was calculated. Sequence positions in the alignment indicating unexpected patterns were recorded. We also successively shortened our alignment by 250 positions and compared the log-likelihood scores for our OV-sorted alignment (partition A) to log-likelihood scores for identical length partitions jackknife resampled from the complete 33289 position alignment. PhyML 3.0 (Guindon et al. 2010) was used for log-likelihood calculations. Seqboot, implemented in the Phylip3.6 package (Felsenstein 2004), was used for jackknife resampling. *Z*-scores were calculated by subtracting the log-likelihood score on the original data from the mean log-likelihood score for the psuedoreplicate data sets and dividing by the standard deviation (SD) of mean scores.

Compositional Heterogeneity

MEGA5.0 (Tamura et al. 2011) was used to calculate the average nucleotide composition of 1) all codon sites, 1st + 2nd codon sites, and 3rd codon sites, and 2) intervals of increasing length (250 bp) beginning from the most varied end of the OV-sorted alignment. The SD of mean nucleotide frequencies was plotted to visualize compositional heterogeneity among taxa.

Phylogenetic Analyses

ML trees were built assuming a GTR + G model implemented in PhyML 3.0 (Guindon et al. 2010). The relative length of branches and extent of heterotachy (lineage-specific differences in evolutionary rate) in these trees was visualized using SplitsTree 4.0 (Huson and Bryant 2006).

Results

Effect of Improved Taxon Sampling

In ML analyses of all codon positions and 1st + 2nd sites, inclusion of the newly determined sequences from three Cupressophyta genomes halved the length of the internal branch subtending Gnetales and Cupressophyta when compared with phylogenetic reconstructions made without these taxa. Inclusion of sequences from these additional genomes did not change the topology. In the trees with additional taxa, the Gnecup hypothesis (fig. 1*B*) was strongly supported (96% and 97% bootstrap support for all positions and 1st + 2nd sites, respectively). However as we show below, support for this hypothesis was also strongly dependent on the inclusion of sites in the data that showed a poor fit to the GTR + G substitution model.

The Impact of Site Removal

We used the OV sorting criterion of Goremykin et al. (2010) to rank site patterns from most varied to least varied. Blocks of columns in steps of 250 sites were then removed sequentially. This produced a series of shortened alignments. ML trees under a GTR + G model were reconstructed for each partition, and the bootstrap support for alternative hypotheses was measured for each partition. This analysis was made for all sites, 1st + 2nd codon position sites, and 3rd codon position sites. Figure 2*A* (all sites) shows that the Gnecup hypothesis was favored only while the 2000 most varied positions were included in the analysis. After these sites were removed, the Gnepine hypothesis became favored until 3,250 sites were removed. After this point, alternative hypotheses were unresolved. With 1st and 2nd codon position data alone, the Gnepine hypothesis was favored after removal of 750 sites and before removal of 1,250 sites (fig. 2*B*). With 3rd codon position data, the Anthophyte hypothesis was initially weakly supported, but this support decreased as sites were removed (fig. 2*C*).

Data Model Fit

To help understand the impact of site removal, we investigated the fit of site patterns to three alternative evolutionary models (Gnecup, Gnepine, and Gnetifer trees) that assumed an optimal GTR + G substitution model. Using MISFITS (Nguyen et al. 2011), we computed the overrepresented and underrepresented site patterns in the OV-sorted data. For the Gnepine hypothesis, we observed that 46%of the sites not fitting the evolutionary model occurred within the 2250 most varied positions (i.e., in 7% of the total alignment length; 15% of all variable sites). About 3.1% (691/22193) of the 1st + 2nd position sites and 15.2% (1687/11096) of the 3rd position sites do not fit the Gnepine tree. A similar poor fit was also obtained for tree topologies that supported the Gnetifer and Gnecup hypotheses (fig. 3), suggesting that in the most varied positions of the OV-sorted alignment, misspecification was a general property of the GTR + G substitution model and not specific to any one hypothesis of evolutionary relationship.

To further evaluate the impact of the most varied positions on data model fit with our three tree models, we also compared the log-likelihood scores for the sequentially shorted (partition A) data sets, with scores for identical length data sets comprised of jackknife resampled site patterns taken from the original 33289 position alignment. The results from this analysis corroborated those obtained with MISFITS in identifying an extremely poor data model fit for sites at the most varied end of the OV-sorted alignment (supplementary fig. S1, Supplementary Material online).

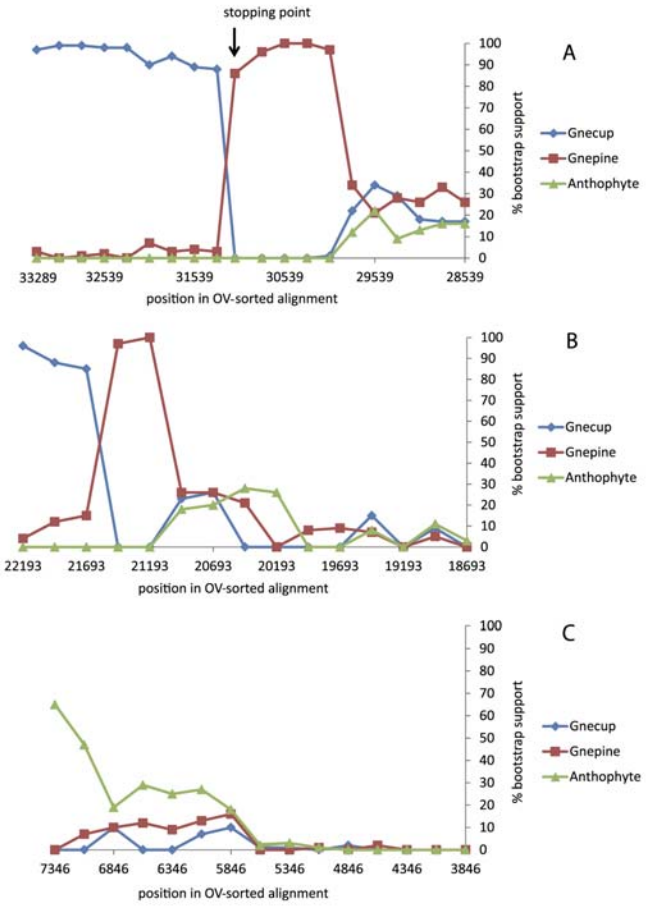

FIG. 2.—Bootstrap support in optimal PhyML trees for three alternative relationships as intervals of 250 bases were successively removed from the OV-sorted alignment. (*A*) all sites, (*B*) 1st + 2nd codon positions, and (*C*) 3rd codon positions.

Compositional Heterogeneity

Figure 4 shows the SD of individual base frequencies from mean (stationary) estimates for intervals increasing in length by 250 bases sampled from the most varied end of the OV-sorted alignment. While the average nucleotide compositional frequencies of all sites, 1st + 2nd sites, and 3rd sites are relatively homogeneous (Results not shown), the most varied OV-sorted sites in the alignment exhibit significant compositional heterogeneity. This decreases incrementally toward the more conserved positions of the OV-sorted alignment.

Heterotachy

Optimal PhyML trees (GTR + G substitution model) were reconstructed for sampling intervals that increased in length by 250 bases from the most varied end of the OV-sorted alignment. The relative length of the Gnetales internal branch separating Gnetales from other species in the 16 taxon data set for each sampling interval is shown in figure 5*A*. The relative length of the branches subtending the Cupressophyta, Pinaceae, and angiosperms in the 13 taxon data set is shown in figure 5*B*. A striking feature of the 16 taxon trees is that the branch leading to the Gnetales

GENOME BIOLOGY AND EVOLUTION
SMBE

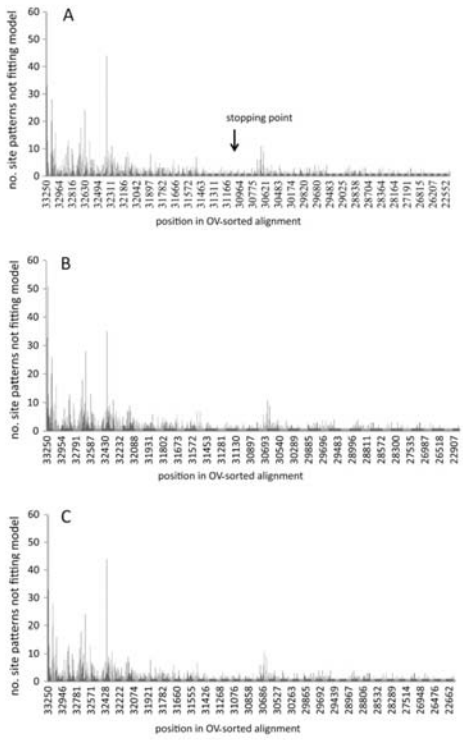

FIG. 3.—Histogram indicating consecutive misfitting site patterns under the (*A*) GTR + G + Gnepine, (*B*) GTR + G + Gnetifer, and (*C*) GTR + G + Gnecup evolutionary model. The height of each histogram indicates the number of unexpected site patterns.

lineage is disproportionately much longer than branches subtending other seed plant lineages (more than 60× longer over the first 1750 bases and between 10×–5× between 2000 and 2500 bases) at the most varied end of the OV-sorted alignment (fig. 5). This extreme branch length difference is a feature of both the 1st + 2nd codon position and 3rd codon position data (not shown).

Removal of Most Varied Sites from the Alignment

We used the stopping criterion of Goremykin et al. (2010) to make an assessment of the number of most varied sites that should be excluded prior to tree building. This criterion considers the alignment partitions created by the sequential shortening steps described previously and compares 1) ML distances for the conserved (A) and the variable (B) bipartition and 2) *p* distances and ML distances for the B partition. The authors have suggested that the removal of

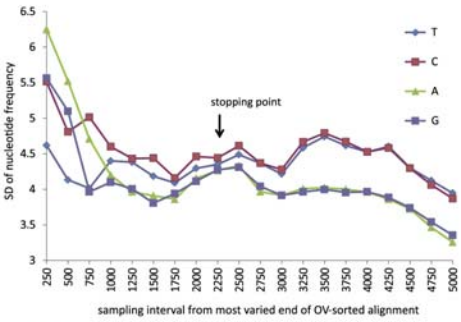

FIG. 4.—Plot indicating nucleotide compositional heterogeneity within intervals sampled from the most varied end of the OV-sorted alignment. Subsequent intervals increased in length by 250 bases per interval.

variable positions should be continued at least until the very end of the sharp rise in Pearson correlation values in either analysis. The stopping criterion identifies the point where the substitution properties of most varied sites (partition B) become more similar to those of the more conserved sites in the alignment (partition A), and where corrected and uncorrected distances for the variable B partition begin to show a strong positive correlation. As such it provides a means to objectively decide a cutoff point for excluding from tree building sites that exhibit site saturation and or model misspecification. Figure 6 indicates change in the correlation coefficient (*r*) and similarity of distances estimates as sites are removed. A sharp rise in (*r*) occurs when 2,000 sites have been removed and it ceases with removal of 2,250 sites in the correlation of *p* distances and ML distances estimated from B partitions. Reference to figure 5 shows that this is accompanied by reduction of heterotachy associated with the Gnetales lineage. It also marks the transition zone for bootstrap support of the Gnecup and Gnepine hypotheses. The Gnepine hypothesis is strongly favored after removal of 2,250 sites (position 31039). It continues to be favored until 3,250 sites are removed when the PhyML trees become unresolved.

Discussion

Most phylogenetic methods assume that DNA sequences have evolved under stationary, reversible, and homogeneous conditions. Violation of this model assumption is well known to lead to inaccurate tree reconstruction (e.g., Lanave et al 1984; Lockhart et al. 1994; Foster 2004; Jermiin et al. 2004; Delsuc et al. 2005; Lockhart and Steel 2005). Our MISFIT analyses indicate a poor fit between the most varied nucleotide sites in the Gnetales chloroplast concatenated data set and a GTR + G model—one of the more

GENOME BIOLOGY AND EVOLUTION
SMBE

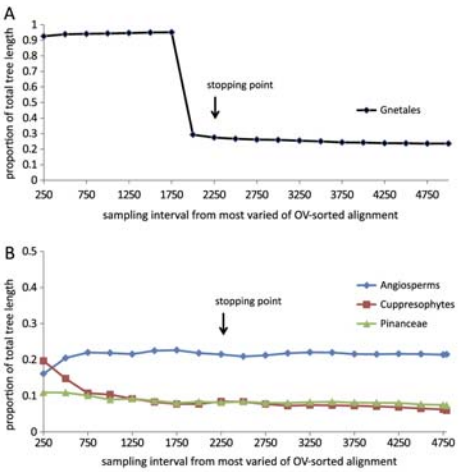

FIG. 5.—Relative length of internal branch leading to (*A*) Gnetales in a 16 taxon data set; (*B*) non-Pinaceae, Pinaceae, and Angiosperms in a 13 taxon data set (this second data set excluded Gnetales). The branch lengths are shown as a proportion of total tree length. Optimal PhyML trees were reconstructed for the same sampling intervals as used in figure 4.

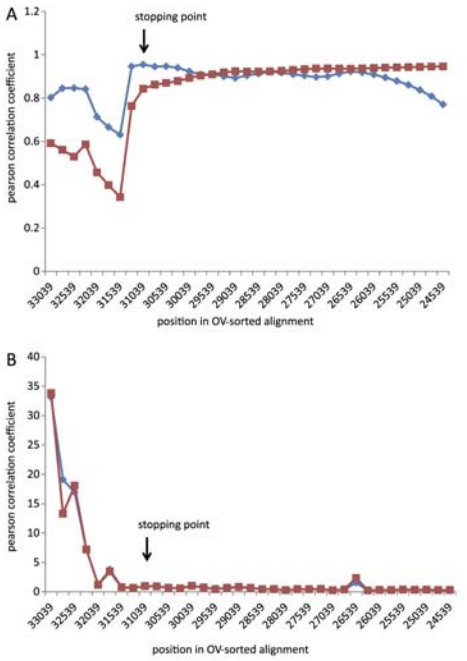

FIG. 6.—(*A*) Pearson correlation analyses. The blue dotted line indicates the Pearson correlation coefficients (*r*) of ML distances for (the more conserved) partition "A" and (less conserved) partition "B". The red dotted line represents *r* value of uncorrected *p* distances and ML distances for partition B. The *r* values begin to increase sharply at the eighth shortening step (31289 position remained). (*B*) Mean deviations of ML distances from *p* distances for B partitions. The red dotted line shows deviations between *p* distances and ML distances calculated using the best-fitting ML model as determined by ModelTest (Posada and Crandall 1998) using the *Akaike information criterion* (the neighbor joining tree was used to estimate ML model parameters). The blue dotted line indicates the deviation between *p* distances and ML distances calculated as above but using an ML tree to fit model parameters.

general models of substitution currently used in phylogenetic reconstruction. Although more complex mixture models exist (e.g., such as the CAT model, Lartillot and Philippe 2004), like GTR + G, they also assume a stationary distribution of base frequencies and have the expectation for a constant proportion of variable sites in all sequences.

Deviation from compositional homogeneity occurs in the most varied regions of the OV-sorted alignment. However, this heterogeneity extends past the OV sorting stopping point and shows no obvious relationship to it. Thus, compositional homogeneity appears an insufficient explanation for the significant increase in value of the Pearson statistic after removal of 2,000 sites and an insufficient explanation for the extent of poor model fit observed in the most varied part of the OV-sorted alignment.

More important for explaining the sharp rise in the Pearson statistic is the extent of substitution rate difference inferred for the Gnetales lineage across the sampling intervals at the most varied end of the OV-sorted alignment. This property of the aligned data causes high variance in ML distance estimation between Gnetales and other species when estimates are made from B partitions. This property of the sorted data explains much of the Pearson coefficient behavior in the correlation analyses. By the final shortening step, at 2250 bases, the relative length of the internal branch separating Gnetales shows approximately 60× reduction in length. This reduction is accompanied by a rapid change in the bootstrap support for the Gnepine hypothesis.

The extreme branch length differences between Gnetales and other lineages for sites at the most varied end of the OV-sorted alignment suggests an issue with alignment of some amino acid positions, despite a conservative approach being used in generating the sequence alignments in the present study. To investigate this further, we also aligned seed plant DNA sequences using the approach of Goremykin et al. (2010) and excluded regions of low sequence similarity (analyses not shown). Working with these alignments, we

GENOME BIOLOGY AND EVOLUTION
SMBE

also obtained very similar results and conclusions regarding heterotachy, compositional heterogeneity, misfit analyses, and bootstrap support. Thus, we conclude that heterotachy is a strong feature of the data and is not a feature of a specific alignment method.

Very recently, a similar study has been undertaken to that reported here. Wu et al. (2011) have determined chloroplast genome sequences for five Cupressophytes and a cycad. They also studied the phylogenetic placement of Gnetales with respect to other seed plants. Our general conclusions are similar to theirs—phylogenetic reconstruction of Gnetales in seed plant phylogeny is misled by non-time reversible properties of aligned chloroplast sequences. From their sampling of taxa, Wu et al. (2011) obtain stronger evidence than we do for lineage-specific change in the Cupressophyta that parallels Gnetales. Our studies also differ in that these authors did not evaluate the relative contribution of compositional heterogeneity and heterotachy in causing problems for tree building. Our analyses suggest that heterotachy is a more significant cause of systematic error in the seed plant sequences analyzed. As we have discussed below, our analyses also suggest that removal of sites rather than individual genes provides a better strategy for dealing with this problem.

Wu et al. (2011) divided chloroplast sequences into L (low heterotachy) and H (high heterotachy) genes and provide evidence that only phylogenetic inference from genes in the L data set is reliable. The H data set contains genes involved in translation including the rpo genes, which previously have been shown to exhibit nonconservative substitutions, indels, and increased proportions of variable sites in green algae (Lockhart et al. 2006). Our analyses indicate that while heterotachy is most pronounced in genes of the H data set, a significant level of heterotachy also occurs in the L data set for conifers that we have studied (not shown). There is also a significant amount of useful phylogenetic information in the H genes, as indicated from our results that favor the Gnepine hypothesis. This conclusion is based on an analysis of 31,039 sites, whereas that of Wu et al. (2011) is based on 21945 DNA positions (7,315 amino acids in the L data set). In general, we suspect that it will be more phylogenetically informative to remove model violating sites rather than genes prior to phylogenetic analyses.

Wu et al. (2011) suggest that the example of Gnetales follows the classic LBA scenario of Felsenstein (1978), wherein there is LBA between Gnetales and Cupressophyta. However, it is important to note that while similar, the LBA scenario for seed plants is likely to differ from this. The properties of seed plant sequences better fit the LBA scenario described by Lockhart and Steel (2005) in which proportions of variable sites change in a lineage-specific fashion, and where parallel changes occur (Zhong et al. 2010) because of similar proportions and convergent patterns of variable sites (modeled in Gruenheit et al. 2008). The significance of the difference in scenarios is important because current methods of tree building do not model lineage-specific change the proportion of variable sites in homologues (Lockhart and Steel 2005; Lockhart et al. 2006; Gruenheit et al. 2008; Shavit Grievink et al. 2008). Although it is possible to model changes in proportions of variable sites using branch length mixtures, these can be complex under some scenarios and thus problematic to identify (Matsen and Steel 2007; Gruenheit et al. 2008; Lartillot et al. 2009). Furthermore, Wu et al. (2011) observe that a mixture branch lengths model was unsuccessful in alleviating LBA with the H data set.

Conclusions

Observations of a poor fit between fast-evolving sites and time reversible models such as the GTR + G model of sequence evolution are not novel (e.g., Sullivan et al. 1995; Goremykin et al. 2004). However, the significance of having a poor fit becomes much more obvious in analysis of concatenated sequences. In the present study, systematic error arising from lineage-specific differences in evolutionary constraint dominates phylogenetic signal and misleads phylogenetic reconstruction. When systematic error contributing to most of the model misfit is removed prior to tree building, our analyses favor the Gnepine hypothesis for seed plant phylogeny (Bowe et al. 2000; Chaw et al. 2000; Finet et al. 2010; Zhong et al. 2010; Soltis et al. 2011; Wu et al. 2011).

We studied site removal in the context of substitution model misspecification and the stopping criterion of Goremykin et al. (2010). With respect to this, our study provides more insight into the performance of this method. Our results indicate that use of the stopping criterion also removes sites that provide a poor fit to tree-building assumptions. Although this criterion does not remove all model violating sites from data, it removes sites that significantly impact on phylogenetic estimates and thus sites most important for misleading tree building. Thus, it provides a useful tool to guide phylogenomic analyses.

Wu et al. (2011) note that improved taxon sampling was insufficient to overcome LBA between Curessophytes and Gnetales. We also obtained this result. However, we wish to be more positive about the contribution that improving taxon sampling of conifers will make to phylogenetic reconstruction of seed plant phylogeny. In our study, addition of sequences from three Cupressophytes reduced the length of the internal branch leading to Gnetales and Cupressophytes 2-fold, even if it was not sufficient to change the topology. Together with international efforts currently underway to sequence and analyze conifer genomes, we believe that analytical approaches such as those used here will be essential for evaluating and mitigating the impact of systematic error in large-scale phylogenomic data sets for seed plants.

GENOME BIOLOGY AND EVOLUTION
SMBE

Supplementary Material

Supplementary table S1, figure S1, and data matrix concatenated gapped alignment are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals. org/).

Acknowledgments

This work was financially supported by the Allan Wilson Centre, Massey University, the New Zealand Royal Society, and the Deutscher Akademischer Austausch Dienst. We thank Jennifer Tate, the anonymous reviewers, and the associate editor for helpful comments.

Literature Cited

Atherton RA, et al. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods. 6:22.

Bowe LM, Coat G, dePamphilis CW. 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci U S A. 97:4092–4097.

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci U S A. 97:4086–4091.

Chaw SM, Zharkikh A, Sung HM, Lau TC, Li WH. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol Biol Evol. 14:56–68.

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 11:485.

Crane PR. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann Mo Bot Gard. 72:716–793.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Doyle JA. 2006. Seed ferns and the origin of angiosperms. J Torrey Bot Soc. 133:169–209.

Drummond AJ, et al. 2011. Geneious v5.4. Auckland (New Zealand): Biomatters, Ltd. [cited 2011 Aug 3]. Available from: http://www.geneious.com/.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Felsenstein J. 2004. PHYLIP (phylogeny inference package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Finet C, Timme RE, Delwiche CF, Marletaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol. 20:2217–2222.

Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.

Goremykin VV, Hirsch-Ernst KI, Woelfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21:1445–1454.

Goremykin VV, Nikoforova SV, Bininda-Emonds OPP. 2010. Automated removal of noisy data in phylogenomic analyses. J Mol Evol. 71:319–331.

Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol. 25:1512–1520.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hansmann S, Martin WT. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol. 50:1655–1663.

Hendy M, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool. 38:297–309.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol. 53:638–643.

Lanave C, Preparata G, Sacone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20:86–93.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for calculating evolutionary substitution rates. Mol Biol Evol. 21:1095–1109.

Lockhart PJ, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.

Lockhart PJ, Steel MA. 2005. A tale of two processes. Syst Biol. 54:948–951.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.

Matsen FA, Steel MA. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst Biol. 56:767–775.

Nguyen MAT, Klaere S, von Haeseler A. 2011. MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. Mol Biol Evol. 28:143–152.

Nickrent DL, Parkinson CL, Palmer JD, DuV RJ. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol Biol Evol. 17:1885–1895.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut A. 2002. Se-Al. Sequence alignment editor v2.0a11. Edinburgh (UK): Andrew Rambaut. [cited 2011 Aug 15] Available from: http://tree.bio.ed.ac.uk/software/seal/.

Shavit Grievink L, Penny D, Hendy MD, Holland BR. 2008. Lineage SpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol. 8(1):317.

Soltis DE, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 98:704–730.

Sullivan J, Holsinger KE, Simon C. 1995. Among-site variation and phylogenetic analysis of 12s rRNA in sigmodontine rodents. Mol Biol Evol. 12:988–1001.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

GENOME BIOLOGY AND EVOLUTION
SMBE

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. Forthcoming 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol.

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. Advance Access published September 19, 2011, doi:10.1093/gbe/evr095.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhong BJ, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol. 27:2855–2863.

Associate editor: Martin Embley

[This page intentionally left blank]

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Robin Amber Atherton

Name/Title of Principal Supervisor: Professor Peter J. Lockhart

Name of Published Research Output and full reference:

"Systematic Error in Seed Plant Phylogenomics"

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V. and Lockhart, P.J. (2011) "Systematic Error in Seed Plant Phylogenomics" Genome Biology Evolution 2011; 3: 1340–1348.

In which Chapter is the Published Work: Appendix

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The candidate developed a chloroplast isolation protocol, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments as well as contributing to and editing the manuscript

Candidate's Signature — 25 July 2013 (Date)

Principal Supervisor's signature — 26 July 2013 (Date)

GRS Version 3– 16 September 2011

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Jul 14, 2013

This is a License Agreement between Robin A Atherton ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3187810931205
License date	Jul 14, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Genome Biology and Evolution
Licensed content title	Systematic Error in Seed Plant Phylogenomics:
Licensed content author	Bojian Zhong, Oliver Deusch, Vadim V. Goremykin, David Penny, Patrick J. Biggs, Robin A. Atherton, Svetlana V. Nikiforova, Peter James Lockhart
Licensed content date	01/01/2011
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Ngā Uru Tāpu - a cultural and genetic study of the karaka/kōpi tree in Aotearoa New Zealand
Publisher of your work	n/a
Expected publication date	Jul 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Robin Amber Atherton

Name/Title of Principal Supervisor: Professor Peter J. Lockhart

Name of Published Research Output and full reference:

"The evolutionary root of flowering plants"

Goremykin, V.V., Nikiforova, S.V., Biggs, P.J., Zhong, B., Delange, P., Martin, W., Woetzel, S., Atherton, R.A., Mclenachan, P.A., Lockhart, P.J. (2013) "The evolutionary root of flowering plants" Systematic Biology 62:1 Pp. 50-61

In which Chapter is the Published Work: Appendix

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The candidate developed a chloroplast isolation protocol, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments as well as contributing to and editing the manuscript

Candidate's Signature — 25 July 2013 (Date)

Principal Supervisor's signature — 26 July 2013 (Date)

GRS Version 3– 16 September 2011