The Chloroplast Genome of Arthropodium bifurcatum.

A thesis presented in partial fulfillment of the requirements
for the degree of
Master of Science in Biological Sciences
Massey University
Palmerston North, New Zealand

Simon James Lethbridge Cox
2010

Abstract

This thesis describes the application of high throughput (Illumina) short read sequencing and analyses to obtain the chloroplast genome sequence of *Arthropodium bifurcatum* and chloroplast genome markers for future testing of hypotheses that explain geographic distributions of Rengarenga – the name Maori give to species of *Arthropodium* in New Zealand.

It has been proposed that *A.cirratum* was translocated from regions in the north of New Zealand to zones further south due to its value as a food crop. In order to develop markers to test this hypothesis, the chloroplast genome of the closely related *A.bifurcatum* was sequenced and annotated. A range of tools were used to handle the large quantities of data produced by the Illumina GAIIx. Programs included the de novo assembler Velvet, alignment tools BWA and Bowtie, the viewer Tablet and the quality control program SolexaQA.

The *A.bifurcatum* genome was then used as a reference to align long range PCR products amplified from multiple accessions of *A.cirratum* and *A.bifurcatum* sampled from a range of geographic locations. From this alignment variable SNP markers were identified.

Putative gene annotations for *A.bifurcatum* were compared to published chloroplast genomes from the Magnoliids and Monocot clades. Interesting similarities and differences have been detected and these have been discussed.

Acknowledgements

I would like to extend many thanks to many people-my supervisors- Lara Shepherd and Pete Lockhart, my co-chloroplast extractor Robin Atherton. Thanks to Trish McLenachan and Olga Kardailsky for their help in the lab and for matters chemical. To the many people on the fifth floor of Science Tower D for answering my questions- Patrick Biggs, Lesley Collins, Bennet McComish, Jing Wang, Murray Cox, Ibrar Ahmed, Nicole Grunheit & Oliver Deusch. Also thanks to Lorraine Cook and Richard Fong from the genome service for help in understanding the Illumina process.

Thanks also goes to the people who took the time to gather Rengarenga, as well as the aforementioned Lara and Robin, a big thank you goes to P. de Lange, P.A. Aspin, J. Collins, J. Hobbs, D. Blake, J. Rolfe and R. Stone.

Thank you to Rob Hallam for the hours of formatting he saved me.

Thank you to Rititia for her love and support and to my parents Michael and Jenny.

Contents

Tables

(Excluding Appendix)

Figures

(Excluding Appendix)

Introduction

Project Outline

The aim of this project was to determine the sequence for the complete chloroplast genome of *Arthropodium bifurcatum*. This genome is approximately 146,000 base pairs (bp) long and contains genes that encode 80 proteins, 42 transfer RNAs (tRNA) and 10 ribosomal RNAs (rRNA). This genome was then used to align PCR products from different accessions of *Arthropodium* in order to discover Single Nucleotide Polymorphisms (SNP's) that would differentiate accessions and to provide data for investigating the translocation patterns of *A.cirratum*.

There are two distinct facets to this study. The first facet deals with the purported translocation of *A.cirratum* and the support from Maori history for these translocations The second facet deals with the molecular science- the extraction of total and chloroplast DNA, the sequencing of the chloroplast DNA, the subsequent primer design, Polymerase Chain reaction (PCR), and further sequencing experiments to identify the SNP's.

This is one part of a larger project that investigates the cultural and genetic aspects of endemic plants that Maori are thought to have translocated, from their hypothesized endemic range in the northern North Island, throughout New Zealand.

Background

Translocation

The Polynesian people doubtlessly translocated plants around the Pacific Ocean on their numerous voyages of exploration and settlement (P. A. Cox & Banack, 1991). These plants would have been of tropical origin and, following their arrival in New Zealand, probably few survived the colder climate (Leach & Stowe, 2005). Rengarenga, the Maori name for *A.cirratum* and *A.bifurcatum* species, appears to be one of the likely endemic replacements for the tropical crop species- taro, cassava, coconut (Fale, 1988)- that did not survive; *Arthropodium* roots being an edible source of starch.

There is a gap in the distribution of Rengarenga in New Zealand. Its endemic range is thought to be north of 38^0 South; it is largely absent down to about 40^0 south where more populations occur. It has been postulated that all populations south of this 38^0 line have been translocated by Maori prior to European arrival (Heenan, Mitchell, & De Lange, 2004).

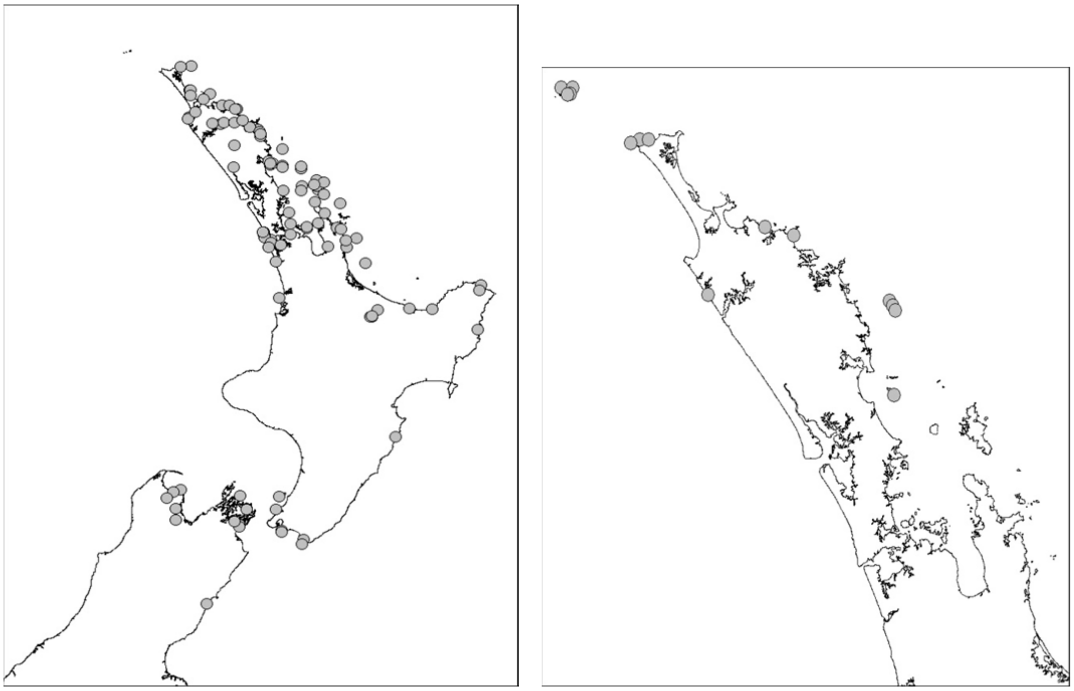


Figure 1: Map of New Zealand showing the distribution of *A.cirratum* (left) and *A.bifurcatum* (right) based on herbarium records and personal observations (L.Shepherd).

Molecular marker development

Molecular data has been used to elucidate the history of domestication and identify the probable source of translocations of both plants and animals in studies that infer relationships between populations (Hamilton, Zug, & Austin, 2010; J. Wang, Pan, Gong, Chiang, & Kuroda, 2011; Zhan, Wang, Gong, & Peng, 2011). These studies have relied upon DNA fingerprinting, microsatellite DNA or short range sequencing of mitochondrial or chloroplast loci and have therefore examined only a small portion of the total amount of sequence variation available for studying recent evolutionary differences.

In studies involving plants, chloroplast DNA (cpDNA) is typically chosen because of its conservative gene organisation, its uniparental (maternal) mode of inheritance and the presence of hotspot regions which can provide high resolution markers (e.g. Lockhart et al., 2001).

In the early 1990's, the rationale was to sequence the spacers and introns that lie between highly conserved coding regions of the chloroplast genome; the idea being that these regions are less functionally constrained and so are more likely to accumulate mutational variations useful for elucidating inter and intra population differences. However a few years later intron and intergenic spacer variability was shown to be highly constrained in many cases due to functional limitations imposed by its secondary structure (Borsch & Quandt, 2009).

Before Next Generation Sequencing became the mainstream approach to sequencing DNA, two main methods were used to extract useful information about population differences in plants – short range sequencing and fingerprinting. Short range sequencing uses generic or custom designed primers, usually from a closely related genus or species whose chloroplast genome has been published, to amplify the region of interest. Fingerprinting amplifies a region of the genome using primers; polymorphisms are visually scored by the absence/presence of a peak or band. The restriction fragments are generated without any knowledge of the DNA sequence. Both methods are limited by the length of DNA sequence that can be amplified by PCR (Polymerase Chain Reaction).

Like any approach, these methods have their limitations. Short range sequencing focuses on the fastest evolving regions of the chloroplast genome, which may be non-coding microsatellites, spacer or intron regions, but there may still be insufficient variation to discriminate between specimens. Further, many of the published generic primers may not provide the suitable kind of resolution required for the taxonomic level being studied. A slowly evolving region may not generate enough difference between accessions while a fast evolving region may mask the differences due to homoplasy. Many studies use both chloroplast and nuclear genome makers to add further depth but at times the signals from the nucleus and the chloroplast disagree (Clarke, Burtenshaw, McLenachan, Erickson, & Penny, 2006; Kårehed, Groeninckx, Dessein, Motley, & Bremer, 2008).

In the search for a fungal DNA barcode both ITS (nuclear ribosomal Internal transcribed spacer) and COX1 (mitochondrial cytochrome c oxidase subunit 1) regions were both deemed unsuitable due to insufficient variation amongst the accessions tested (Seifert, 2009). The same problem arose in choosing a genetic barcode for plants, the authors suggest using more markers but then highlight the perils of paraphyly in the choice of a universal barcode (Fazekas et al., 2009). Use of the COX1 gene, extracted from goldfish, to use as a marker to identify differing strains of this invasive fish species met a similar fate, insufficient variation (Knox, Hicks, Banks, & Hogg, 2008).

With fingerprinting the signal to noise ratio may be low and stochastic error associated with the data quality can mask true signals leading to a loss of resolving power. Fingerprinting errors may occur from

- technical facets of profile generation like PCR-stutter and non-specific amplification
- human error or subjectivity in reading or scoring a profile
- disparities in peak iterations and intensities in the fingerprint profiles (Meudt & Clarke, 2007).

Limitations of the fingerprint process mean that the fingerprinting process lends itself to studies where there are clear differences between accessions, such as in inter rather than intra population enquiries (McLenachan et al., 2000).

The main reason for focusing on characterising small regions of the genome has been the prohibitive price of the infrastructure, and the cost per bp, needed for comparative genome sequencing (Rokas & Abbot, 2009). However, costs have been falling and the deeper the coverage when sequencing a genome, and the more individual genomes that are able to be sequenced, the better the quality of the analysis that can be undertaken.

Taxonomy

The taxonomic place of *Arthropodium* is unresolved. At the familial level it is currently placed in either Laxmanniacae or Asparagaceae. The problem revolves around the discrimination of difficult-to recognize families that feature in the Liliopsida. This was dealt with by introducing different family names for plants with similar features. It appears as though plants that are hard to discriminate within the Asparagaceae/Laxmanniaceae boundary are noted as *sine loco* (without place). Currently the Angiosperm Phylogeny Group (APG) recognises only Asparagaceae (The Angiosperm Phylogeny, 2009).

Arthropodium

Arthropodium is a genus of nine species. Four species occur in Australia, three in New Zealand and one each in New Caledonia and Madagascar. The four Australian species are as follows- *A.milleflorum* grows in the south east corner of Australia with some specimens growing up to 2m tall. Records state that Aboriginals who went into the mountain regions of Victoria and New South Wales to feast on Bogong moths also harvested this species of *Arthropodium*, whose roots were large and plentiful (Gott, 2008). *Arthropodium minus* has a similar range to *A.milleflorum* and is differentiated by its smaller size and single flower per node (Brittan, 1987). *Arthropodium dyeri* grows in Western Australia in eucalypt woodlands and under *Acacia aneuria* (a native scrub) in the drier regions. *Arthropodium curvipes* grows in the south of Western Australia and prefers growing in open, sandy, gravel washes or under shrubs. *Arthropodium noucaledonicum* is endemic to New Caledonia and is described as an alluvium scrub (London., 1920). *Arthropodium caesioides* is found in dense mountain forest or grasslands in Madagascar (Goodman & History., 1996).

Arthropodium in New Zealand

The following three species occur within New Zealand:

Arthropodium cirratum (G.Forst.) R.Br., *Bot.Mag. 49*, t. 2350 (1822)

ʃ *Anthericum cirratum* G.Forst, *Florulae Insularum Australium Prodromus*, 24 (1786).

Arthropodium bifurcatum Heenan, A.D.Mitch. et de Lange, *New Zealand Journal of Botany, 42, (2004)*

Arthropodium candidum (Raoul) Ann.Sci.Nat., Bot. sér. 3, 2:117. (1844)

Arthropodium cirratum was first described by Georg Forster in 1773. This in itself is quite curious as Georg Forster was appointed as an artist to his naturalist father, Johann R. Forster whose job it was to record new plants. Both sailed on James Cook's second major expedition whose purpose was to discover the existence of the southern continent of Antarctica. So an artist first describes *Arthropodium cirratum* on a trip to discover Antarctica. Described as *Anthericum cirratum* in G.Forster's specimen card, the sample was collected by J. Forster in Ships Cove, Queen Charlotte Sound which their ship visited 3 times; the only other stopping point in New Zealand on this expedition was in Dusky Sound, Fiordland (Aughton, 2004). No *A.cirratum* or *A.bifurcatum* have been found growing south of Kaikoura.

Arthropodium bifurcatum was described as a new species in 2004 after collectors noticed a bigger, more fully-fleshed variety, typical of two offshore island groups namely Poor Knights Islands and Three Kings Islands, than plants from some areas of mainland New Zealand. The separation of this species was also supported by AFLP DNA fingerprinting and a range of morphological characters. (Heenan, et al., 2004).

A.candidum was first described by Etienne Raoul on his visit to Akaroa in 1842/3 where he found a specimen on the edge of some “white woods”. Back in France he corresponded with both Hooker and Colenso to confirm the new species from his collection (Raoul, 1846).

Biogeography

There are a number of possibilities for the origin of the *Arthropodium* genus in New Zealand. It may derive from vicariance, with the breakup of the southern continents that were known as Gondwanaland. That there is *Arthropodium* species on Madagascar, that split from Gondwanaland roughly 160 mya, makes tectonic vicariance a less likely hypothesis as flowering plants were evolving at about this time. This tectonic vicariance is, however, a plausible theory for many plant genera, such as *Agathis* (Knapp, Mudaliar, Havell, Wagstaff, & Lockhart, 2007). Dispersal of the seed across the ocean is another possibility for both New Zealand and Madagascan populations (Queiroz, 2005)

Arthropodium in New Zealand may have moved south along a theorized land bridge from New Caledonia that may have been severed 60 million years ago(mya) (Ladiges & Cantrill, 2007). After this separation, an area known as the Norfolk Ridge was an active volcanic zone and is thought to have risen and fallen through a 30 million year period after this. At times the Norfolk ridge is thought to have been totally submerged, while at others the peaks of the Ridge appeared as a long chain of islands possibly to a fully raised bridge (Herzer et al., 1997). At present there is no consensus on the impact that these events had on New Zealand’s biota as discussed in a recent article on the phylogeny of New Zealand earthworms (Buckley et al., 2011). If *Arthropodium* did arrive here via this land bridge then it has been isolated here since the end of that period, 30 million years ago.

It is possible that both of these processes, vicariance and dispersal, may explain the transoceanic dispersal of *Arthropodium* as suggested for *Nothofagus* (Knapp, et al., 2007). Australia is likely to be the source of *Arthropodium*; tectonic vicariance may partly account for the current distribution of *Arthropodium* as may dispersal on the ocean or air currents. It has been suggested that the 8200km distance from South America to New Zealand, against prevailing weather patterns, was enough to establish a new population of *Tetrachondra.* A similar distance, 5800km, would need to be covered to account for *Arthropodium* in Madagascar, if Australia was indeed the centre of origin (Winkworth, Wagstaff, Glenny, & Lockhart, 2002).

There is evidence of *Arthropodium* pollen in the Kowai Formation in the Lake Pukaki area of the Mackenzie Basin. The author tentatively assigns the late Pliocene / early Pleistocene, 1.5-2 mya, as the period when this pollen was deposited. He infers that the climate conditions were so harsh that no trees were able to survive concluding that the environment was sub-alpine (Mildenhall, 2001). As the only *Arthropodium* species suitable for these conditions would be *A.candidum*, the pollen may be from this source. *A.candidum* has been observed growing in the wild around Queenstown, Lake Wakitipu (L.Shepherd, *pers comm*.).

Ecology

Arthropodium cirratum is chiefly a species found on or near the coast. It occurs in fertile areas on and around the outskirts of cliffs, rocky outcrops and slips on limestone, greywacke, basalt and schist substrates. It usually grows in open areas and is known to proliferate under stands of ngaio (*Myoporum laetum*) and pohutukawa (*Metrosideros excelsa*). Two inland populations grow at 300 m.s.l. (Raukumara Range) and 500 m.s.l. (Rotoehu and Otakaina).

Arthropodium bifurcatum is also mainly a coastal species. It primarily favors islands and rocky headlands that are exposed. It reaches its greatest abundance just above the spray zone, growing on exposed rock stacks, platforms that have been cut by the waves and steep cliff faces; in these favored places it appears dense and is almost mono-specific. Associates are sparse and are known to include New Zealand ice plant (*Disphyma australe* subsp. *austral)*, toetoe (*Cortaderia splendens)*, New Zealand climbing spinach (*Tetragonia implexicoma)*, taupata (*Coprosma repens)*, mahoe (*Melicytus novae-zelandiae)*, knotted sedge (*Isolepis nodosa)*, flax (*Phormium tenax*), *Samolus repens* var. *strictus* as well as various other herbs and grasses. It is also found appreciable distances inland on the islands it inhabits, growing on the ground in canopy gaps and scrub, around shearwater and petrel burrows, as well as on cliff faces, boulder falls and rocky outcrops (Heenan, et al., 2004).

Arthropodium candidium is a species that grows from sea level to 1,300m a.s.l. especially in limestone country. It is very hardy and grows from seed in the alpine garden. It is significantly smaller than its sister species growing to a maximum of 36cm high with leaves less than 1cm wide (Richards, 1949).

Reproductive Biology

There appears to be a gap in the literature regarding the reproductive biology of *Arthropodium* in New Zealand. However, in a recent doctoral thesis, Australian *Arthropodium* were stated as being hermaphrodites (produce flowers with both male and female functions) and are generally rhizomatous or tufted-caespitose (Donnon, 2009). More than likely, Arthropodium is pollinated by unspecialised insect groups like many New Zealand flowers; *Arthropodium* flowers share the features of other New Zealand flowers that have unspecialised pollinators, simple flower structure and small flowers with non-showy colours (Lloyd, 1985). It is likely that *Arthropodium* propagules are dispersed aneomochoricaly due to the attachment of a long funicle (Thorsen, Dickinson, & Seddon, 2009) .

Distribution of New Zealand *Arthropodium*

The distribution of *A. cirratum* is shown in Figure 1 and it is most common in Northland, Auckland and Marlborough. There are only a few known locations in Taranaki, Hawke's Bay, Gisborne, Wellington, and Nelson (Heenan, et al., 2004). The natural distribution of *A. cirratum* is difficult to ascertain due to the plants value to Maori as a food and a medicine, thus the plant is likely to have been translocated. Also of note is that *Arthropodium* is very palatable to introduced browsers, such as goats and snails, therefore some of the present distribution may be recently diminished, with survival limited to areas that are inaccessible or inhospitable to these grazers.

The Rotorua, Hawke's Bay, Wellington, Nelson and Marlborough populations are suggested to have been translocated by Maori and have persisted there naturally (Harris & TeWhaiti, 1996; Heenan, et al., 2004). In particular, the Cape Palliser populations are associated with Maori archaeological sites that haven't been actively occupied since the 1600's. Also there is evidence that stone walls were erected at one of the Cape Palliser sites, east of the Cape Palliser lighthouse (Harris & TeWhaiti, 1996), and that they may have been solely for the propagation of *A. cirratum*. There is now no *A. cirratum* to be found in the garden ruins but it is growing on the nearby sheer bluffs that are out of reach of introduced herbivores. The Kairakau population on the east coast below Hawke's Bay is the most isolated of all populations; it has been suggested that this population too is associated with a Maori archaeological site (Heenan, et al., 2004) as well as the southernmost population at Oaro, south of Kaikoura. In the Rotorua, Cape Palliser and Kaikoura populations, karaka (*Corynocarpus laevigatus*) trees, another source of food for the Maori and likely also cultivated and translocated, are to be found very close by (Clarkson, 1991). A similar pattern is apparent in the more numerous Nelson and Marlborough Sound populations. The Nelson populations are associated with sites where pits, artefacts and middens have been discovered. The 41 sites where *A.cirratum* occurs in the Marlborough Sounds are directly connected with some of the 330 known Maori archaeological sites in the area (Heenan, et al., 2004).

The countering distributional hypothesis is that the current distribution of *A. cirratum* is natural. In this case the absence of many lower North Island populations is due to anthropogenic habitat modifications and removal by browsing animals. The markers developed in this thesis may help determine whether the southern populations of *A.cirratum* derive from translocations or whether the current distribution is naturally occurring.

The distribution of *A.bifurcatum* is mainly insular (Figure 1) - Three Kings Islands, Poor Knights Islands, Cavalli Islands, and Hen and Chicken Islands; there are also scattered mainland populations. The southern limit is Whatapuke Island, one of the Northern Hen and Chicken Islands south of Whangarei. Some of the *A.bifurcatum* populations are sympatric with *A.cirratum* populations.

The distribution of *A.candidum* is predominantly alpine in both of the major islands with some coastal populations being recorded at Dunedin and Te Waewae Bay, Fiordland (Institute., 1868).

Maori, Rengarenga and Oral History

Maori have a tradition of oral history (Haami, 2004). Tohunga, who held onto the history of the people, would select children who would retain the knowledge in their time; and so the histories of the people were transmitted. Most of the knowledge transmitted was to do with whakapapa – the genealogical relationship between people and their ancestors, and the ancestors and their gods. Rengarenga has little mention in written histories of Maori.

It is considered that the bulk of writing about Maori life appears in manuscripts (Sturm, 1991, p. 9) that are held by family members and are jealously guarded; to the extent that family members that are considered unsuitable will not be told of the existence or whereabouts of these manuscripts. They may typically contain- genealogy, prayers, songs, daily entries, tribal narratives, letters and the such-like.

There is a convention that manuscripts shall not contain information about food and tribal histories. This stems from the concepts of tapu and noa (sacred and everyday); food is noa and tribal histories/genealogies are tapu. Noa objects are able to destroy the power of tapu so that, if held together in the same manuscript, the information about food would effectively destroy the "living link" to the ancestors that those pages contain (Haami, 2004). Perhaps this is why there is scant written knowledge about the translocation of food crops as all records of journeys are related to whakapapa.

There is however information that has been passed on about the origins of particular Karaka groves, another food source. Whether this information was

- considered to be common property for Maori and so tapu restrictions were unable to be applied personally
- or whether the cultivation of Karaka was common and Rengarenga was rare, as indicated by the volume of information that has been passed on and is publically accessible,
- or whether there was a reason at all, is unknown.

One reference to Rengarenga concerns its symbolic power as a store for the mauri- the spiritual energy that moves- of the Maori people. The reference concerns a time where spiritual energy was being transferred from Hawaiiki – ancestral homeland of the Maori- to New Zealand. The energy of Hawaiiki was transferred into stones and plants, and these stones and plants represented a bridge between the two island nations (Tregear, 1904). Another spiritual reference refers to Maori souls waving fronds of Rengarenga on their passage north to Cape Reinga (the traditional departing spot for Maori souls on their trip back to Hawaiiki) indicating that they had died of natural causes (Riley, 1994).

Interestingly, the place where the Rengarenga was imbued on a sacred altar is Whangara, a small place about 30 kilometers (km) north of Gisborne. Supposing it was part of a natural population, it would be on the southern most limits of the proposed endemic, untranslocated range.

Another reference to Rengarenga comes from the observations of Colenso. He noted the use of a pattern called Rengarenga in one of the wharenui – Maori meeting house - he visited. Wharenui are decorated with bold, repeatable patterns; he states that the pattern he saw was in imitation of the curved anthers typical of Rengarenga (Colenso, 1891).

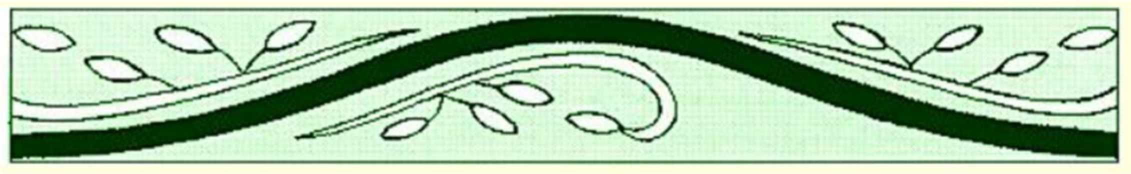


Figure 2: Drawing of a panel depicting the curved anthers typical of Rengarenga (Harris, 1996)

Maori used Rengarenga as a rongoa (medicine) (Palmer, 1988).Rengarenga roots were roasted, beaten then prepared as a poultice and used for the treatment of ulcers and used to soothe the swelling of joints and limbs, while unbroken tumors, abscesses, boils were treated by the lower ends of the leaves after being beaten then mashed into a poultice (Harris & TeWhaiti, 1996).

Maori actively cultivated food plants and one of these was Rengarenga (Colenso, 1881). Consensus appears to be that settlers to the Pacific brought their own food plants with them (Biggs, 1997; Finney, 2006) as well as other items of value like obsidian (A. J. Anderson, 2000; Walter, Jacomb, & Bowron-Muth, 2010). It seems a reasonable supposition that when ancient Maori travelled they took food plants with them. Therefore, if whakapapa states that an ancestor arrived at a place in New Zealand, it would also be possible that food plants were taken and planted at these places. This is the proposed method of translocation.

Internal voyages would probably have the same way of operating in regards to food plants. One account states that traditionally, karaka trees were on the voyaging canoes and that the local people distributed them, planting them by tracks as a source of food and as markers of the sacred (Haami, 2004). Another author suggests that the majority of voyages made in the first 200 years of Maori settlement were made on land (Walter, et al., 2010).Unexpected events often happen, like changes in the weather, and having a ready source of food at strategic locations would appear to be a survival tool.

Initial travels in New Zealand were taken by eponymous ancestors- generally from their homes in the north of the North Island to unsettled places further south. Tamatea-pokai-whenua was born just out of Kaitaia, in the north of the North Island. He travelled extensively with features named after him in Kawhia, Wanganui, Porirua, and Hawke's Bay in the North Island; Marlborough Sounds, Christchurch and Invercargill in the South Island. As these places are remembered in oral histories centuries later, it would appear he did not travel alone (Biggs, 1997).

Ngai Tahu, the iwi (tribe) that lives in the South Island, had its eponymous ancestor, Tahu-potiki, call Te Tai Rawhiti (East Coast of the North Island) the place of his upbringing. Ihenga and Kahu-mata-momoe travelled from the north of the Coromandel into the Rotorua, Kawhia and Waikato regions, naming places as they went. The Tainui and Takitimu waka (two of the seven legendary boats that brought the Maori to New Zealand) both made many stops around New Zealand with sites in Northland being stops early on in their journeys (Biggs, 1997).

Maori have continued their many voyages from their initial landing in the mid 13th century, although this date is often disputed (A. Anderson, 1991; Higham, Anderson, & Jacomb, 1999). Archeological records show that there were only 37 Pa (Maori villages), 18 in the North Island, 19 in the South, that are more than 600 years old (Biggs, 1997). After that time, expansion was rapid and the demand for food increased rapidly. The Maori concept of "mana" (worth of an individual or family or tribe) was measured by how many generations of your ancestors, had been constantly holding and using the cultivations and other resources of an area (Haami, 2004). Pre-European Maori moved between their cultivations and resource zones depending on the time of year each resource required attention. Management of resources were a priority and cultivation of food plants a key tool for survival (Colenso, 1881).

Archaeological evidence suggests that long journeys of 1200 km were made, in this case, from Mayor Island, in the Bay of Plenty, to sites on the west coast of the South Island (Walter, et al., 2010). Both sites were small and coastal, occupied for perhaps a few decades sometime in the late 14th century. Flakes of Mayor Island obsidian were found at both sites with 60% of the flakes being unused (Sheppard, 1993). This might indicate that a sea voyage of this distance was held in small regard if valuable cutting tools sourced so far away were left unused. Maybe sea voyages of this distance were easily made; if so, the potential translocation of Rengarenga from Northland to regions further south would appear to be easily accomplished.

Although not explicitly stated in any printed record, the associations of Rengarenga with archaeological sites makes it highly likely that Maori transplanted Rengarenga from its location of highest density in Northland to locales further south.

Next Generation Sequencing (NGS)

The advent of high throughput sequencing provides a rich resource for molecular marker development – in particular the high coverage, accuracy and relative lower cost of Illumina sequencing - has considerable potential for enabling determination of translocation patterns. As Single Nucleotide Polymorphism (SNP) discovery is a way to distinguish between different accessions, a technology that delivers DNA sequence in a massively parallel fashion is going to be a tool of choice.

The technology for sequencing DNA began to blossom with the introduction of Sanger (dideoxy) sequencing in the late 1970's. Fredrick Sanger and colleagues developed a method of terminating the chain-building effect of DNA polymerase with dideoxynucleotide triphosphates. Deoxynucleotide triphosphates are included in the reaction and one of these is typically isotopically labelled, so that the size of the synthesised strands can be visualised after electrophoresis (Sanger, Nicklen, & Coulson, 1977). Sanger sequencing machines, implementing fluorescent detection, instead of using isotopes, currently can sequence up to 1000bp of DNA very reliably and cheaply; however this technology is limited by the length of DNA that can be sequenced, making the construction of genomes that are millions of bps long a very tedious and time consuming process. Using Sanger technology, the human genome project took 10 years to sequence, assemble and cost three billion dollars.

NGS technologies are also based on strand synthesis protocols. Illumina sequencing involves imaging fluorescent emissions that occur as each dNTP is incorporated progressing 1 base/1 photo at a time (Turcatti, Romieu, Fedurco, & Tairi, 2008). This process is discussed in more detail in the Methods section. The end result is over 16 million "reads" of length 32-76bp.

Error rates per bp are over 10 times higher than traditional Sanger platforms (Shendure & Ji, 2008) with an error rate of 0.5%. The error rate for fingerprinting has been estimated as being in the 2-5% range (Bonin et al., 2004). The advantage that NGS has is in the depth of coverage. Each bp in a sequence may be reproduced over 10,000 times in different reads so that, when aligned to a reference sequence or a *de novo* contig (sequence generated by overlapping reads), the position of the nucleotide (nt) in a sequence is confirmed by the depth of the coverage.

The higher error rate in the raw reads, produced by Illumina sequencing, is ameliorated by the use of bioinformatics tools that allow for the

- exclusion of the end bases of a read, a region that tends to have a higher error rate
- exclusion of reads that are under a certain length
- ascertainment of the quality of a certain run thereby allowing for trimming of reads that have a lower quality score (Deschamps & Campbell, 2010)

NGS is proficient at precisely locating DNA polymorphisms that are present within single and multiple samples (Rokas & Abbot, 2009). The great depth of coverage increases the accuracy for SNP discovery. SNP's discovered in this project were generated from the Illumina GAIIx.

Materials and Methods

Samples and Collection

For the chloroplast extractions, fresh leaf material was harvested from two species of *Arthropodium*, namely *A.cirratum* and *A. bifurcatum* under cultivation at the Turitea Campus of Massey University, Palmerston North. These accessions were chosen over field collected material after chloroplast extraction proved to be difficult and large amounts of fresh materials were needed. University records do not state the origin of these plants other than stating that they were purchased from a nursery (G.Mack, *pers comm*.).

Genomic DNA was extracted from fresh or silica-gel dried leaf tissue from 18 *A. bifurcatum*, 101 *A cirratum* and 2 *A. candidum*. Sample details are provided in Table1 in the Appendix.

Genomic DNA Extraction Protocol

DNA was extracted using either a DNEasy Plant Mini Kit (Qiagen) following the manufacturer's instructions or a CTAB procedure (Doyle & Doyle, 1990). This procedure was modified by using liquid nitrogen to freeze the silica dried samples which were then ground using a pestle and mortar before incubating in the CTAB buffer (2% hexadecyltrimethylammonium bromide [CTAB: Sigma H-5882], 1.4 M NaCl, 0.2% 2-mercaptoethanol, 20 mM EDTA, 100 mM Tris-HCl, pH 8.0) at 60°C for 40 minutes. After incubation, chloroform was used to denature proteins and solubalise membranes; cold isopropanol was used to precipitate the DNA and ethanol (80% EtOH, 10 mM ammonium acetate) to clean the polysaccharides from the DNA.

Chloroplast Isolation Protocol

Leaf samples were cut from living plants and the chloroplasts isolated using a modified procedure based on one designed for isolating chloroplasts from *Arabadopsis thaliana* (Aronsson & Jarvis, 2002). During this procedure all materials were kept at 4°C. Fresh plant material weighing 5g was homogenized with 20ml isolation buffer (0.3M sorbitol, 5 mM MgCl2, 5 mM EGTA, 5 mM EDTA, 20 mM HEPES/KOH, pH 8.0, 10 mM NaHCO3) in a 50ml test tube, using an Ultra-Turrex homogeniser with an N18 rotor (Janke & Kunkel IKA, Hamburg, Germany).

The resulting homogenate was then filtered through a double layer of washed and autoclaved nappy liner (Johnson & Johnson Ltd.) and then loaded onto a two-step Percoll gradient that had been prepared the previous day in 50ml Corex tubes (Dupont). The gradients consisted of a bottom layer (20ml) comprising of a Percoll solution [30% w/v Percoll, 70% MilliQ, 120µl 0.5M EDTA (Ethylenediaminetetraacetic acid**)** pH8,30 µl 1M $MgCl_{2,}$ 3ml 0.5M Tricine, 1.8gm Sorbitol], a top layer (20ml) comprising of a Percoll solution [60% w/v, 40% MilliQ, 120µl 0.5M EDTA (Ethylenediaminetetraacetic acid**)** pH8,30 µl 1M $MgCl_{2,}$ 3ml 0.5M Tricine, 1.8gm Sorbitol] then 2ml of the filtered homogenate. The gradients were centrifuged using a Sorvall SS32 swinging bucket rotor.

The intact chloroplasts were found in between the two Percoll gradients and could be removed using a Gilson pipette. DNA was extracted from the chloroplast suspension using a DNEasy Plant Mini Kit (Qiagen) following the manufacturer’s instructions (Atherton et al., 2010).

Multiply-primed rolling circle amplification

Multiply-primed rolling circle amplification (RCA) was performed to make a profusion of purified chloroplast DNA template in preparation for sequencing. This system involves isothermal, strand-displacement amplification using multiple primers and is proficient at yielding a large quantity of DNA from very little starting template. Phi29, the DNA polymerase used in multiply-primed RCA, has the property of

- generating a very high proportion of specific amplicons (no nonspecific amplification artefacts)
- giving complete coverage of loci
- generating DNA of lengths > 10kb (Dean et al., 2002)

Chloroplast-enriched DNA (cpDNA) from both Rengarenga samples was amplified in this way using a REPLI-g Mini Kit (Qiagen) following the manufacturer’s instructions.

In brief, 2.5 µL of DNA was added to 2.5 µL of Buffer D1 (containing KOH & EDTA) to denature the DNA. The sample was incubated for three minutes at room temperature, before adding 5 µL of Buffer N1 (neutralizing buffer). Then, 40 µL of Master Mix, containing the REPLIg DNA polymerase, were added to 10 µL of the denatured DNA solution. The mixture was incubated overnight at 30°C. After amplification, the REPLI-g DNA polymerase was heat-inactivated during a ten minute incubation step at 65°C.

EcoR1 Digestion

Prior to PCR, the purified and enriched cpDNA was digested by EcoR1 Restriction enzyme in order to confirm the presence of plastid DNA. Because of the small size of the plastid genome in comparison with the nuclear genome, when sample DNA is digested by a 6kb cutter such as EcoR1, the resultant agarose gel will show discrete bands for plastid DNA and a smear for nuclear DNA.

The method was to add 5 µL of DNA to 5 µL 10× Template buffer (27.5nM $MgCl_2$; Roche Applied Science, Auckland), 2 µL EcoR1 (10U/µL; Roche Applied Science, Auckland) and 38µL MilliQ. This was put on a Biometra Thermocycler with the following conditions- 37°C for 180 min to cut the DNA then 70°C for 15 min to inactivate the enzyme. Results were then visualized on agarose gels.

Illumina GAIIx Sequencing

Arthropodium chloroplast DNA was sequenced by the Massey Genome Service using an Illumina GAIIx. The Illumina GAIIx uses second generation sequencing technology; a process typified by cyclic array sequencing. This process may be summarised as the sequencing of thickly packed clusters of DNA by repeated cycles of enzymatic manipulation and photographic data collection (Shendure & Ji, 2008).

Initially, a library has to be constructed. The DNA to be sequenced is randomly sheared into fragments approximately 800bp long; this is done by passing compressed air (>32psi) into a nebulizer that contains the DNA and TE. The ends then need to be repaired; T4 DNA polymerase fills in the 5’ overhangs and the exonuclease action of a Klenow enzyme removes the 3’ overhangs (QIAquick PCR Purification Kit [QIAGEN, #28104]). “A” bases are added to the 3’ end of the fragments using the polymerase action of a Klenow fragment; this prepares the fragments for attachment to the adaptors which have a single “T” base overhang (MinElute PCR Purification Kit [QIAGEN, #28004]).

This product is then cleaned to remove unattached adaptors and to select a size range of templates to go on the flowcell. 30µl of the purified ligation product from the last step is loaded onto a 2% agarose gel (2% agarose, MilliQ, 1XTAE, ethidium bromide [400ng/ml]) and run at 120V for 120 minutes. Template of the desired size is then excised from the gel and eluted in 30ml of EB Buffer (QIAGEN).

The purified product is then amplified by PCR using primers (PCR Primer PE 1.0 & 2.0) that anneal to the ends of the adaptors – 30 seconds at 98^0C, followed by 18 cycles of: 10 seconds at 98^0C, 30 seconds at 65^0 C and 30 seconds at 72^0 C, then 5 minutes at 72^0 C, hold at 4^0 C. 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.

This sample is moved to a fluidics device that hybridises samples onto a flowcell and amplifies them ready for sequencing. This device is known as a cluster station. Clusters are generated by the following process-

- template DNA is hybridised onto the oligonucleotide-coated surface of the flowcell by the adaptors that were attached in a previous step
- this DNA is then amplified, using a technique known as bridge PCR, which generates clonal DNA clusters, Bridge PCR is a technique in which both forward and reverse PCR primers are attached to the flowcell by a linker; all amplicons generated from that particular sequence are therefore fixed to a physical location on the flowcell and result in a cluster of that sequence of approximately 1000 copies. Several million clusters can be identifiable in a single lane, out of 8 lanes on a "cell".
- chemically linearize the dsDNA clusters; this begins the conversion of dsDNA to ssDNA that is suitable for sequencing
- free 3' OH ends are blocked thereby preventing the sequencing of nonspecific sites
- denaturing the dsDNA and hybridising the sequencing primers onto the linearized and blocked clusters

The flowcell is then ready for sequencing.

Each amplicon in a cluster then has a sequencing primer that is hybridised onto the universal sequence that lies adjacent to the DNA fragment. A modified DNA polymerase and a mixture of the 4 nucleotides begin their cycle; the nucleotides have been altered in 2 ways- reversible terminators have been added so that only one nucleotide may be added during each cycle and each nucleotide has one of 4 fluorescent labels attached. So, all the single stranded clusters have a complementary nucleotide added before a photo is taken. The fluorescent label and blocking agent are cleaved for the next cycle to begin. The series of photographs taken capture the addition of each nucleotide to the growing reads generated at each physical cluster location (Deschamps & Campbell, 2010; Garvin, Saitoh, & Gharrett, 2010; Shendure & Ji, 2008) in a process that has been likened to "photographing the twinkling of city lights over the course of a night" (Rokas & Abbot, 2009, p. 1).

Evaluation of Raw Read Data Quality

In the first instance, with the sequencing of the *A.bifurcatum* genome, there was no evaluation of the raw read data other than the analysis with the proprietary Illumina pipeline v1.3 software. This software gives operators of the Illumina sequencer detailed information about data quality for each lane, tile and nucleotide position but is usually not released to the user. At the time my sequencing project was undertaken there was no readily available data quality evaluation tool available to the non bio-informaticist. This changed with the release of SolexaQA (M. Cox, Peterson, & Biggs, 2010).

Error profiles of high throughput short sequence reads, like the Illumina GAIIx, have a propensity to display a steep exponential increase in error rate along the length of the read (M. Cox, et al., 2010). As such, with the *A.bifurcatum* genome, the 3' ends of the reads were trimmed to remove this poor quality sequence using the program FASTX-Toolkit (Hannon, 2010). This effectively chops the specified length of nucleotides off the end of each read.

In the second instance, with the sequencing of the long-range PCR products, data quality was evaluated using the SolexaQA program. This suite of analytical tools uses the quality information that is stored in the fourth line of the FASTQ file format to analyse data quality. The program randomly selects 10,000 reads from each tile at each nucleotide position per cycle and calculates the mean quality scores. These scores may be presented as a heat map, a line graph or in a tabular form. From this output, low quality data can be easily identified and discarded; a histogram of maximized read lengths is also produced from which users can deduce the optimal read length to which to trim their reads. Incorporated into the suite is a program called DynamicTrim; this trims the reads to their longest contiguous read segment based on a user defined quality score limit (M. Cox, et al., 2010).

Assembly of Chloroplast Contigs and Genome

Initial assembly of the chloroplast reads into contigs (contiguous sequences) was done using the de novo assembler, Velvet (Zerbino & Birney, 2008). Velvet is a series of different algorithms that have been generated for constructing long contigs from short read sequences.

First the reads are hashed according to a predefined kmer length. The key concept that allows Velvet to construct large contigs is that rather than using the reads directly, Velvet uses a portion of the read (known as a word or a kmer) to look for similar words in other reads that differ by one nucleotide at the beginning or end of the word. This way, there has to be a near total agreement in the nucleotides (bar the additional nucleotide) before an additional kmer from a different read is added to a growing sequence (contig).

Velvet notes from which read the kmer derives; it also notes the position of the kmer within the read. This information is determined for all kmers in all reads and the relationship is known as a "Roadmap". In de Bruijn graphs these relationships are denoted using the concept of nodes and arcs (blue boxes and arrows). A highly simplified illustration of this process is shown in Figure 3 below.

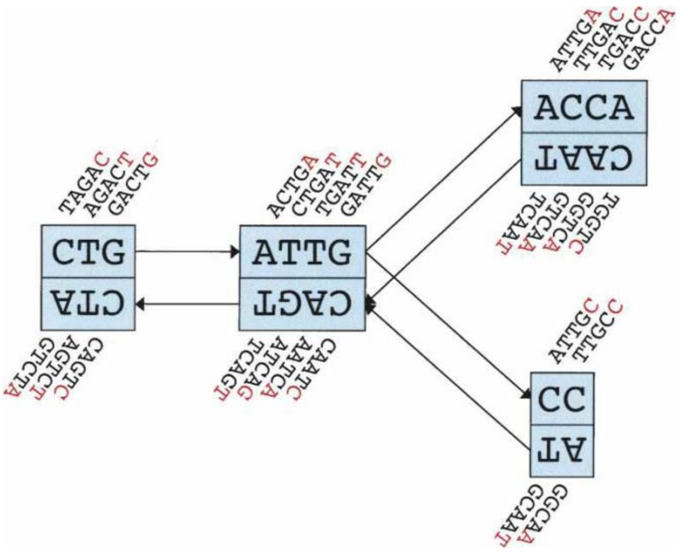


Figure 3: A simplified schematic of sequence construction using Velvet (Zerbino & Birney, 2008).

Boxes represent nodes, which are a collection of kmers 5 nucleotides in length, and arrows the arcs that join the boxes together. The agreement between the nucleotides in the kmers continues from one node to another; these kmers may come from the same or different reads. The nucleotides in red represent the sequence of the growing contig (Imelfort & Edwards, 2009; Zerbino & Birney, 2008).

In analysing *A.bifurcatum* reads, a range of kmer values were trialled in an effort to maximise the contig length and the N50 values. The N50 is calculated by arranging the contigs, by ascending length, end-to-end. The mid-point of this concatenation identifies a contig whose length is the N50 value of that Velvet run.

In the assembly of *A.bifurcatum* reads the following parameters were also considered -

- minimum contig length,
- depth of coverage before contig included in final output,
- expected coverage, which may be calculated when the approximate size of the genome is known,
- coverage cut-off which removes nodes with low coverage

Velvet outputs a file with a series of contigs based on the chosen parameter values. In the present work this file was then imported into Geneious software (Drummond et al., 2009) where these individual contigs were analysed in BLAST searches (Tatusova & Madden, 1999) to validate their chloroplast origin and discover the closest chloroplast genome to align the contigs against.

Geneious was used to align the contigs to the closest reference genome using the default Geneious algorithm. DOGMA software (Wyman, Jansen, & Boore, 2004) was then used to annotate the genome. DOGMA works by finding the orthologues in other annotated plant species using BLAST comparisons with a limited set of reference genomes. It identifies probable start and stop positions of genes.

PCR and Long-range PCR

In the first instance, PCR was used to amplify the three regions of chloroplast DNA where gaps were indicated in the assembled contigs. Primers were designed using Primer3 (Rozen & Skaletsky, 1999) to amplify these regions.

Table 1: Shows short range primers used to fill the gaps

Location	Primer name	Primer sequence (5' to 3')
24,820	Insert1 F	TGGATGGAGACGGGGCGGAG
25,276	Insert1 R	TGCCCATGTTGCCCCAAGTGA
47,912	Doubt1 F	ACCAAGCGTCTCCCATGGGTTCT
48,915	Doubt1 R	AGCGAAACCTAACCCACCGCT
105,439	NoContig F	ACTATTTACACGGCATCGCGCC
105,784	NoContig R	GCCGCCACTCGGACTCGAAC

PCR amplification was performed in 20 µl volumes containing 2 µl of 10× PCR buffer (10 mM Tris–HCl, 50 mM KCl, pH 8.3; Roche Applied Science, Auckland), 3.8 µl of 1 M betaine, 1 µl of 5 mM dNTPs, 1 µl of 10 ρmol/ µl of each primer, 0.2 µl of 1 U Taq polymerase (Roche Applied Science, Auckland) and 1 µl of 50 ng/ µl of template DNA to which 10 µl of MilliQ was added.

The PCR cycling conditions were as follows- for Doubt 1, template denaturation at 94°C for 3 min followed by 35 cycles of denaturation at 94°C for 30 seconds (secs), primer annealing at 50°C for 30 secs, and primer extension at 72°C for 45 secs; followed by a final extension step of 5 min at 72°C. For NoContig and Insert1 a long-range cycling condition was followed – template denaturation at 94°C for 3 min followed by 35 cycles of denaturation at 94°C for 30 secs, primer annealing at 50°C for 30 secs, and primer extension at 68°C for 45 secs per kilo base (kb) of sequence; followed by a final extension step of 10 min at 68°C. NoContig was estimated at being 5,500 base pairs (bp) long and Insert at 1,700 bp long.

Amplified PCR products for the gaps 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 Windows (Gene Codes Corporation, Ann Arbor, MI).

In the second instance, PCR was used to amplify the entire DNA sequence in both the Long Single Copy (LSC) and Short Single Copy (SSC) regions of the chloroplast genome for 18 different accessions (see Appendix, Table 4) in an attempt to discover SNP's. Primers were designed using Primer3 software and may be found in the Appendix, Table 2.

PCR amplification was performed in 25 µl volumes containing 2.5 µl of 10× Expand Long Template buffer 3 (27.5nM $MgCl_2$; Roche Applied Science, Auckland), 4 µl of 5M betaine, 2.5 µl of 5mM dNTPs, 2 µl of 10 ρmol of each primer, 0.38 µl of 1 U Taq polymerase (Roche Applied Science, Auckland) and 1 µl of 50 ng/ µl of template DNA.

The PCR cycling conditions were as follows: template denaturation at 93°C for 2 min followed by 10 cycles of denaturation at 93°C for 30 secs, primer annealing at 50°C for 30 secs, and primer extension at 68°C for 6 minutes. A further 24 cycles of denaturation at 93°C for 30 secs, primer annealing at 50°C for 30 secs, and primer extension at 68°C for 6 minutes increasing by 20 secs a cycle; this was followed by a final extension step of 10 min at 68°C.

Sequencing and Mapping of PCR Products to *A.bifurcatum* Chloroplast Genome

Amplified PCR products were sequenced on the Illumina GAIIx then the reads were aligned to the *A.bifurcatum* chloroplast genome using Bowtie software (Langmead, Trapnell, Pop, & Salzberg, 2009). This indexes the reference genome using a Burrows-Wheeler index then quickly aligns the reads producing a SAM file. This file was visualized using TABLET software (Milne et al., 2010). Visualization allowed for identification of potential SNP's as well as providing confirmation of PCR products that covered the gaps in the contigs.

Results

Chloroplast preparation

Sucrose gradient

A number of different sucrose gradient combinations were trialled to determine which gave the highest and cleanest whole chloroplast count (Table 2). Chloroplast counts were taken from the smallest square on the haemocytometer covering an area of 0.0025 mm^2. The two gradients selected for subsequent chloroplast extraction were 20:40 for the *A.cirratum* (tube 1) and 30:60 for the *A.bifurcatum* (tube 7); both chloroplast samples were siphoned from the top layer.

Table 2: Shows results from chloroplast distribution in sugar gradients. The heading "upper" refers to the upper band of chloroplasts that lies on top of the lightest layer of sugar solution. "Middle" refers to the band of chloroplasts that lie between the two sugar gradients. The "green colour intensity" is a subjective scale indicating the intensity of chloroplasts in each sample. Debris is a relative count of the ruptured chloroplasts, when viewed by light microscopy.

Tube no.	Gradient	Species	Upper	Middle	Green colour intensity/10	Debris/10	Count
1	20:40	cir	x		6	0	34
2	20:40	cir		x	1	2	9
3	30:60	cir	x		4	1	18
4	30:60	cir		x	5	0	26
5	20:40	bif	x		9	0	50
6	20:40	bif		x	2	1	3
7	30:60	bif	x		10	0	57
8	30:60	bif		x	3	1	6

Chloroplast visualization on agarose gel

Figure 4 shows the gel electrophoresis of (i) molecular weight marker and (ii) an EcoR1 digest of chloroplast enriched *A.bifurcatum* DNA (Sample 1378, see Appendix, Table 1). Although the photograph lacks resolution, discrete bands could be observed above the background smear in lane (ii). This result is consistent with the presence of both chloroplast and nuclear DNA in the sample extract.

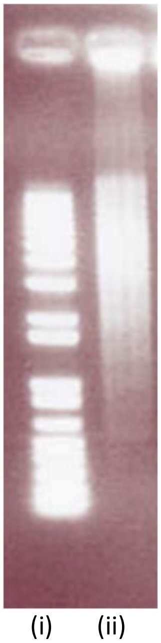


Figure 4: Shows the results of EcoR1 digestion.

SolexaQA and de Novo Assembly

SolexaQA

Reads from the long-range PCR were analysed with SolexaQA with the indexed adaptors still attached. Figures 5 & 6 show the mean quality per tile and cycle. Figure 5 is a heat map, showing that there have been complete tile failures, as displayed by the grey lines, for tiles 1, 2, 3, 5, 9, 11, 12, 13, 15, 91, 92, 95, 96, 98, 100, 103-107, 109, 113, 114 & 117. The gradual shading of yellow near the end of the read length indicates that data quality decreases near the end of the reads.

Figure 6 shows the mean probability of an error at each nucleotide position, represented by the red circle. Individual tile scores are represented by the black dots. The line indicates that the mean probability of an error at each nucleotide position increases along the length of the read. The peak at the beginning of Figure 6 shows the region where the indexed adaptors lie and indicates low quality in the adaptor not the read itself.

Figure 7 shows that the bulk of the reads passed the quality threshold (p= 0.05) and are 74 or 75 nucleotides in length. Thus the data appears largely to be comprised of reads of good quality.

Dynamic trim was used to separate high quality reads from lower quality reads. Only the former were used for subsequent analyses.

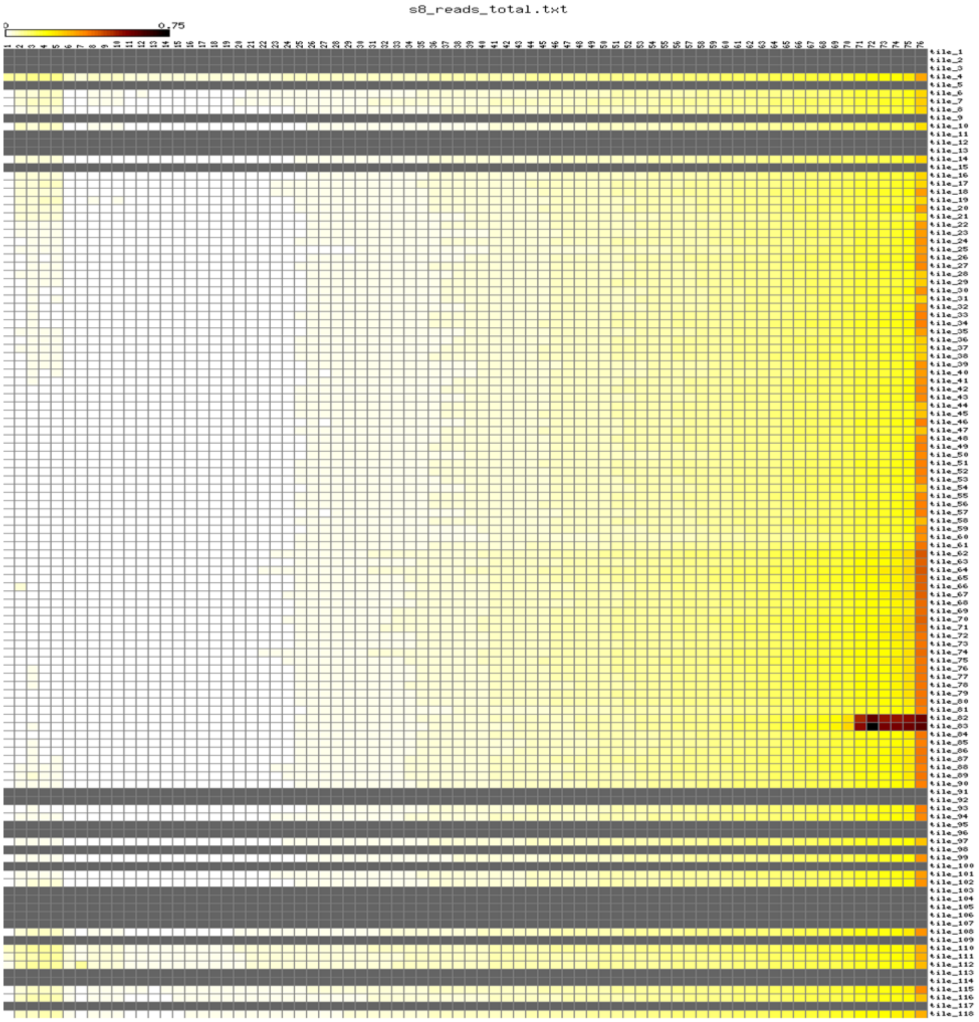


Figure 5: Heat Map – shows quality degenerating near the end of the reads

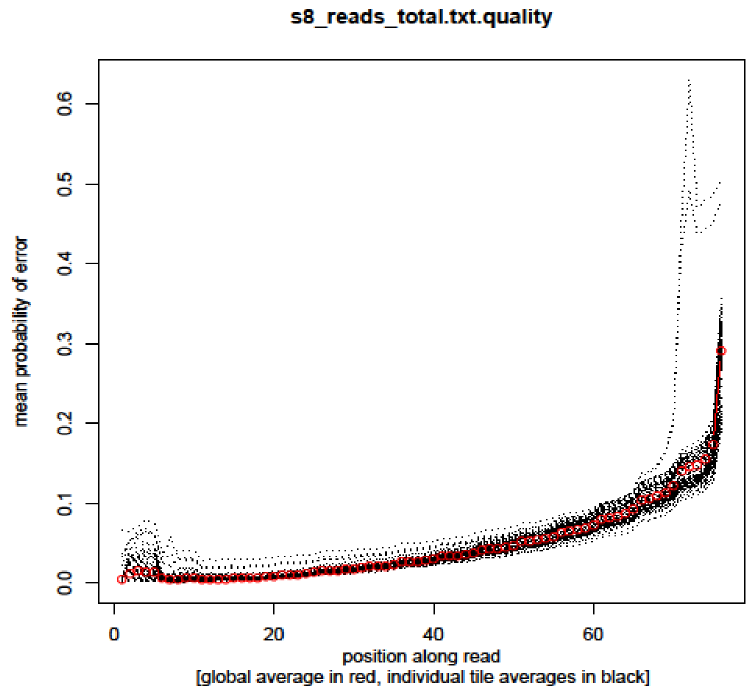


Figure 6: Nucleotide mean quality

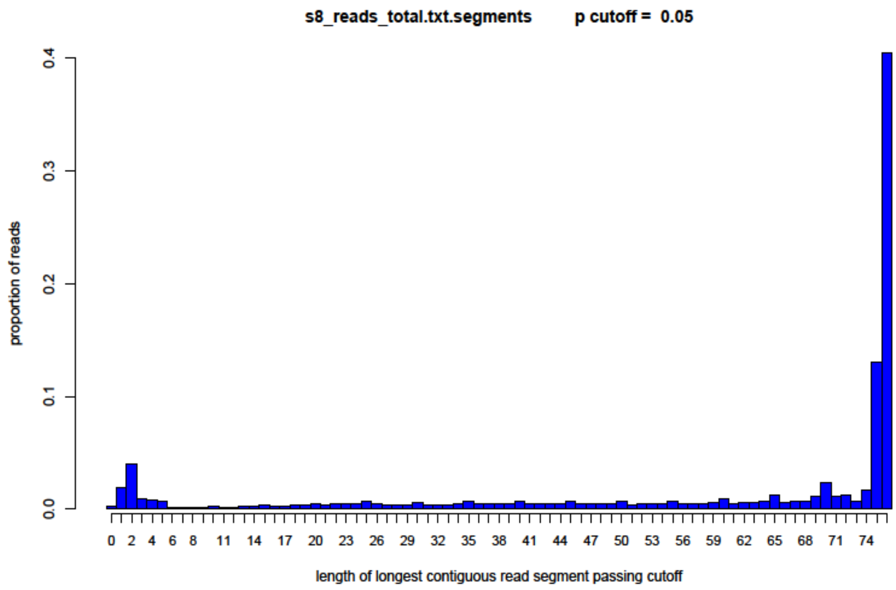


Figure 7: Distribution of longest reads that pass quality assessment

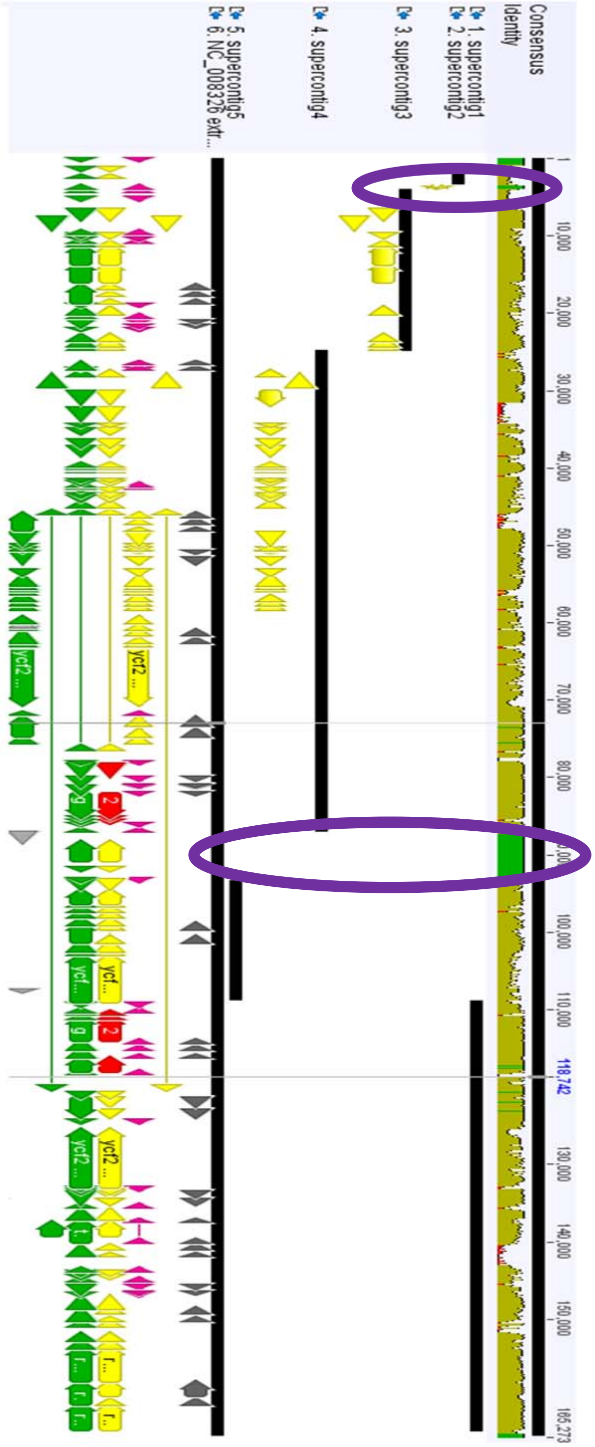
de Novo Assembly

Velvet was run initially with minimum contig length set to 100 nucleotides and expected coverage was set to default. A range of kmer lengths were trialled. Kmers of 47, 49 & 51 gave the largest contig length and N50 value. Coverage cut-off values were then maximised to see if larger contigs could be obtained (see Appendix, Table 3). Optimal settings were found to be: kmer = 47, coverage cut-off = 25 and expected coverage = 150.

Table 3: Shows results from differing kmer lengths.

k	Nodes	N50	Max Contig Length	Total
35	21899	257	9066	1091194
37	13114	698	11855	930780
39	7180	1643	18405	822993
41	4293	2487	18942	762844
43	2457	3124	18940	715016
45	1497	3726	18938	682439
47	1035	5103	23699	657981
49	743	5968	19155	642131
51	599	5536	19252	628132
53	547	4530	19252	614186
55	666	2442	20425	590825
57	983	876	9834	483356
59	232	1059	11520	111771

The contigs from Velvet were imported into Geneious. Contigs confirmed by BLAST analyses as being of a chloroplast origin, were assembled into five large super contigs (Figure 8). These were subsequently aligned to the closest reference genome, *Liriodendron tulipifera* (Genbank: NC008326) which was the most similar available at the time of study. There were four regions where there were gaps between the supercontigs; three of these were resolved using a combination of short range PCR followed by Sanger sequencing, aligned to the reference genome using Bowtie and visualised using Tablet. Although a number of primers were trialled, PCR products to close the fourth gap were not obtained, illustrated in Figure 8 by the lower highlight.



Highlights an “assembly gap”, discussed p48.

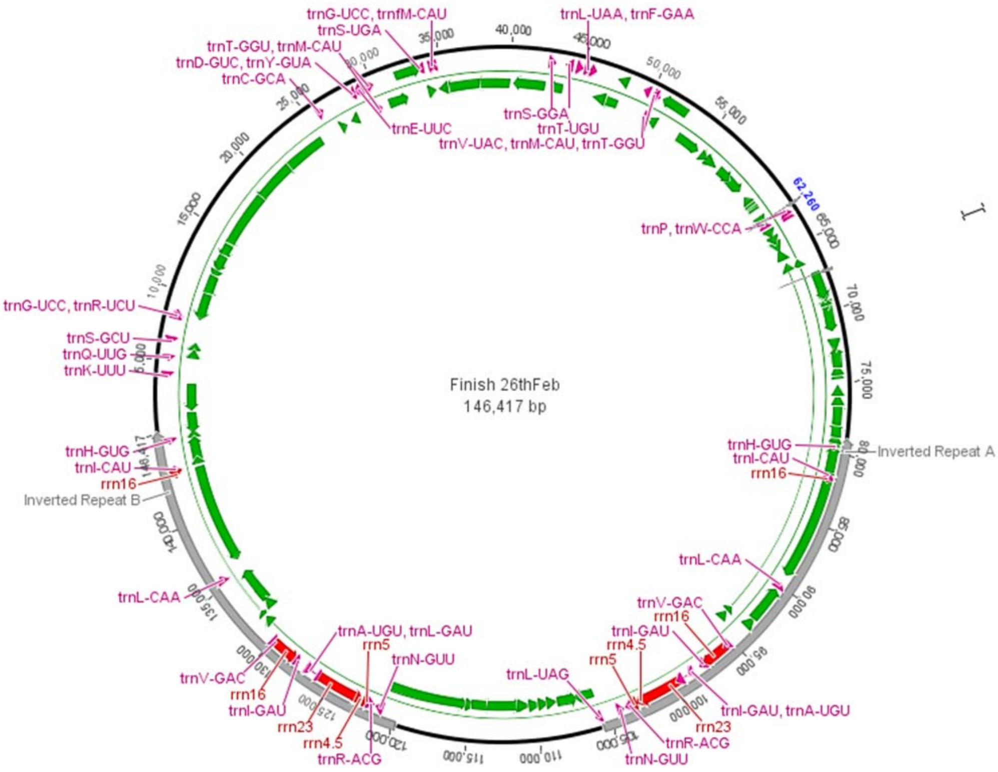
Highlights the IRA/SSC junction gap, discussed p47.

Figure 8: Shows the alignment of the supercontigs to reference genome NC 008326. The purple circles represent two of the major gaps found in the alignment.

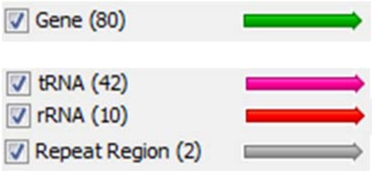
Annotated *A.bifurcatum*

The complete sequence length of the *A.bifurcatum* chloroplast genome was found to be 146,417 bp. The genome includes a pair of inverted repeats (IR) 26,413 bp in length separated by a SSC of 14,559 bp and a LSC of 79,032 bp. The genome comprised of 80 genes, 7 of which are repeated in the IR, 42 tRNA (transfer RNA) and 8 rRNA (ribosomal RNA), 4 in each IR. The overall GC content across all contigs was found to be 38.5%.

A.bifurcatum Chloroplast Genome Annotated



Key



☑ Gene (80)

☑ tRNA (42)

☑ rRNA (10)

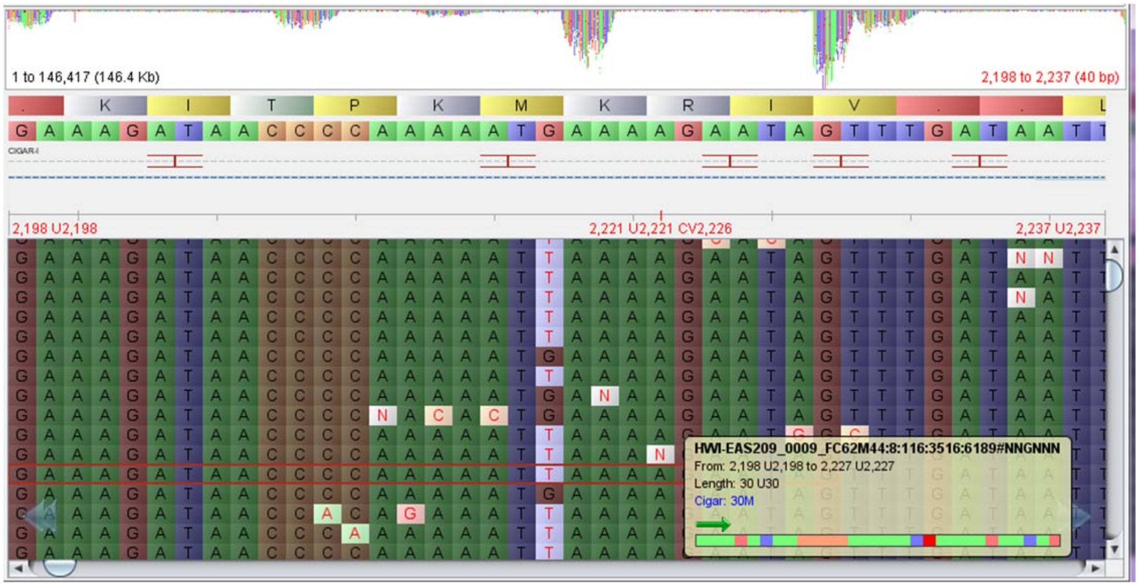
☑ Repeat Region (2)

Figure 9: Shows the annotated *A.bifurcatum* genome

A.bifurcatum Chloroplast Genome with SNP's

SNPs between the long range PCR products from multiple samples and the *A. bifurcatum* chloroplast genome were visualised in Tablet. Visualisation with Tablet made it relatively easy to identify SNPs in relation to read coverage (illustrated in Figure 11 below). In the single copy regions of the genome, coverage was extremely high for some regions. Inferred SNPs have been tabulated in the appendix, their location in the genome has been illustrated in Figure 11.

Figure 10: Example of a putative SNP visualized using Tablet. The putative SNP "T" found in the central column is highlighted in light blue. The text box indicates the length of the read and details read coordinates from the Illumina cell.

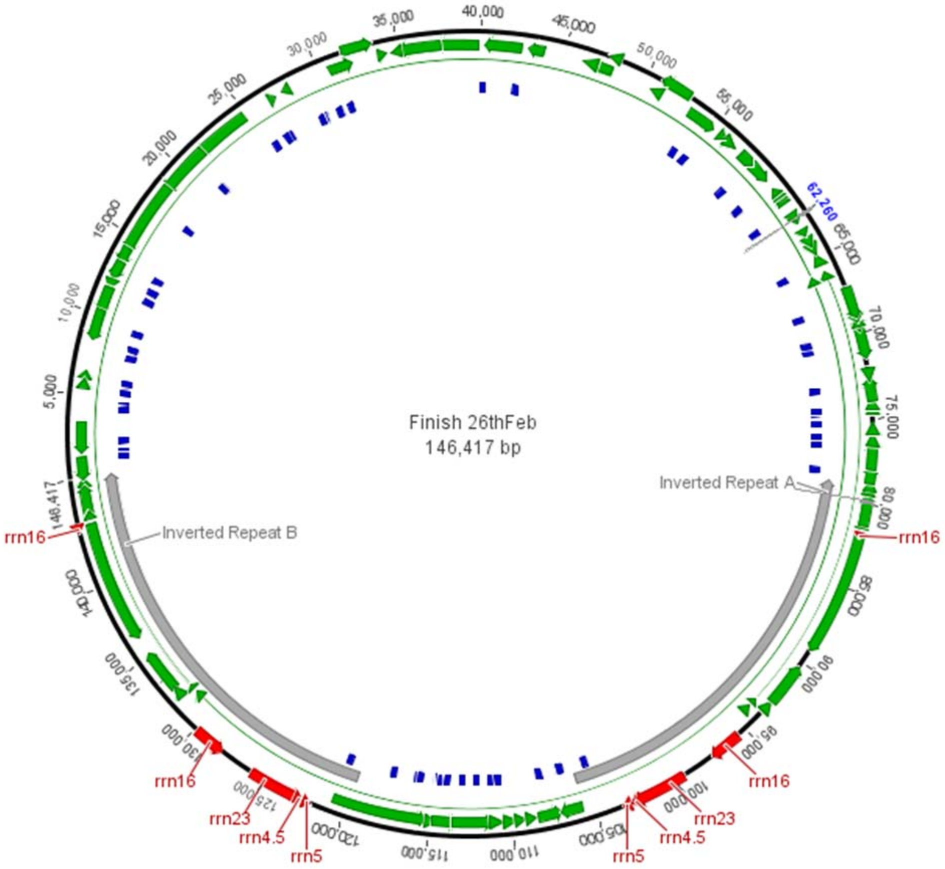


HWI-EAS209_0009_FC62M44:8:116:3516:6189#NNGNNN
From: 2,198 U2,198 to 2,227 U2,227
Length: 30 U30
Cigar: 30M

Table 4, below, indicates which accessions amplified successfully using long range PCR. PCR amplification of the IRA/SSC boundary (rps12-ndhF) was not successful.

Table 4: Shows long range PCR products that successfully amplified. The numbers along the top row indicate the different accessions and are listed in the appendix, Table 4.

PrimerPair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
LSC																		
rpl2-psbl	X	X	X	X	X				X	X	X							
psbl-rpoC2				X		X				X			X					
rpoC2-rpoB	X	X	X		X			X	X									
rpoB-rps14		X							X	X								
rps14-rps4	X	X		X		X		X		X			X	X			X	
rps4-rbcl	Some smaller frags 1-13													X		X	X	
rpcL-petG	X	X	X					X	X	X				X	X			
petG-petD		X		X						X								
petD-rpl2	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
SSC																		
rps12-ndhF	None but some smaller frags																	
ndhF-ndhl	X	X	X	X	X	X		X	X		X		X		X	X	X	X
ndhl-ycf1	X	X		X	X	X		X	X	X			X	X	X	X	X	X
ycf1-rps12	X	X		X		X			X	X			X	X	X	X	X	



Key

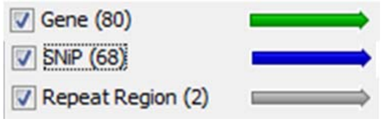

☑ Gene (80) 🟩
☑ SNiP (68) 🟦
☑ Repeat Region (2) ▨

Figure 11: shows *A.bifurcatum* genome with SNP's

Discussion

SolexaQA and de Novo Assembly

SolexaQA

Next generation short range sequencing technologies have been noted for their higher positional error rates when compared to the traditional Sanger sequencing with some reports stating a tenfold increase in the error rate of the former technology (Shendure & Ji, 2008). Error rates in the data obtained in the current study were around the 0.5% mark. The higher error rates in second generation sequencing are offset by the depth of coverage and bioinformatic analysis during and after sequencing; SolexaQA is a bioinformatic tool that allows the investigator to examine the qualities of sequence reads. SolexaQA was used to examine the quality of data determined for the present project.

The results, described in Figures 5, 6 & 7, show that the Illumina sequencing of long range PCR products, was of a high quality. The loss of average quality near the end of the reads, which can be observed in Figure 6, is typical of Illumina runs and is thought to result from incomplete cleavage of fluorescent labels later stages of the Illumina sequencing process (Shendure & Ji, 2008).

Figure 5 also indicates that data was lost completely from some tiles. As the tiles are spatially arranged in a U-shape when sequenced (i.e. tiles 1 and 100 are located next to each other at one end of the flowcell, tiles 50 and 51 at the other end), the failure of tiles near the beginning and end of the run (Figure 5) potentially indicates that an oil loss or spill may have affected this physical location on the flowcell (M. Cox, et al., 2010). Never-the-less, despite these issues 3,139,526 reads of high quality were obtained for the present study.

de Novo Assembly

The kmer length of 47 was regarded as optimal (see Table 3, pp42) because it gave the longest contig. Although the N50 was higher for the kmer length 49, the maximum contig length was higher for k = 47. The coverage cutoff variable has considerable effect on the maximum contig length (see Appendix, Table 3) values for maximum contig length varying over a 20,000 base range. The optimal coverage cutoff was 25 in that it gave the longest contigs. The expected coverage was 150.

Five super contigs were constructed for these optimal parameters, the maximum length of a contig being 56,034 bases. When aligned to the closest available relative, there was significant mismatch resulting in some gaps with the largest gap being 3,739 bases at the junction of the IRA/SSC region. Notwithstanding these gaps, this initial alignment helped to confirm the order of the supercontigs and identify regions where Velvet had failed to assemble the initial contigs. These results were the backbone in the construction of the *A.bifurcatum* genome.

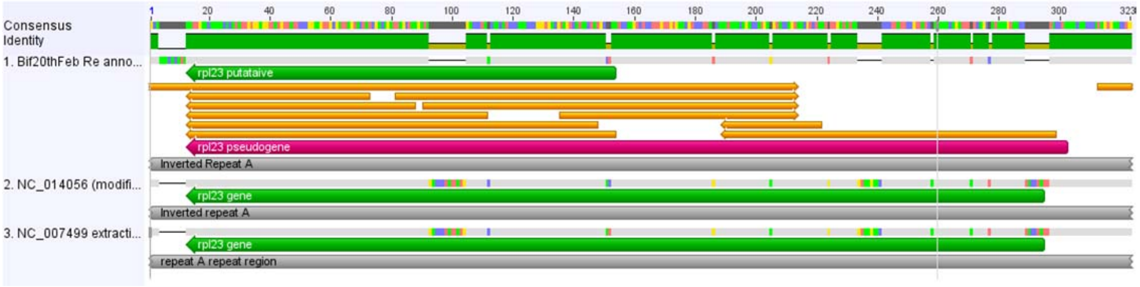
A major criticism of short sequence assemblers has been their inability to span nucleotide repeats longer than the read length, which results in an “assembly gap” (Chaisson, Brinza, & Pevzner, 2009; Pop & Salzberg, 2008; Shendure & Ji, 2008).This was the case in one of the gaps as illustrated by the upper highlight in Figure 8, pp40. This problem has been addressed by the introduction of paired end sequencing; a technique where two reads in the same sample are known to be separated by a DNA sequence of known length but unknown sequence. This positional data can help span difficult to assemble repeat regions. Unfortunately, this approach was not available for the initial sequencing of the *A.bifurcatum* genome.

Tablet was used to visualise the reads that aligned to the putative genome once Bowtie had done the aligning. Summary statistics from this alignment indicate that of the 4,709,998 raw reads that are in the read file from Illumina, only 26.86% of these successfully aligned with the reference genome. The major factor contributing to this low statistic is suspected to be mitochondrial and nuclear DNA present in the sample prep. Tablet and Bowtie proved to be excellent tools for visualising gaps that the Velvet assembled contigs failed to span. For three out of the four cases short range PCR products sequenced by Sanger technology provided the missing sequences.

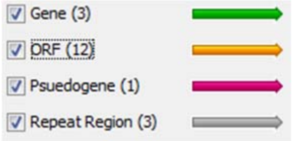
DOGMA was used to annotate the genome. Genes with low sequence identity were manually annotated after confirming both the start and stop codons using the open reading frame tool in Geneious, utilizing the chloroplast/bacterial genetic code. One of the genes (rpl23) had a stop codon in the middle of the gene and so was annotated by using the nearest open reading frame that agreed with the final stop codon proposed by Dogma. Further checks included -

- alignment of the gene to the chloroplast genomes of closest relatives, discussed below
- using the translation function in Geneious to check for alternative codon starts
- realigning all reads against the rpl23 gene to check for sequencing errors

This is illustrated in Figure 12 below. The gaps in the Consensus sequence indicate where there are disagreements between the organisms. The open reading frames show where there are start and stop codons by the length and direction of the arrow. The stop codon appears in the codon before the start of the putative *A.bifurcatum* rpl23 gene.



Key-



☑ Gene (3) 🟩

☑ ORF (12) 🟧

☑ Psuedogene (1) 🟥

☑ Repeat Region (3) ⬜

Figure 12: Shows alignment of the rpl23 genes of *A. bifurcatum*, NC 0007499 *Phalaenopsis aphrodite subsp. formosana* and NC 014056 *Oncidium Gower Ramsey*.
Open reading frames for *A.bifurcatum* and the potential pseudogene are also shown.

All genes identified as core genes, in that they are present in all chloroplast genomes from photosynthetic organisms, are present in the *A.bifurcatum* genome (Barbrook, Howe, Kurniawan, & Tarr, 2010). These include genes coding for RNA Polymerase (rpo), ribosomal proteins (rpl & rps), ycf4, Rubisco (rbcL), ribosomal RNA (rRNA) and all of the genes required for photosynthesis- photosystem I (psa), photosystem II (psb), cytochrome *b6f* complex (pet) and ATP synthase (atp).

Chloroplast Comparisons

Most angiosperm chloroplast genomes range from 120-160 kb in length and consist of four distinct portions- a LSC of 80-90 kb, a SSC of 16-27 kb and two IR regions of 20-28 kb. In most instances the gene content is conserved, encoding 4 rRNAs, 30 tRNAs, and 80 unique proteins (Yang et al., 2010). In recent times, many rearrangements of this basic pattern have been discovered including huge IR expansions, rearrangements (Chumley et al., 2006) and frequent gene losses (Chang et al., 2006; Hansen et al., 2007; Mardanov et al., 2008).

The most similar available chloroplast genomes to *A.bifurcatum* are listed in Table 5 below. All of these genomes appear in results from BLAST searches of varying portions of the *A.bifurcatum* genome. “Max score” represents the alignment score between the *A.bifurcatum* chloroplast genome and the other organisms. It is calculated by BLAST from the sum of the match rewards and the mismatch, gap open and extend penalties independently for each alignment.

Table 5: Whole chloroplast sequences most closely related to that obtained from *A.bifurcatum* as discovered through the NCBI database.

Organism	Genbank ID	Max Score	Class	Order	Length (bp)
Arthropodium bifurcatum	n.a.	n.a.	Monocots	Asparagales	146,417
Phoenix dactylifera	GU811709	1.79E+04	Monocots	Arecales	158,462
Typha latifolia	GU195652	1.71E+04	Monocots	Poales	161,572
Liriodendron tulipifera	DQ899947	1.62E+04	Magnoliids	Magnoliales	159,886
Oncidium Gower Ramsey	GQ324949	1.57E+04	Monocots	Asparagales	146,484
Phalaenopsis aphrodite	AY916449	1.56E+04	Monocots	Asparagales	148,964
Drimys granadensis	DQ887676	1.56E+04	Magnoliids	Canellales	160,604
Oryza sativa	AY522329	1.12E+04	Monocots	Poales	134,494

In a comparison of *A.bifurcatum, Phoenix dactylifera* and *Liriodendron tulipifera* chloroplast genomes, chosen because they consistently appeared in the top ranks of gene searches using BLAST, there is a large segment of approximately 4,100 bases located at the juncture of IRA and SSC that is not in the *A.bifurcatum* chloroplast genome but is in the other two chloroplast genomes. This IRA/SSC region proved very problematic in the multiple attempts made to PCR amplify it. This unsecured region has been indicated by a chain of 7N's used to represent this unfilled gap. Three colleagues are also finding this portion of the chloroplast genome problematic to sequence in their respective projects (T. McLenachen, O.Deusch , I. Ahmed, pers comm.). Whether this region does in fact occur in the chloroplast genome of *A.bifurcatum* remains unclear. Most likely it does not; a region of DNA 4,000bp long would be sequenced by the Illumina GAIIx.

In the search of Genbank for whole chloroplast genomes two genome sequences of similar length were found- *Phalaenopsis aphrodite subsp. formosana* and *Oncidium Gower Ramsey,* both orchids from the same order as *A.bifurcatum* namely Asparagales. A Geneious alignment (Appendix- Figure 1) of these two chloroplast sequences with *A.bifurcatum* highlights some interesting features. The gene order was found to be largely conserved across most of the genome in this comparison. Some exceptions are discussed below.

Interesting Features

In *A.bifurcatum* the ycf1 gene spans the SSC/IRB junction with the 5' end located in the repeat region by some 730bp. There is considerable variation in the position of the ycf1 gene in many angiosperms, with its location ranging from being completely within the SSC region as in *Phalaenopsis aphrodite,* to having between 156 bp as in *Nymphaea alba* and 1,583 bp as in *Amborella trichopoda*, in the repeat region *(Raubeson et al., 2007).* In *O.Gower Ramsey* this gene was not annotated with no reason given for this omission (Wu et al., 2010).

The LSC/IRA and LSC/IRB junctions in *A.bifurcatum* conform to results found for Asparagales in a study using comparisons of the location of this junction to distinguish basal angiosperms (R.-J. Wang et al., 2008). Plant chloroplast genomes from the Asparagales order have their LSC/IR junction either between genes rps19 and rpl22. However in a number of instances the rpl22 gene is largely contained within the IRA region.

The SSC region of the chloroplast genome, where most of the ndh genes are located, was where most differences occurred between the chloroplast genome of *A.bifurcatum* and other members of the Asparagales. The chloroplast NADH (ndh) dehydrogenase complex is a homologue of the mitochondrial complex 1, which has greater than 15 subunits, with 11 of these subunit homologues (ndhA-K) encoded by the chloroplast genome of most land plants: notable exceptions being pines and some parasitic plants (Yukawa & Sugiura, 2008). Many of the ndh genes in this alignment of the orchids and *A.bifurcatum* are of varying lengths, orientations and presence (Table 6).

Table 6: Shows the presence/absence and orientation of dehydrogenase subunit genes in *Phalaenopsis aphrodite subsp. formosana*, *Oncidium Gower Ramsey* and *A.bifurcatum.* Positive (+ve) and negative (-ve) symbols refer to the orientation of genes in the genome; (2) indicates that the gene contains an intron; the length of each gene is given in base pairs.

Gene	Location	*A.bifurcatum*	*O. Gower Ramsey*	*P.aphrodite*
ndhA	SSC	1095(2) +ve	550 +ve	absent
ndhB	IR	3075(2)+ve	2457 +ve	2291 +ve
ndhC	LSC	374 -ve	373 +ve	373 -ve
ndhD	SSC	1152 -ve	1140 +ve	1503 -ve
ndhE	SSC	308 -ve	308 -ve	156 -ve
ndhF		absent	absent	absent
ndhG	SSC	539 -ve	584 -ve	absent
ndhH	SSC	1182 -ve	326 -ve	absent
ndhI	SSC	549 -ve	275 -ve	514 -ve
ndhJ	LSC	479 -ve	550 -ve	479 -ve
ndhK	LSC	812 -ve	absent	2724 -ve

The variable presence and orientation of ndh genes among the three genomes is possibly the result of redundancy in the coding potential of ndh genes across chloroplast, nuclear and mitochondrial genomes. Previously it has been shown that all ndh genes in *O.Gower Ramsey* , except ndhE, lack apparent function, and are characterised by deletions and other mutations (Wu, et al., 2010). The ndh genes of *P.aphrodite* are either absent, truncated, dispersedly deleted or frame shifted and are considered non- functional (Chang, et al., 2006). These findings in orchids may also explain substitutions and indels in *A.bifurcatum* ndh genes which would presumably disrupt their function.

In analyses of the *A.bifurcatum* genome, rps16, accD and clpP genes were not identified by DOGMA. In the alignment with the orchids, deletions are apparent in *A.bifurcatum* where these genes are expected. Pairwise tBLASTn comparisons of proteins for rps16, accD and clpP against the *A.bifurcatum* genome did not recover sequences with significant e-values.

Another feature in annotating the *A.bifurcatum* chloroplast was the inability of DOGMA to identify small 5' exons (<12bp) separated from the rest of their gene by the presence of an intron. This limitation, observed in the present study, has been noted previously (O.Deusch, *pers comm*.). Here the problem occurred with rpl16 and petB which have starting exons of nine and six bp respectively. Additionally, the middle exon (26bp) of rps12 gene was not identified by DOGMA. In all instances, the sequence and location of these small exons were identical between the orchids and *A.bifurcatum*.

DOGMA identified open reading frames corresponding to ycf15 and ycf68 genes. The validity of ycf15 and ycf68 as protein coding genes has been questioned previously. However, they have been identified in a wide range of angiosperms including eudicots (*Ranunculus macranthus*), monocots (*Zea mays*) and Amborellales (*Amborella trichopoda*) (Raubeson, et al., 2007).

SNP's

Table 5 (Appendix) lists the 68 putative SNPs found in this study. There are zones within the single copy regions where there are more SNP's than at others. However, on average there are approximately seven SNP's for every 10,000 bases in the single copy regions and an overall average of 4.85 SNPs per 10,000 bases. This compares favourably with the rate of five SNPs per 10,000 bases, calculated from studies of differing accessions of maize and wheat (Muse, 2000). Only two SNPs (or one pair) were found in the inverted repeat region; this SNP was discovered in the ycf1 gene that spans the SSC/IRB junction.

Of the 68 SNPs that were found among the 18 different accessions, 20 were transitions and 48 were transversions (see Figure 13, below). It is interesting that the majority of the SNPs are transversions as this type of mutation changes the chemical structure of the DNA having more severe consequences than transitions. Usually transitions are more common than transversions. As these SNPs remain invalidated, the high proportion of transversions remains unsubstantiated.

14 of the SNP's discovered were found in simple sequence repeats. 25 SNP's were found in the coding regions of genes, five in their introns; this caused amino acid changes in 18 of these genes, seven were in the third codon position and elicited no change (see Appendix, table 7).

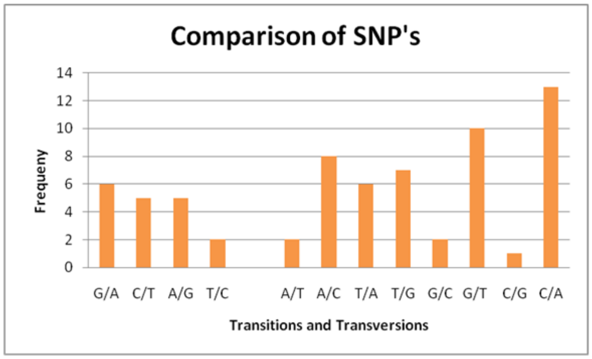


Figure 13: Comparison of SNP transitions and transversions in 18 accessions of *Arthropodium* chloroplast genomes.

It is well established that introns and spacers from the SSC regions of the chloroplast can provide high resolution markers for evolutionary studies. In a recent review of chloroplast mutational dynamics (Borsch & Quandt, 2009), the authors reviewed a range of mutational hotspots commonly used in phylogenetic analyses. Table 7 shows the location of SNPs inferred in the present study in relation to these hotspot regions.

Table 7: Location of *Arthropodium* SNPs in putative hot spot regionsBorsch & Quandt, 2009

Region Name	Position	SNP
trnT–trnF	44,038-45,660	0
trnS–G	6,528-8,207	1
atpB–rbcL	51,121-54860	2
psbA–trnG	1,138-8,172	10
Introns		
petD	no intron	n/a
rpl16	76,947-77,818	0
rps16	pseudogene 5133-6350	3
trnK	1,362- 4,031	2

The single copy regions adjacent to the inverted repeat regions indicate a relatively high occurrence of SNP's (Table 8). The average of seven SNPs (mentioned at the beginning of this section) is exceeded in all these four hotspot regions when 10,000 bases (from the start of the junction into the respective single copy region) are counted but this value drops when 5,000 bases are counted but only in the SSC region (data highlighted in purple). In this sense the *A. bifurcatum* genome SNPs conform to this hotspot trend around the junctions.

Table 8: The number of SNP's in 5,000 &10,000 bp zones into the single copy regions from the 4 junction regions.

Region	5000 bases	10000 bases
IRB-LSC	6	12
LSC-IRA	9	12
IRA-SSC	2	11
SSC-IRB	3	12

Some potential problems have been identified with using NGS for SNP discovery. First among these challenges is the need to be able to handle the enormous amounts of data that these technologies produce, shifting the emphasis from data acquisition to bioinformatic analysis. Secondly is the problem of ascertainment bias or identification of uninformative SNPs. Variation discovered between a few individuals cannot be presumed to represent all of the variation across the range of that species (Garvin, et al., 2010).

Another problem involved in SNP discovery is that of the trade off between the number of individuals sampled in a NGS run and the depth of coverage needed to ascertain SNPs. Indexed samples are a cost effective way of increasing the number of individuals able to be sampled per NGS run but the above trade off applies. Pooled samples, such as the mix of long PCR products used in this study, may be a way around this but representation by samples of a populations true variation requires careful consideration.

Further issues concern coverage of the pooled long range PCR products. Figure 14 shows that when reads were mapped back to the reference genome, there was a 100x range in the depth of coverage. The significance of this is that it may impact on the reliability of SNP discovery.

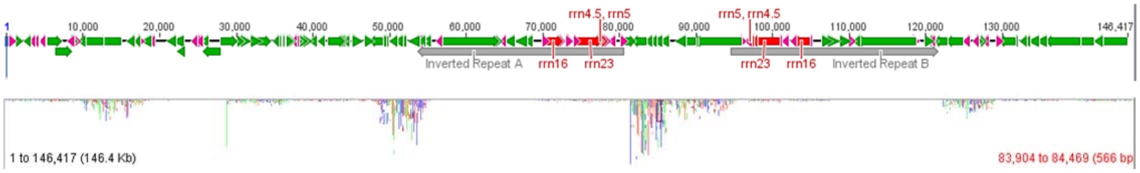


Figure 14: Relative coverage of mapped reads to the *A.bifurcatum* genome visualized in Tablet.

Illumina Sequencing

Illumina sequencing is one of the major second generation platforms for DNA sequencing. High throughput technologies have revolutionised the way we view the blue print for all life and further technological advancements mean that the thousand dollar genome and even the hundred dollar genome are achievable in the not too distant future. Studies of population history and the inference of relatedness will then be greatly advanced, such as is the hope in this study.

Using Illumina sequencing to de novo assemble a chloroplast genome and then find markers to distinguish different accessions using this genome has been an approach that is becoming more widespread. Slowly there are more peer reviewed articles appearing that have used a similar approach on a variety of platforms using differing approaches to SNP discovery (see Table 9). Most of these studies used transcriptome analysis to identify SNP's.

Table 9: Studies using high throughput sequencing for SNP discovery.

Adapted from (Ekblom & Galindo, 2010).

Organism	Platform	Author
Glanville fritillary butterfly (*Melitaea cinxia*)	454	(Wheat, 2010)
Gum tree (*Eucalyptus grandis)*	454	(Novaes et al., 2008)
Lake sturgeon (*Acipenser fulvescens*)	454	(Hale, McCormick, Jackson, & DeWoody, 2009)
Flesh fly (*Sarcophaga crassipalpis*)	454	(Hahn, Ragland, Shoemaker, & Denlinger, 2009)
Coral (*Acropora millepora*)	454	(Meyer et al., 2009)
Sugarcane (*Saccharum officinarum*)	454	(Bundock et al., 2009)
Apple Maggot (*Rhagoletis pomonella*)	454	(Schwarz et al., 2009)
Giant Panda (*Ailurapoda melanoleura*)	Solexa	(Li et al., 2010)
Stickleback (*Gasterosteus aculeatus*)	Solexa	(Hohenlohe et al., 2010)
Whitefish (*Coregonus* spp.)	Solexa	(Renaut, Nolte, & Bernatchez, 2010)
Moscow salsify (*Tragopogon miscellus*)	Solexa/454	(Buggs et al., 2010)
Great tit (Parus major)	Solexa	(van Bers et al., 2010)

Several problems have arisen in the application of new sequencing technologies that produce short read sequences. Prime amongst them is the reliance on a close reference sequence for contig assembly. This problem can be alleviated by paired end sequencing as previously mentioned, but ultimately can be resolved by the increase in the length of sequence reads. In this respect, PacBio's 'ZMW technology' promises multi kilo base read lengths (Zhou et al., 2010).

Sequencing errors have a similar signature to low frequency SNP alleles which can mislead SNP identification. False positives may also occur from alignment error. However, if a putative SNP is present in multiple overlapping reads then true SNPs should be more easily identified. Post sequence quality assurance tools, such as SolexaQA as used in this study, can also be used to estimate and minimize the impact of sequencing error rate. While a conservative approach was used in the present study – only reads greater than 65 bases in length and with a high quality threshold (p=0.05) were used to describe putative SNPs – SNP calling error is yet to be evaluated.

There is a growing trend towards using pooled samples for high throughput sequencing and SNP discovery because of cost effectiveness (Everett, Grau, & Seeb, 2011; Futschik & Schlotterer, 2010; Seeb et al., 2011; Timmermans et al., 2011). The pooled data from the present study was generated from one third of an Illumina flowcell lane. This produced more than three million reads on a GAIIx, and with current sequencing protocols the number of reads per lane now exceeds 30 million. The low sequencing error rate of the Illumina GAIIx makes this approach a potentially very valuable one (Futschik & Schlotterer, 2010).

Future Steps

The screening of the 129 samples, extracted to date, with primers designed from the denser polymorphic regions detected during this study provide a basis for ongoing studies needed to test translocation hypotheses of New Zealand *Arthropodium* species. However, validation of the putative SNPs identified in the present work is an important next step before this can be done. This will involve PCR and Sanger sequencing from the accessions collection.

Completion of the *A.bifurcatum* genome is also desirable. To achieve this, cloning of PCR products covering the gap region may be necessary. The sequence of the *A.cirratum* genome would be a further check on both the unsequenced region in *A.bifurcatum* and add weight to the location of the putative SNP's.

The possible origins of *Arthropodium* in New Zealand could be tested in the future with molecular data and molecular clock analyses. Molecular dating would give an indication of the divergence time of the NZ species of *Arthropodium* from its Australian relatives. The branching orders would also be informative about whether the NZ taxa are more closely related to those in New Caledonia, Madagascar or Australia.

Bibliography

Anderson, A. (1991). The chronology of colonization in New Zealand. *Antiquity, 65*, 767-795.

Anderson, A. J. (2000). Implications of prehistoric obsidian transfer in south Polynesia. *Bulletin of the Indo-Pacific Prehistory Association, 20*, 117-123.

Aronsson, H., & Jarvis, P. (2002). A simple method for isolating import-competent *Arabidopsis* chloroplasts. *FEBS Letters, 529*(2-3), 215-220.

Atherton, R., McComish, B., Shepherd, L., Berry, L., Albert, N., & Lockhart, P. (2010). Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods, 6*(1), 22.

Aughton, P. (2004). *Resolution: The story of Captain Cook's second voyage of discovery*. London: Weidenfield & Nicolson.

Barbrook, A. C., Howe, C. J., Kurniawan, D. P., & Tarr, S. J. (2010). Organization and expression of organellar genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1541), 785-797.

Biggs, B. (Ed.) (1997) New Zealand Historical Atlas- Ko Papatuanuku e Takoto Nei. Auckland: Batemans.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology, 13*(11), 3261-3273.

Borsch, T., & Quandt, D. (2009). Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant systematics and evolution = Entwicklungsgeschichte und Systematik der Pflanzen, 282*(3-4), 169-199.

Brittan, N. H. (1987). *Arthropodium*. In A. S. George (Ed.), *Flora of Australia* (Vol. 45). Canberra: Australian Government Publishing Service.

Buckley, T. R., James, S., Allwood, J., Bartlam, S., Howitt, R., & Prada, D. (2011). Phylogenetic analysis of New Zealand earthworms (Oligochaeta: Megascolecidae) reveals ancient clades and cryptic taxonomic diversity. *Molecular Phylogenetics and Evolution, 58*(1), 85-96.

Buggs, R. J. A., Chamala, S., Wu, W. E. I., Gao, L. U., May, G. D., Schnable, P. S., et al. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology, 19*, 132-146.

Bundock, P. C., Eliott, F. G., Ablett, G., Benson, A. D., Casu, R. E., Aitken, K. S., et al. (2009). Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal, 7*(4), 347-354.

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research, 19*, 336-346.

Chang, C.-C., Lin, H.-C., Lin, I.-P., Chow, T.-Y., Chen, H.-H., Chen, W.-H., et al. (2006). The Chloroplast Genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. *Molecular Biology and Evolution, 23*(2), 279-291.

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The Complete Chloroplast Genome Sequence of *Pelargonium × hortorum*: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants. *Molecular Biology and Evolution, 23*(11), 2175-2190.

Clarke, A. C., Burtenshaw, M. K., McLenachan, P. A., Erickson, D. L., & Penny, D. (2006). Reconstructing the Origins and Dispersal of the Polynesian Bottle Gourd (*Lagenaria siceraria*). *Molecular Biology and Evolution, 23*(5), 893-900.

Clarkson, B. D. (Ed.). (1991). *Coastal plants inland*. Rotorua, N.Z.: Forest Research Institute.

Colenso, W. (1881). On the vegetable food of the ancient New Zealanders. *Transactions and Proceedings of the New Zealand Institute., 13*, 3-38.

Colenso, W. (1891). *Vestiges, Reminiscences, Memorabilia Of Works, Deeds, And Sayings Of The Ancient Maoris.* Paper presented at the Hawke's Bay Philosophical Institute, Hawkes Bay.

Cox, M., Peterson, D., & Biggs, P. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*(1), 485.

Cox, P. A., & Banack, S. A. (1991). *Islands, Plants and Polynesians. An Introduction to Polynesian Ethnobotany*. Portland, Oregon: Dioscorides Press.

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America, 99*(8), 5261-5266.

Deschamps, S., & Campbell, M. (2010). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding, 25*(4), 553-570.

Donnon, M. J. (2009). *Molecular Systematics of the Lomandra Labill. Complex (Asparagales : Laxmanniaceae).* University of Adelaide, Adelaide.

Doyle, J. J., & Doyle, J. D. (1990). Isolation of plant DNA from fresh tissue *Focus, 12*, 13-15.

Drummond, A. J., Ashon, B., Cheung, M., Heled, J., Kearse, M., Moir, R., et al. (2009). Geneious v4.7. Auckland: Biomatters Ltd.

Ekblom, R., & Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity, Advance on-line publication*, 1-15.

Everett, M. V., Grau, E. D., & Seeb, J. E. (2011). Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources, 11*, 93-108.

Fale, F. (1988). *Experimental tropical crops gardening in New Zealand*. Paper presented at the Contributions to an International Workshop on Ethnobotany.

Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C. H., et al. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources, 9*, 130-139.

Finney, B. (Ed.). (2006). *Ocean Sailing Canoes*. Auckland: Batemans.

Futschik, A., & Schlotterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics, 186*(1), 207-218.

Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources, 10*(6), 915-934.

Goodman, S. M., & History., F. M. o. N. (1996). *A Floral and faunal inventory of the eastern slopes of the Réserve naturelle intégrale d'Andringitra, Madagascar : with reference to elevational variation / Steven M. Goodman, editor* (Vol. n.s. no.85(1996)). [Chicago, Ill.] :: Field Museum of Natural History.

Gott, B. (2008). Indigenous use of plants in south-eastern Australia. *Telopea, 12*(2), 215.

Haami, B. (2004). *Putea whakairo: Maori and the written word.* Wellington: Huia Publishers in association with the Ministry for Culture and Heritage.

Hahn, D., Ragland, G., Shoemaker, D., & Denlinger, D. (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics, 10*(1), 234.

Hale, M., McCormick, C., Jackson, J., & DeWoody, J. A. (2009). Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics, 10*(1), 203.

Hamilton, A. M., Zug, G. R., & Austin, C. C. (2010). Biogeographic anomaly or human introduction: a cryptogenic population of tree skink (*Reptilia: Squamata*) from the Cook Islands, Oceania. *Biological Journal of the Linnean Society, 100*(2), 318-328.

Hannon, G. L. (2010). FASTX-Toolkit. from http://hannonlab.cshl.edu/fastx_toolkit/

Hansen, D. R., Dastidar, S. G., Cai, Z., Penaflor, C., Kuehl, J. V., Boore, J. L., et al. (2007). Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution, 45*(2), 547-563.

Harris, G. F., & TeWhaiti, H. (1996). Rengarenga lillies and Maori occupation at Matakitaki-a-Kupe (Cape Palliser). *Journal of the Polynesian Society, 105*(3), 271-286.

Heenan, P. B., Mitchell, A. D., & De Lange, P. J. (2004). *Arthropodium bifurcatum* (Asparagaceae), a new species from northern New Zealand. *New Zealand Journal of Botany, 42*, 233-246.

Herzer, R. H., Chaproniere, G. C. H., Edwards, A. R., Hollis, C. J., Pelletier, B., Raine, J. I., et al. (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.

Higham, T., Anderson, A., & Jacomb, C. (1999). Dating the first New Zealanders: the chronology of Wairau Bar. *Antiquity, 73*(280), 420(428).

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genet, 6*(2), e1000862.

Imelfort, M., & Edwards, D. (2009). De novo sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics, 10*(6), 609-618.

Institute., N. Z. (1868) Transactions and proceedings of the New Zealand Institute. *Vol. v.28 (1895)* (pp. Page 577). Wellington :: New Zealand Institute.

Kårehed, J., Groeninckx, I., Dessein, S., Motley, T. J., & Bremer, B. (2008). The phylogenetic utility of chloroplast and nuclear DNA markers and the phylogeny of the *Rubiaceae* tribe *Spermacoceae*. *Molecular Phylogenetics and Evolution, 49*(3), 843-866.

Knapp, M., Mudaliar, R., Havell, D., Wagstaff, S. J., & Lockhart, P. J. (2007). The Drowning of New Zealand and the Problem of *Agathis*. *Systematic Biology, 56*(5), 862-870.

Knox, M. A., Hicks, B. J., Banks, J. C., & Hogg, I. D. (2008). *Fish biosurveillance by genetic methods: a feasibility study*. Hamilton, New Zealand: Centre for Biodiversity and Ecology Research, Department of Biological Sciences, School of Science and Engineering, The University of Waikato. .

Ladiges, P. Y., & Cantrill, D. (2007). New Caledonia–Australian connections: biogeographic patterns and geology. *Australian Systematic Botany, 20*, 383-389.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology 10*(R25).

Leach, H., & Stowe, C. (2005). Oceanic arboriculture at the margins - the case of the karaka (*Corynocarpus laevigatus*) in Aotearoa. *Journal of the Polynesian Society, 114*(1), 7-28.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature, 463*(7279), 311-317.

Lloyd, D. G. (1985). Progress in understanding the natural history of New Zealand plants. *New Zealand Journal of Botany, 23*, 707-722.

Lockhart, P. J., McLenachan, P. A., Havell, D., Glenny, D., Huson, D., & Jensen, U. (2001). Phylogeny, Radiation, and Transoceanic Dispersal of New Zealand Alpine Buttercups: Molecular Evidence under Split Decomposition. *Annals of the Missouri Botanical Garden, 88*(3), 458-477.

London., L. S. o. (1920) The Journal of the Linnean Society. *Vol. v.45 (1920-1922)*. London: Williams and Norgate.

Mardanov, A., Ravin, N., Kuznetsov, B., Samigullin, T., Antonov, A., Kolganova, T., et al. (2008). Complete Sequence of the Duckweed (*Lemna minor*) Chloroplast Genome: Structural Organization and Phylogenetic Relationships to Other Angiosperms. *Journal of Molecular Evolution, 66*(6), 555-564.

McLenachan, P. A., Stöckler, K., Winkworth, R. C., McBreen, K., Zauner, S., & Lockhart, P. J. (2000). Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. *Molecular Ecology, 9*(11), 1899-1903.

Meudt, H. M., & Clarke, A. C. (2007). Almost Forgotten or Latest Practice? AFLP applications, analyses and advances. *Trends in Plant Science, 12*(3), 106-117.

Meyer, E., Aglyamova, G., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J., et al. (2009). Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics, 10*(1), 219.

Mildenhall, D. C. (2001). Pollen analysis of Pliocene-Pleistocene Kowai Formation (Kurow Group), Mackenzie Basin, South Canterbury, New Zealand. *New Zealand Journal of Geology & Geophysics, 44*, 97-104.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., et al. (2010). Tablet—next generation sequence assembly visualization. *Bioinformatics, 26*(3), 401-402.

Muse, S. V. (2000). Examining rates and patterns of nucleotide substitution in plants. *Plant Molecular Biology, 42*(1), 25-43.

Novaes, E., Drost, D., Farmerie, W., Pappas, G., Grattapaglia, D., Sederoff, R., et al. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics, 9*(1), 312.

Palmer, J. (1988). *New Zealand adventive plants as medicinal herbs*. Paper presented at the Contributions to an International Workshop on Ethnobotany.

Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics, 24*(3), 142-149.

Queiroz, A. d. (2005). The resurrection of oceanic dispersal in historical biogeography. *Trends in Ecology & Evolution, 20*(2), 68-73.

Raoul, E. (1846). *Choix de plantes de la Nouvelle- Zelande* (Vol. T11). Paris: Fortin, Masson et Cie.

Raubeson, L., Peery, R., Chumley, T., Dziubek, C., Fourcade, H. M., Boore, J., et al. (2007). Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics, 8*(1), 174.

Renaut, S., Nolte, A. W., & Bernatchez, L. (2010). Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus spp*. Salmonidae). *Molecular Ecology, 19*, 115-131.

Richards, E. C. (1949). *Our New Zealand Trees and Flowers*. Christchurch: Simpson & Williams Ltd.

Riley, M. (1994). *Maori Healing and Herbal*. Paraparaumu, N.Z.: Viking Sevenseas N.Z. Ltd.

Rokas, A., & Abbot, P. (2009). Harnessing genomics for evolutionary insights. *Trends in Ecology & Evolution, 24*(4), 192-200.

Rozen, S., & Skaletsky, H. (1999). Primer3 on the WWW for General Users and for Biologist Programmers. In S. Misener & S. A. Krawetz (Eds.), *Bioinformatics Methods and Protocols* (Vol. 132, pp. 365-386): Humana Press.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74*(12), 5463-5467.

Schwarz, D., Robertson, H., Feder, J., Varala, K., Hudson, M., Ragland, G., et al. (2009). Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics, 10*(1), 633.

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., & Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources, 11*, 1-8.

Seifert, K. A. (2009). Progress towards DNA barcoding of fungi. *Molecular Ecology Resources, 9*, 83-89.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotech, 26*(10), 1135-1145.

Sheppard, P. J. (1993). *Lapita lithics: trade/exchange and technology. A view from the Reefs/Santa Cruz* (Vol. 28). Sydney: University of Sydney.

Sturm, T. (Ed.). (1991). *The Oxford history of New Zealand literature in English* Auckland, N.Z.: Oxford University Press.

Tatusova, T. A., & Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters, 174*(2), 247-250.

The Angiosperm Phylogeny, G. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society, 161*(2), 105-121.

Thorsen, M. J., Dickinson, K. J. M., & Seddon, P. J. (2009). Seed dispersal systems in the New Zealand flora. *Perspectives in Plant Ecology, Evolution and Systematics, 11*(4), 285-309.

Timmermans, M. J. T. N., Dodsworth, S., Culverwell, C. L., Bocak, L., Ahrens, D., Littlewood, D. T. J., et al. (2011). Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*.

Tregear, E. (1904). *The Maori Race*. Wanganui Archibald Dudingston Willis

Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A.-P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research, 36*(4), e25.

van Bers, N. E. M. V., Oers, K. V., Kerstens, H. H. D., Dibbits, B. W., Crooijmans, R. P. M. A., Visser, M. E., et al. (2010). Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology, 19*, 89-99.

Walter, R., Jacomb, C., & Bowron-Muth, S. (2010). Colonisation, mobility and exchange in New Zealand prehistory. *Antiquity, 84*(324), 497(417).

Wang, J., Pan, Y.-Z., Gong, X., Chiang, Y.-C., & Kuroda, C. (2011). Chloroplast DNA variation and phylogeography of *Ligularia tongolensis* (Asteraceae), a species endemic to the Hengduan Mountains Region of China. *Journal of Systematics and Evolution*, no-no.

Wang, R.-J., Cheng, C.-L., Chang, C.-C., Wu, C.-L., Su, T.-M., & Chaw, S.-M. (2008). Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology, 8*(1), 36.

Wheat, C. (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica, 138*(4), 433-451.

Winkworth, R. C., Wagstaff, S. J., Glenny, D., & Lockhart, P. J. (2002). Plant dispersal N.E.W.S from New Zealand. *Trends in Ecology & Evolution, 17*(11), 514-520.

Wu, F.-H., Chan, M.-T., Liao, D.-C., Hsu, C.-T., Lee, Y.-W., Daniell, H., et al. (2010). Complete chloroplast genome of *Oncidium Gower Ramsey* and evaluation of molecular markers for identification and breeding in *Oncidiinae*. *BMC Plant Biology, 10*(1), 68.

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics, 20*(17), 3252-3255.

Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., et al. (2010). The Complete Chloroplast Genome Sequence of Date Palm *Phoenix dactylifera PLoS ONE, 5*(9), e12762.

Yukawa, M., & Sugiura, M. (2008). Termination codon-dependent translation of partially overlapping ndhC-ndhK transcripts in chloroplasts. *Proceedings of the National Academy of Sciences, 105*(49), 19550-19554.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research, 18*(5), 821-829.

Zhan, Q.-Q., Wang, J.-F., Gong, X., & Peng, H. (2011). Patterns of chloroplast DNA variation in *Cycas debaoensis* (Cycadaceae): conservation implications. *Conservation Genetics*, 1-12.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology: A technology review and future perspective. *SCIENCE CHINA Life Sciences, 53*(1), 44-57.

Appendix

Table 1 – Location of Accessions

ID	No.	Locality Details	Latitude	Longitude
A.cir	5277	Whangaroa Harbour	35 02 04.770 S	173 43 20.79 E
A.bif	1019	Massey University	40 23 19.040 S	175 37 07.55 E
A.cir	1053	Landcare, Massey Uni, Palm Nth	40 23 9.23 S	175 37 03.82 E
A.bif	1054	Marsden, Massey Uni, Palm Nth	40 23 6.50 S	175 37 02.81 E
A.cir	1055	Te Henga	36 53 26.439 S	174 26 37.997 E
A.cir	1056	Te Henga	36 53 26.439 S	174 26 37.997 E
A.cir	1057	Te Henga	36 53 26.439 S	174 26 37.997 E
A.cir	1058	Te Henga	36 53 26.439 S	174 26 37.997 E
A.cir	1061	Tauwhare Pa SR, BOP	37 58 53.269 S	177 04 09.610 E
A.cir	1062	Tauwhare Pa SR, BOP	37 58 52.654 S	177 04 09.575 E
A.cir	1063	Tauwhare Pa SR, BOP	37 58 52.654 S	177 04 09.575 E
A.bif	1065	Hen Island	35 57 54.850 S	174 43 06.580 E
A.bif	1066	Three Kings Islands	34 9 34.300 S	172 08 9.610 E
A.bif	1074	Poor Knights Islands	35 27 39.410 S	174 44 17.110 E
A.bif	1075	Surville Cliffs	34 24 10.770 S	173 00 5.740 E
A.bif	1078	Poor Knights	35 27 39.410 S	174 44 17.11 E
A.cir	1101	Ocean Bay, Port Underwood	41 19 50.128 S	174 06 04.294 E
A.cir	1102	Ocean Bay, Port Underwood	41 19 50.297 S	174 06 03.523 E
A.cir	1103	Ocean Bay, Port Underwood	41 19 49.720 S	174 06 02.868 E
A.cir	1104	Ocean Bay, Port Underwood	41 19 49.720 S	174 06 02.868 E
A.cir	1105	Ocean Bay, Port Underwood	41 19 48.892 S	174 06 04.704 E
A.cir	1106	Ocean Bay, Port Underwood	41 19 48.892 S	174 06 04.704 E
A.cir	1116	Abel Tasman Track, Wainui Bay	40 48 14.530 S	172 57 12.376 E
A.cir	1117	Abel Tasman Track, Wainui Bay	40 48 14.530 S	172 57 12.376 E
A.cir	1118	Abel Tasman Track, Wainui Bay	40 48 12.682 S	172 57 12.036 E
A.cir	1128	Hanson Winter Scenic Reserve	40 49 54.057 S	172 53 19.834 E
A.cir	1129	Hanson Winter Scenic Reserve	40 49 54.155 S	172 53 20.474 E
A.cir	1132	Hanson Winter Scenic Reserve	40 49 53.991 S	172 53 18.511 E
A.cir	1133	Hanson Winter Scenic Reserve	40 49 54.056 S	172 53 18.510 E
A.cand	1148	Paynes Ford Scenic Reserve	40 53 07.980 S	172 48 44.778 E
A.cand	1149	Paynes Ford Scenic Reserve	40 53 07.980 S	172 48 44.778 E
A.cir	1161	Marlborough Sounds ex. Arnold Dench	41 00 00.00 S	173 59 60.000 E
A.cir	1208	Whangaruru North Scenic Reserve	35 22 44.859 S	173 47 06.500 E
A.cir	1214	Mahinepua Peninsula Scenic Reserve	34 59 46.373 S	173 46 48.783 E
A.cir	1215	Mahinepua Peninsula Scenic Reserve	34 59 48.086 S	173 43 25.806 E
A.bif	1216	Tauranga Bay (nth end)	35 00 29.486 S	173 43 26.308 E
A.bif	1217	Tauranga Bay (nth end)	35 00 29.486 S	173 43 26.308 E
A.cir	1228	Ranfurly Bay Scenic Reserve	35 00 54.597 S	173 43 11.033 E
A.cir	1229	Ranfurly Bay Scenic Reserve	35 00 54.597 S	173 43 11.033 E
A.cir	1230	Ranfurly Bay Scenic Reserve	35 00 54.646 S	173 43 05.039 E
A.cir	1231	Ranfurly Bay Scenic Reserve	35 00 51.369 S	173 43 04.999 E
A.cir	1232	Ranfurly Bay Scenic Reserve	35 00 51.369 S	173 43 04.999 E
A.cir	1233	Ranfurly Bay Scenic Reserve	35 00 21.516 S	174 29 51.620 E

A.cir	1234	Ranfurly Bay Scenic Reserve	35 00 21.520 S	174 29 51.344 E
A.cir	1235	Ranfurly Bay Scenic Reserve	35 00 17.858 S	174 33 58.384 E
A.cir	1244	Otito Bay Scenic Reserve	35 33 30.486 S	174 34 43.294 E
A.cir	1245	Otito Bay Scenic Reserve	35 33 30.713 S	174 34 43.299 E
A.bif	1248	Bream Tail	36 03 41.780 S	174 23 05.097 E
A.bif	1249	Bream Tail	36 04 02.949 S	173 51 50.580 E
A.bif	1250	Bream Tail	36 04 02.949 S	173 51 50.580 E
A.cir	1264	Awhitu, Hamilton Road	37 06 38.400 S	174 33 54.910 E
A.cir	1265	Maioro, SW of Maioro Rd	37 20 00.730 S	174 41 13.860 E
A.cir	1304	Bream Head, Whangarei	35 50 57.082 S	174 34 56.355 E
A.cir	1305	Bream Head, Whangarei	35 50 57.082 S	174 34 56.355 E
A.cir	1306	Bream Head, Whangarei	35 50 57.065 S	174 34 48.902 E
A.cir	1329	Whangape Ridge	35 21 48.145 S	173 13 08.102 E
A.cir	1330	Whangape Ridge	35 21 45.743 S	173 13 08.603 E
A.cir	1331	Whangape Ridge	35 21 45.743 S	173 13 08.603 E
A.cir	1332	Whangape Ridge	35 21 45.743 S	173 13 08.603 E
A.cir	1340	Ahipara	35 11 05.525 S	173 07 21.503 E
A.cir	1341	Ahipara	35 11 05.525 S	173 07 21.503 E
A.cir	1342	Ahipara	35 11 05.525 S	173 07 21.503 E
A.cir	1343	Ahipara	35 11 05.525 S	173 07 21.503 E
A.cir	1357	Whatawhiwhi	34 52 51.467 S	173 23 57.911 E
A.cir	1358	Whatawhiwhi	34 52 51.467 S	173 23 57.911 E
A.cir	1359	Whatawhiwhi	34 52 32.281 S	173 23 01.054 E
A.cir	1360	Whatawhiwhi	34 52 32.281 S	173 23 01.054 E
A.bif	1370	Great Island, Three Kings	34 9 17.17 S	172 8 28.19 E
A.bif	1371	Great Island, Three Kings	34 9 17.17 S	172 8 28.19 E
A.bif	1372	Great Island, Three Kings	34 9 17.17 S	172 8 28.19 E
A.bif	1373	Great Island, Three Kings	34 9 17.17 S	172 8 28.19 E
A.bif	1374	Great Island, Three Kings	34 9 17.17 S	172 8 28.19 E
A.cir	1375	Awhitu, Hamilton Rd	37 8 47.30 S	174 36 09.58 E
A.cir	1376	Maioro, SW of Maioro Rd	37 20 08.57 S	174 41 13.38 E
A.cir	1377	Massey University, outside Landcare	40 23 09.23 S	175 37 03.82 E
A.bif	1378	Massey University, outside Marsden	40 23 06.50 S	175 37 02.81 E
A.cir	1397	Miramar Peninsula	41 17 42.348 S	174 49 33.852 E
A.cir	1398	Miramar Peninsula	41 17 42.348 S	174 49 33.852 E
A.cir	1411	Tokerau Beach	34 52 32.281 S	173 23 01.054 E
A.cir	1705	South of Sandy Bay, Coromandel	36 31 34.871 S	175 27 45.699 E
A.cir	1706	South of Sandy Bay, Coromandel	36 31 34.871 S	175 27 45.699 E
A.cir	1726	Aotea Cliffs, Shelly Bay	36 34 22.35 S	174 22 37.21 E
A.cir	1727	Waihoka, Otakawhe Bay	36 50 54.934 S	175 08 19.542 E
A.cir	1728	Maunganui Bluff	35 46 1.86 S	173 34 24.17 E
A.cir	1729	Hauturu Island	37 12 56.22 S	175 53 29.15 E
A.cir	1730	Waitamata Harbour	36 49 29.77 S	174 42 18.07 E
A.cir	1731	Rapatiotio Pt, Waihi Beach	36 49 29.77 S	174 42 18.07 E
A.cir	1732	Maungaraho	37 23 37.50 S	175 56 20.90 E
A.cir	1739	Nihotipu	36 57 01.87 S	174 33 52.80 E
A.cir	1740	Chathams	43 57 8.65 S	176 33 50.67 W
A.cir	1741	Nihotipu	36 57 01.87 S	174 33 52.80 E
A.cir	1742	Nihotipu	36 57 01.87 S	174 33 52.80 E
A.cir	1825	Whanarua Bay	37 40 57.2 S	177 47 20.1 E
A.cir	1826	Haparapara River	37 47 57.4 S	177 40 07.6 E
A.cir	1827	Onepoto	37 35 53.1 S	178 17 58.8 E

A.cir	1828	Otiki Hill	37 41 22.0 S	178 32 53.6 E
A.cir	1829	Hicks Bay	37 34 09.8 S	178 17 19.8 E
A.cir	1830	Motu	37 52 07.5 S	177 36 51.1 E
A.cir	1831	Tohora Pirau	37 33 08.9 S	178 09 53.8 E
A.cir	1844	Rotoehu	38 01 49.993 S	176 30 57.067 E
A.cir	1845	Rotoehu	38 01 49.993 S	176 30 57.067 E
A.cir	1846	Rotoehu	38 01 49.993 S	176 30 57.067 E
A.cir	1847	Tryphena point, GBI	36 19 08.800 S	175 28 47.200 E
A.cir	1848	Needle Rocks, GBI	36 14 27.000 S	175 29 07.000 E
A.cir	1849	Memory Is GBI	36 15 57.400 S	175 29 37.500 E
A.cir	1850	Raupuke Pt, GBI	36 07 3.900 S	175 21 52.900 E
A.cir	1851	Okataina	38 08 23.383 S	176 23 30.966 E
A.cir	1852	Titoki point, Lake Okataina	38 08 12.342 S	176 24 07.876 E
A.cir	1853	Titoki point, Lake Okataina	38 08 12.342 S	176 24 07.876 E
A.cir	1854	Titoki point, Lake Okataina	38 08 12.342 S	176 24 07.876 E
A.cir	1863	Waikareka Rd, Limestone Downs	37 27 31.502 S	174 44 35.757 E
A.cir	1864	Waikareka Rd, Limestone Downs	37 27 31.914 S	174 44 36.337 E
A.cir	1865	Limestone Downs	37 27 31.310 S	174 44 33.595 E
A.cir	1866	Limestone Downs	37 27 32.299 S	174 44 34.637 E
A.cir	1989	Kairakau	39 56 34.71 S	176 55 31.25 E
A.cir	1993	Kairakau	39 56 34.71 S	176 55 31.25 E
A.cir	1994	Kairakau	39 56 34.71 S	176 55 31.25 E
A.cir	2009	Orpheus Bay	37 00 35.455 S	174 34 41.206 E
A.cir	2010	Kapiti Island	40 51 19.903 S	174 55 52.585 E
A.cir	2043	Wainuiomata River Mouth	41 24 45.001 S	174 53 30.001 E
A.cir	2044	Wainuiomata River Mouth	41 24 44.100 S	174 53 42.301 E

Key

A.cir – *Arthropodium cirratum*
A.bif – *Arthropodium bifurcatum*
A.cand- *Arthropodium candidium*

Table 2 – Long-range Primers

Primer name	Sequence
LSC	
rpl2- psbI (F)	CCATGGAGGCGGGGAAGGGA
rpl2-psbI(R)	TGCGGCCGGGGTCGTTAGAT
psbI-rpoC2 (F)	ATCTAACGACCCCGGCCGCA
psbI-rpoC2 (R)	TTGCGAGCGGAACGAGCAGG
rpoC2-rpoB (F)	TGGAATCCACCTCTACGGTCCCA
rpoC2-rpoB (R)	TGCTCCGGAATGGAAATGAGGGA
rpoB-rps14 (F)	TCCCTCATTTCCATTCCGGAGCA
rpoB-rps14 (R)	ACCGGGCGCAACAAGATCCA
rps14-rps4 (F)	TGGATCTTGTTGCGCCCGGT
rps14-rps4 (R)	CGCCGTCTGGGGGCTTTACC
rps4-rbcl (F)	CGGTAAAGCCCCCAGACGGC
rps4-rbcl (R)	GCCCCTGCTTCTTCAGCGGG
rpcL-petG (F)	CCCGCTGAAGAAGCAGGGGC
rpcL-petG (R)	GGTCCAACTGATCACCACGTCTGT
petG-petD (F)	ACAGACGTGGTGATCAGTTGGACC
petG-petD (R)	AGCTACTGGACGGCGAAATGGA
petD-rpl2 (F)	CGCCGTCCAGTAGCTACAACAGT
petD-rpl2 (R)	TCCCCATGGAGGCGGGGAAG
SSC	
rps12-ndhF (F)	TGCGATATCTCACACCGGGCA
rps12-ndhF (R)	ACAAACGGGGTCGGCCTTGC
ndhF-ndhI (F)	TGCCGCAATCGGTCGTGTGA
ndhF-ndhI (R)	CGTGTATGTCCCATAGATCTACCCGT
ndhI-ycf1 (F)	TGGATTCGCCCGCGGAAACG
ndhI-ycf1 (R)	GGAATACGTTTGGCTTGGGGGAGG
ycf1-rps12 (F)	CGCAAGTGTTGCGCTTGGCA
ycf1-rps12 (R)	ACACCGGGCAAATCCTGAACCC

Table 3 – Maximising Coverage Cutoff

Coverage Cutoff for Kmer=47

Cov Cutoff	Nodes	N50	Max Contig Length	Total
1	1035	5103	23699	657981
2	959	5351	23699	657384
3	803	6224	23699	652308
4	645	6732	23699	640948
5	541	7571	23699	632321
6	503	8054	23699	627737
7	449	8054	23699	624029
8	434	8054	23699	622546
9	416	8054	23699	620440
10	392	8054	23699	618228
11	380	8054	23699	616704
12	365	8054	23699	614990
13	346	8120	23699	613327
14	338	8120	23699	612237
15	339	8120	23699	612202
16	325	8131	23699	600141
17	319	8131	23699	599935
18	311	8361	23699	590418
19	298	8412	23913	562937
20	286	8523	23913	547717
21	271	8531	37470	530722
22	254	8531	37470	521218
23	246	8531	37470	520241
24	229	9307	37470	515149
25	210	10569	37519	489564
26	207	9998	37519	485418
27	195	9307	29362	462784
28	186	10648	35586	433104
29	185	10648	35586	431682
30	181	9998	29362	415864
31	168	11725	29362	394781
32	153	12386	29362	369567
33	149	12386	29362	343633
34	135	11725	29362	320155
35	122	11725	29362	314703
36	113	10648	29362	308729
37	103	10648	29362	304645
38	99	9653	25707	291789
39	96	9209	21660	265640
40	93	9209	19193	262513
41	87	9209	19193	262130
42	78	8544	19193	250786
43	77	8131	19193	233730
44	76	8125	19193	222705
45	71	7870	17937	200774
46	68	7559	20693	200584
47	59	7870	20693	178291
48	56	7870	20693	178239
49	55	7870	20693	177287
50	52	10671	20693	173260
51	47	7870	21697	155824
52	47	7870	21697	155774
53	46	7870	21697	155660
54	44	10671	21697	151040
55	44	10671	21697	151040
56	42	10671	21697	146318
57	42	10671	21697	144479
58	44	10671	21697	144561
59	43	10671	21697	143789
60	42	10671	21697	136230
61	41	10635	21697	133658
62	41	10635	21697	133658
63	40	9627	21697	124029
64	41	9627	21697	123352
65	39	10330	25337	120835
66	40	10330	25337	120791
67	39	9627	25337	116465
68	37	9627	25337	116151
69	37	9627	25337	115614
70	36	9627	25337	115476
71	37	9627	25337	115420
72	36	9627	25337	115165
73	36	9627	25337	115165
74	35	9627	25337	114501
75	35	9627	25337	114501
76	32	9627	25337	114083
77	31	9627	25337	113988
78	31	9627	25337	111860

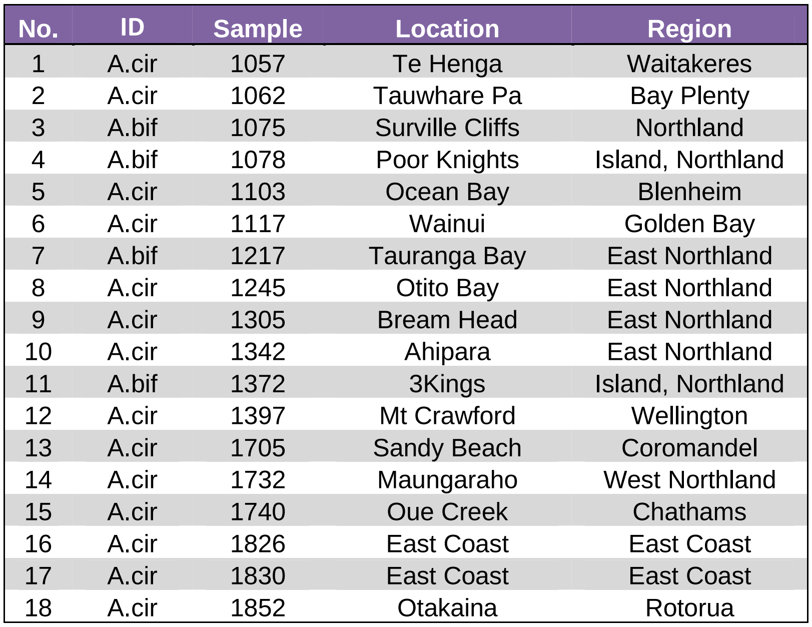
79	30	9627	25337	109706
80	30	9627	25337	109706
81	31	7406	21886	109668
82	31	7406	21886	109668
83	29	9627	21886	106297
84	30	8637	21886	103314
85	32	6565	18318	103200
86	31	6565	18318	101522
87	31	6565	18318	99095
88	30	6528	18318	98718
89	29	5526	18318	94493
90	29	5526	18318	94492
91	29	5526	18318	94492
92	29	5526	18318	94492
93	29	5526	18318	94492
94	29	5526	18318	94492
95	30	5526	18318	94219
96	31	5526	18318	94218
97	30	5526	18318	93356
98	30	5526	18318	93323
99	31	5526	18318	93222
100	31	5526	18318	93222

Coverage Cutoff for Kmer=49

Cov Cutoff	Nodes	N50	Max Contig Length	Total
1	743	5968	19155	642131
2	711	5968	19155	641765
3	641	6732	19155	638147
4	532	6805	19252	630746
5	479	7938	19252	626177
6	420	7938	21752	621413
7	400	8006	21752	620177
8	385	8006	21752	617247
9	367	8006	21752	613849
10	361	8006	21752	613159
11	345	8006	21752	611068
12	334	8014	21752	608596
13	330	8014	21752	605485
14	320	8072	21752	598318
15	310	8072	21752	592292
16	293	8204	21752	577634
17	275	8359	21752	546373
18	261	8359	21752	532715
19	238	8531	21752	523199
20	225	9309	21752	521777
21	204	10277	21655	499555
22	196	10277	21655	486308
23	179	10277	26477	456996
24	173	10486	26477	435724
25	170	10486	26477	429408
26	159	10486	26477	416546
27	143	11331	26477	380722
28	133	11331	26477	346300
29	112	12483	26477	320251
30	102	10650	26477	296160
31	94	10650	26477	291985
32	90	11727	26477	288345
33	86	11727	26477	275197
34	76	11727	26477	271578
35	76	10864	26477	262470
36	70	10486	26477	243145
37	70	9879	26477	241541
38	67	10332	26477	204662
39	60	10332	26477	199976
40	52	12483	26477	186580
41	51	12483	26477	185625
42	48	12483	26477	173698
43	41	13862	26477	162274
44	41	13862	26477	162274
45	40	13862	26477	157998
46	37	14405	26477	145168
47	37	14405	26477	143593
48	35	14405	26477	141425
49	35	14365	26477	141385
50	36	13862	26477	141345
51	35	13862	26477	133756
52	34	13862	26477	131304
53	36	10332	26477	122598
54	37	10332	26477	121917
55	37	10332	26477	121917
56	36	10332	26477	121799
57	35	10332	26477	117365
58	32	10332	27234	116835
59	32	10332	27234	116835
60	31	9625	25251	116535

61	30	10332	25251	116283
62	30	10332	25251	116283
63	30	10332	25251	116283
64	29	10332	25251	116174
65	26	10332	25251	114583
66	27	9625	25251	113217
67	25	9625	25251	111099
68	25	9625	25251	111059
69	26	9625	25251	111019
70	26	9625	25251	111019
71	27	9625	24497	108079
72	28	8639	21884	101658
73	28	6565	21884	99889
74	28	6565	21884	97462
75	26	6526	21884	94247
76	27	6526	21884	94246
77	27	6526	21884	94246
78	27	6526	21884	94275
79	27	6526	21884	94275
80	27	6526	21884	94275
81	27	6526	21884	94275
82	27	6526	21884	94073
83	27	6526	21884	94042
84	28	6526	21884	93939
85	28	6526	21884	93503
86	28	6526	21884	93466
87	27	6526	21884	93126
88	27	6012	21884	78230
89	28	6012	21884	78165
90	27	6012	21884	78007
91	27	6526	21884	73463
92	27	6526	21884	73463
93	27	6526	21884	73463
94	26	6012	21884	66107
95	25	6012	21884	64800
96	25	9625	21884	62358
97	25	9625	21884	62354
98	25	9625	21884	62354
99	24	9625	21884	56342
100	24	9625	21884	56213

Table 4 – List of accessions that were used in long range PCR



No.	ID	Sample	Location	Region
1	A.cir	1057	Te Henga	Waitakeres
2	A.cir	1062	Tauwhare Pa	Bay Plenty
3	A.bif	1075	Surville Cliffs	Northland
4	A.bif	1078	Poor Knights	Island, Northland
5	A.cir	1103	Ocean Bay	Blenheim
6	A.cir	1117	Wainui	Golden Bay
7	A.bif	1217	Tauranga Bay	East Northland
8	A.cir	1245	Otito Bay	East Northland
9	A.cir	1305	Bream Head	East Northland
10	A.cir	1342	Ahipara	East Northland
11	A.bif	1372	3Kings	Island, Northland
12	A.cir	1397	Mt Crawford	Wellington
13	A.cir	1705	Sandy Beach	Coromandel
14	A.cir	1732	Maungaraho	West Northland
15	A.cir	1740	Oue Creek	Chathams
16	A.cir	1826	East Coast	East Coast
17	A.cir	1830	East Coast	East Coast
18	A.cir	1852	Otakaina	Rotorua

Key

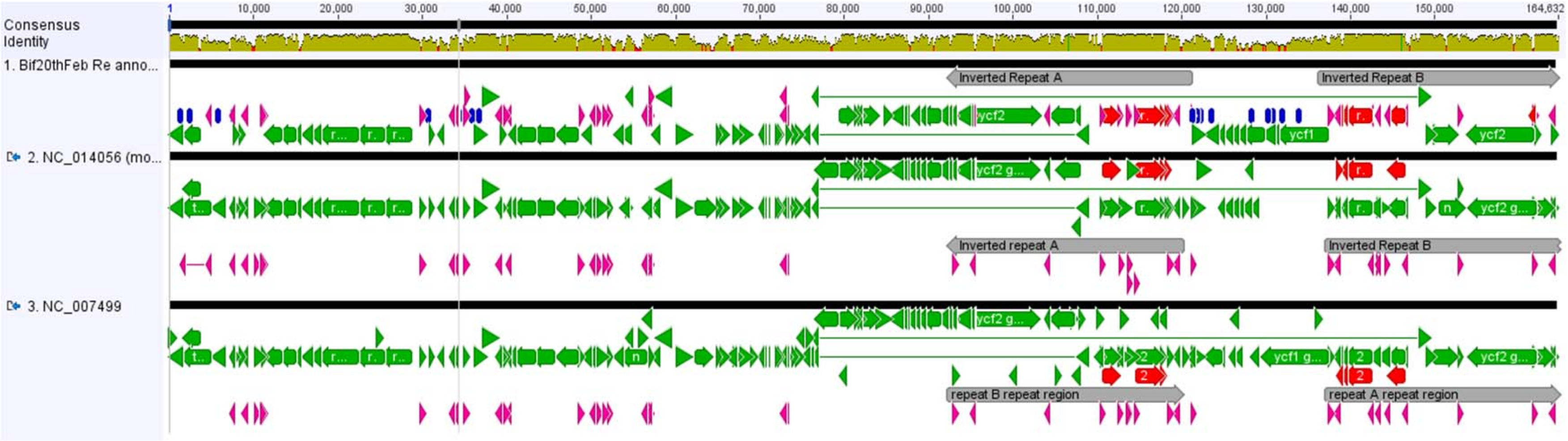
A.cir – *Arthropodium cirratum*

A.bif – *Arthropodium bifurcatum*

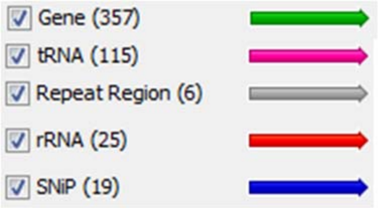
Table 5- SNP Positions

Position	Nucleotide in Reference	Nucleotide in Reads	SNP Depth	Total Depth	Percentage of SNPs in Total Reads
1158	C	A	20	109	18%
2217	G	T	26	80	33%
4239	G	T	684	2534	27%
4298	A	G	1008	2317	44%
4541	A	C	1092	2569	43%
5332	C	A	495	2326	21%
5424	A	C	51	89	57%
6021	C	A	344	401	86%
7671	A	C	39	81	48%
8273	A	T	37	67	55%
9383	T	C	19	86	22%
11654	G	T	70	135	52%
12352	G	A	32	70	46%
13180	A	C	57	105	54%
17167	G	A	51	64	80%
20852	G	C	61	79	77%
20968	A	C	149	157	95%
25535	G	T	98	123	80%
26470	G	C	222	267	83%
26696	A	G	156	357	44%
29095	T	G	269	314	86%
29287	C	A	183	222	82%
30321	G	T	170	239	71%
30364	T	A	190	266	71%
30414	T	G	252	320	79%
31200	C	T	177	204	87%
40114	T	A	599	3746	16%
42351	C	A	421	1680	25%
53848	A	C	120	253	47%
54636	T	A	324	463	70%
58003	A	G	183	832	22%
59742	G	T	472	919	51%
59802	G	T	464	573	81%
59823	A	C	707	1088	65%
61679	T	A	373	482	77%
65361	T	A	46	114	40%
68181	C	T	20	167	12%
70046	C	A	37	195	19%
70464	C	A	41	210	20%
73076	T	G	3752	5608	67%

74425	C	G	1219	6096	20%
74448	A	T	2355	7067	33%
74592	T	G	4509	7174	63%
75251	G	A	4109	8139	50%
75297	C	A	2374	8616	28%
76125	G	T	4101	6003	68%
76584	G	A	2444	8693	28%
76623	G	T	3360	6970	48%
78279	C	A	3436	6349	54%
94961	C	T	43	49	88%
104737	A	C	115	132	87%
106402	C	A	3297	14206	23%
106520	C	A	2144	11043	19%
110659	C	A	1277	2273	56%
111151	G	A	666	735	91%
112091	C	A	2179	2868	76%
113034	A	G	1111	3556	31%
114023	C	A	771	2258	34%
114102	T	C	2217	2578	86%
114484	A	G	788	2804	28%
114642	T	C	242	2281	11%
114713	T	G	2445	2964	82%
115777	G	T	2330	2565	91%
116064	C	T	539	1342	40%
117657	A	C	992	1157	86%
120712	T	G	121	139	87%
125484	G	A	97	186	52%



Key



☑ Gene (357) 🟩

☑ tRNA (115) 🟪

☑ Repeat Region (6) ⬜

☑ rRNA (25) 🟥

☑ SNiP (19) 🟦

Figure 1- Alignment of *A. bifurcatum*, NC 0007499 *Phalaenopsis aphrodite* subsp. *formosana* and NC 014056 *Oncidium Gower Ramsey*.

Table 7- SNP’s in coding regions of genes

Gene	SNP	Change in amino acid
matK	G/A	H-N
atpA	T/C	I-V
rpoC2	G/A	A-V
psbT	C/T	none, 3rd codon
rbcL	A/C	Y-S
	T/A	L-I
cemA	A/G	I-V
psbB	C/T	none, 3rd codon
petB	C/A	F-L
rpoA	T/G	Q-P
rpl14	T/G	T-P
rpl16	G/A	none, 3rd codon
	G/T	none, 3rd codon
rps3	C/A	M-I
ccsA	C/A	F-L
	C/A	none, 3rd codon
ndhI	G/A	H-Y
ndhA	A/G	none, 3rd codon
ndhH	C/A	G-A
	T/C	R-G
	A/G	none, 3rd codon
	T/C	T-A
ycf1	G/T	S-I
	C/T	none, 3rd codon
	A/C	F-L