THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Genomic characterisation of Australian wild rice species

Marta Katarzyna Brozynska

MSc of Biotechnology

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

Queensland Alliance for Agriculture and Food Innovation

Abstract

Rice (*Oryza sativa* L.) is one of the most important crops in the world and a basic food source for about two-thirds of the population. It has been grown in the tropics and sub-tropics probably for thousands of years. Owing to the fast growing world population, there is a high demand for increased and improved rice production. Wild relatives of cultivated rice, the related A genome species, carry a vast genetic diversity and are a significant resource for rice improvement. Recent advances in both molecular breeding programmes and next-generation sequencing have potential for application in genetic improvement of elite rice lines with valuable traits from wild rice species. Numerous genes have already been transferred from wild species into selected rice lines and new varieties have been released for commercial use. Very recently two distinct groups of wild rice have been discovered in northern Queensland. They could be new Australian wild rice species and represent novel gene reservoirs in rice. These populations have remained isolated from the great genetic impact of the domesticated rice populations from Asia. The characterisation, collection and conservation of this wild genetic resource may be critical for global food security. The focus of this study was, firstly, the characterisation of the novel Australian wild rice relatives in terms of their relationship with other *Oryza* species, including the cultivars, using full chloroplast sequence as well as nuclear genome regions; secondly, creating genomic resources of these wild rice relatives by means of whole genome sequencing and generating reference genomes for the two novel Australian wild rice taxa. The draft genomes generated were used for whole genome characterization studies including phylogenetic reconstruction with high resolution, divergence time estimation, structural descriptions, gene annotation, and repeat content exploration. The evolutionary inference results revealed that northern Australia might be the centre of diversity of the A genome *Oryza* and highlighted the need for better understanding and exploitation of these populations.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Singh AK, Furtado A, Brozynska M, Mishra NS and Henry RJ (2016) Phylogeny and molecular evolution of miR820 and miR396 microRNA families in *Oryza* AA genomes. *Scientific Reports*. Accepted.

Brozynska M, Furtado A and Henry RJ (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal* **14**:1070-85.

Wambugu PW, Brozynska M, Furtado A, Waters DL and Henry RJ (2015) Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Scientific Reports* **5**:13957.

Brozynska M, Furtado A and Henry RJ (2014) Direct Chloroplast Sequencing: Comparison of Sequencing Platforms and Analysis Tools for Whole Chloroplast Barcoding. *PLoS ONE* **9**:e110387.

Brozynska M, Omar ES, Furtado A, Crayn D, Simon B, Ishikawa R and Henry RJ (2014) Chloroplast Genome of Novel Rice Germplasm Identified in Northern Australia. *Tropical Plant Biology* **7**:111-120.

Publications included in this thesis

Brozynska M, Furtado A and Henry RJ (2014) Direct Chloroplast Sequencing: Comparison of Sequencing Platforms and Analysis Tools for Whole Chloroplast Barcoding. *PLoS ONE* **9**:e110387 – incorporated as Chapter 3.

Contributor	Statement of contribution
Brozynska M (Candidate)	Designed experiments (40%) Performed the experiments (100%) Analysed the data (70%) Wrote and edited the paper (50%)
Furtado A	Designed experiments (30%) Analysed the data (30%)

	Wrote and edited the paper (20%)
Henry RJ	Designed experiments (30%) Wrote and edited the paper (30%)

Brozynska M, Omar ES, Furtado A, Crayn D, Simon B, Ishikawa R and Henry RJ (2014) Chloroplast Genome of Novel Rice Germplasm Identified in Northern Australia. *Tropical Plant Biology* **7**:111-120 – incorporated as Chapter 4.

Contributor	Statement of contribution
Brozynska M (Candidate)	Designed experiments (40%) Performed the experiments (100%) Analysed the data (70%) Supported field collections (30%) Wrote and edited the paper (50%)
Omar ES	Analysed the data (10%)
Furtado A	Designed experiments (30%) Analysed the data (20%) Wrote and edited the paper (10%)
Cryan D	Supported field collections (10%) Wrote and edited the paper (5%)
Simon B	Supported field collections (20%) Wrote and edited the paper (5%)
Ishikawa R	Supported field collections (40%) Wrote and edited the paper (5%)
Henry RJ	Designed experiments (30%) Wrote and edited the paper (25%)

Contributions by others to the thesis

Dr Ryuji Ishikawa from The Hirosaki University, Hirosaki, Aomori, Japan facilitated wild rice collections in northern Queensland and provided information on collection sites.

The sequencing data for *Oryza sativa* ssp. *japonica* var. Nipponbare was provided by Dr Catherine Nock from Southern Cross Plant Science, Southern Cross University, Lismore, NSW, Australia.

The local support team from Life Technologies provided technical assistance with Ion Torrent sequencing data.

The set of putatively single copy orthologs (6,015 genes) used in the phylogenetic inference between Australian wild rice taxa and other A genome *Oryza* species was provided by Prof. Rod Wing from Arizona Genomics Institute (AGI), The University of Arizona, Tucson, AZ, USA. This set was generated as a part of a study under the international initiative I-OMAP – International *Oryza* Map Alignment Project.

Dr Dario Copetti from AGI contributed towards genome annotation of Australian wild rice taxa.

Prof. Robert Henry critically revised this thesis as a whole prior to submission.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

This project was jointly supported by the Department of Agriculture and Fisheries, the University of Queensland, and Rural Industries Research and Development Corporation (RIRDC), Australia. I sincerely thank the University of Queensland for the financial support through University of Queensland International Scholarship (UQI) that covered the cost of my tuition fees and UQ Research Scholarship that provided me with the living allowance stipend. I would like thank the UQ Graduate School for a Graduate School International Travel Award (GSITA) for assistance in collaborative research in USA and QAAFI RHD Travel Award for allowing a participation in International Plant & Animal Genome XXIII Conference (San Diego, CA, USA). I wish to acknowledge the International Rice Research Institute for awarding me with a Young Rice Scientist Award to present an oral paper at the 4th International Rice Congress (Bangkok, Thailand).

I wish to thank my principal advisor: Prof. Robert Henry for his valuable contribution to my PhD research, his guidance and invaluable advice. I also want to thank my associate advisors: Dr Agnelo Furtado and Dr Daniel Waters for their support and assistance.

I am grateful to Dr Ryuji Ishikawa (Hirosaki University), Dr Katunori Ichitani (Kagoshima University) and Masahiro Sotowa (Hirosaki University) for their help during the field trip in tropical Queensland and providing valuable information on the novel Australian wild rice populations. I am thankful to the local support team from Life Technologies for technical assistance with Ion Torrent sequencing data analysis and interpretation. I sincerely thank the staff at the Arizona Genomics Institute (University of Arizona, Tucson, AZ, USA), including: Prof. Rod Wing and Dr Dario Copetti, for the opportunity to visit the institute and analyse a part of my research data.

Keywords

wild rice, *Oryza*, crop wild relatives, chloroplast sequence, nuclear sequence, phylogeny, genome sequencing, genome assembly, genome annotation, genomics

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408, Genomics, 60%

ANZSRC code: 060309, Phylogeny and Comparative Analysis, 20%

ANZSRC code: 070305, Crop and Pasture Improvement (Selection and Breeding), 20%

Fields of Research (FoR) Classification

FoR code: 0604, Genetics, 30%

FoR code: 0607, Plant Biology, 30%

FoR code: 0703, Crop and Pasture Production, 40%

Table of Contents

List of Figures

List of Tables

List of Abbreviations used in the thesis

ABC	Advanced backcross
AFLP	Amplified fragment length polymorphism
BI	Bayesian inference
BUSCO	Benchmarking Universal Single-Copy Orthologs
CA	Celera Assembler
CCD	Charge-coupled device
CEG	Core Eukaryotic Gene
CEGMA	Core Eukaryotic Genes Mapping Approach
CSSL	Chromosome segment substitution lines
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
GA	Genome Analyzer
GBS	Genotyping-by-sequencing
GO	Gene ontology
GPM	Genome Puzzle Master
GWAS	Genome-wide association studies
HPD	Highest Posterior Density
indels	Insertions and deletions
I-OMAP	The international *Oryza* map alignment project
IR	Inverted repeat
IRGSP	International rice genome sequencing project
ISSR	Inter-simple sequence repeat

KEGG	Kyoto Encyclopedia of Genes and Genomes
LSC	Large single copy
MAS	Marker-assisted selection
MCMC	Monte Carlo Markov Chain
ML	Maximum likelihood
miRNA	micro RNA
MNV	Multi-nucleotide variant
MP	Maximum parsimony
mya	Million years ago
NGS	Next-generation sequencing
NIL	Nearly isogenic line
NJ	Neighbour-joining
NNI	Nearest-neighbour interchange
OLC	Overlap-layout-consensus
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PE	Paired end reads
PGM	Personal Genome Machine
QTL	Quantitative trait locus
QUAST	Quality Assessment Tool for Genome Assemblies
RAPD	Random amplified polymorphic DNA
RLFP	Restriction fragment length polymorphism
RIL	Recombinant inbred lines

RNA	Ribonucleic acid
rRNA	ribosomal RNA
SBS	Sequencing-by-syntheses
SMRT	Single-molecule, real-time
snoRNA	small nucleolar RNA
SNP	Single nucleotide polymorphism
SPR	Subtree pruning and regrafting
sRNA	spliceosomal RNA
SSC	Small single copy
SSR	Simple sequence repeat
TBR	Tree bisection reconnection
TE	Transposable element
tRNA	transfer RNA
UPGMA	Unweighted pair group methods using arithmetic averages

Chapter 1

General Introduction

1.1 The importance of rice and the need for rice improvement

Rice (*Oryza sativa* L.) is one of the most important crops in the world and the basic food source for about two-thirds of the population. It has been grown in the tropics and sub-tropics for thousands of years. Almost 90% of overall rice consumption and production takes place in Asia (Rockwood & Chandler 2001), where 60% of the world's inhabitants live. The human populations in Asia rely on rice for up to 80% of their daily calorie requirement. Rice crops are important for the socioeconomic and political stability in many countries in Asia, Africa and Latin America, as well as being a substantial source of employment in those countries.

Research on rice is significant for two reasons: as a model for plant science, especially in genomic studies supporting research on other species, and because of its role in food security for a fast growing world population. It is estimated that by 2050 there will be an additional 2 billion people living on Earth. The dilemma that humans are facing nowadays is how to ensure there is enough food to feed an increasing world population. In answering this question rice plays a very important role.

There is an ongoing demand for increased rice production. On one hand, the growth in rice production required by the year 2030 is 25%, as it has been calculated that for every additional 1 billion people about 100 million additional tons of rice has to be produced (Global Rice Science Partnership 2010). On the other hand, this intensification has to occur using less land, water, labour and chemicals, i.e. pesticides, fungicides and herbicides, owing to the climate change and reduced surface of arable land (Abberton *et al.* 2016).

The challenges faced by rice growers currently include environmental stresses and adaptation to climate change. Rice production can be affected by pollution (air, water, heavy metals), soil degradation, global warming, decreasing availability of arable land, decreasing water availability, and fertilisation costs due to high-labour and high-energy requirements. The main stresses affecting plant production are both biotic (insect, bacterial, fungal and viral diseases) and abiotic (drought, salinity, floods). Furthermore, current cultivated varieties are considered to have reached almost the limits of their production (Food and Agriculture Organization 2016).

In order to overcome these difficulties and to produce enough rice to feed the growing global population, new and better rice varieties are required. These improved cultivars need to be better adapted, faster growing with greater yield, enriched nutritional grain quality and have reduced environmental impact. In addition, they have to be resistant to biotic and abiotic stresses. Agronomic potential of rice has been increased through biotechnological techniques and conventional breeding. The first requirement for creating these new rice varieties is a deep understanding of rice biology and plant development. Recent progresses in molecular biology technologies and development of new tools have led to advances in rice genomics and genetics. Another important step is to find and describe valuable genes that can be used in rice improvement. A thorough characterisation of rice is important not only for improvement of cultivated rice but also for better management of rice germplasm for future utilisation and conservation.

The other great importance of rice is its use as a model system for plant genomic studies. Owing to a small and compact genome compared to that of the major crops, rice has become a model organism facilitating research on other cereal grasses and understanding their biology, domestication and eventually, improvement. Rice is highly suitable for genomic studies due to its ease of transformation, colinearity of the genome with other cereal crops and rich genetic resources, including mutants, cultivated land races and wild rice relatives (Paterson *et al.* 2005; Shimamoto & Kyozuka 2002; Xu *et al.* 2005). Additionally, finished high-quality genome sequences and available genomic tools and resources increase the value of rice as a model system.

1.2 The importance of wild rice in food security

Wild relatives of cultivated rice are more genetically diverse than domesticated species and represent an important resource for rice improvement. In comparison to the commercial cultivars, wild rice is characterised by low grain yield, poor grain type and grain

shattering in natural habitats. Moreover, there is enormous diversity between wild rice species with respect to morphological characteristics (plant height, production of side shoots, clum, leaf, panicle, seed features), flowering time, distribution and growth habitat, adaptation to different environments and stresses, and agronomic traits. Wild rices have poorer yield than domesticated varieties, but despite this, wild rice species have genes with potential to provide a significant enhancement of agricultural productivity of cultivated rice (Sweeney & McCouch 2007). The desirable gene pool for rice improvement includes genes for biotic and abiotic stresses as well as yield-enhancing traits.

Numerous genes have been already transferred from wild species into elite rice lines and new varieties have been released for commercial use. The improvements in the novel varieties include resistance to grassy stunt, brown plant hopper, tungro virus, bacterial blight, blast, as well as tolerance to acid sulphate soils and salinity (Brar & Singh 2011). This has been carried out mostly in rice-growing countries in Asia. In African rice species new traits that have been introduced include high yield, earlier harvest time, competitive ability with weeds, and tolerance to abiotic stresses. Also there are two new varieties tested in the USA for heat tolerance (Sanchez *et al.*, personal communication). Little work has been done on transferring genes for tolerance to abiotic stresses.

Genetic diversity in wild rice is declining due to modern agriculture practices (Zhang & Wing (eds.) 2013). Some of the wild rice species are already extinct or close to extinction which has led to loss of biodiversity causing concerns for preservation of wild rice germplasm.

1.3 High throughput sequencing technology

DNA sequencing and genomic technologies have advanced considerably in the last decade. Nowadays, these technologies are used to create whole genomes, transcriptomes, exomes and even epigenomes in both model and non-model species. Currently sequencing has multiple applications in a vast number of fields including plant science. Sequencing methodologies have had an enormous impact on plant research revolutionising the study of plant evolution and diversity as well as accelerating crop breeding. Numerous molecular markers have been created [Single Nucleotide Polymorphism (SNP), Simple Sequence Repeat (SSR), etc.] that have led to the discovery and mapping of new quantitative trait loci (QTLs) and alleles [e.g. Su *et al.* (2015), Kaya *et al.* (2016), Hou *et al.* (2015)], and to advancement in marker-assisted selection (MAS) and breeding [e.g. Gramazio *et al.* (2016), Leite *et al.* (2016), Iwata *et al.* (2016)], genome-wide

association studies [GWAS; e.g. Marcotuli *et al.* (2016), Körber *et al.* (2016), Liu *et al.* (2016)] , and genotyping-by-sequencing [GBS; e.g. Bajgain *et al.* (2016), Boutet *et al.* (2016), Yu *et al.* (2016)]. Sequencing technologies also have been used to understand domestication events (Henry 2012), mechanisms of biological and molecular processes within the cell, and whole genome duplication events [e.g. Qin *et al.* (2014), Wang *et al.* (2014a), Wang *et al.* (2013)]. Genomics also allows for gene annotation, discovery of novel genes, studies of gene expression and biosynthetic pathways [e.g. Schnable *et al.* (2009), The Potato Genome Sequencing Consortium (2011), Brenchley *et al.* (2012)].

The first plant genome sequence of *Arabidopsis thaliana* was published in 2000 (Arabidopsis Genome Initiative 2000). It was generated using Sanger sequencing, which is a first generation sequencing method. The second plant genome sequenced using this technology was rice (Goff *et al.* 2002). The Sanger method uses dideoxy analogues of triphosphatates that inhibit DNA polymerase – the process is known as dideoxy chain termination (Sanger *et al.* 1977). First, the genome is randomly fragmented and subsequently each fragment has to be cloned into a vector (cosmid, bacterial artificial chromosome or yeast artificial chromosome) and every clone is sequenced separately. Automated platforms for Sanger sequencing were introduced in the early 1990s that utilised four-colour fluorescent dyes and capillary electrophoresis in contrast to previously used polyacrylamide gel electrophoresis. This method led the sequencing field for almost twenty years and delivered a number of completed genome sequences of human, animals, plants and microorganisms. Despite the improvements, the Sanger technology was very expensive, time-consuming, labour-intensive and using the technique, sequencing of large and complex genomes, e.g. human genome would have taken more than a decade to complete (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001). These limitations resulted in introduction of second- or next-generation sequencing (NGS) methods in the mid-2000s. They produce large volumes of data at cheaper per-base cost, in much shorter timeframes and with less labour investment. Next-generation methods allowed the broader scientific community greater access to sequencing and the launch of numerous large-scale studies. The most common high throughput sequencing methods include 454 (Roche), SOLiD (ThermoFisher Scientific) and Illumina. Currently, Illumina dominates the second generation sequencing market.

Each of the sequencing technologies can be divided into several stages of template preparation, sequencing and signal detection. The library construction involves the shotgun method that randomly splits the genomic DNA into small pieces and creates the

templates for sequencing (Fleischmann *et al.* 1995). The library originates from a single DNA molecule but requires clonal amplification and immobilisation prior to sequencing. Roche 454 and SOLiD use emulsion PCR and DNA capture onto beads. The templates in Roche 454 are placed into wells, and in SOLiD system they are linked to a glass surface. In Illumina the amplification (called solid-phase or bridge amplification) and immobilisation take place on a glass slide. Roche 454 uses so called pyrosequencing based on a series of enzymatic reactions initiated by a release of inorganic pyrophosphate when a single nucleotide is incorporated into the DNA library. As a consequence a visible light is released and measured by an optical detection system. In this platform each of the nucleotides (cytosine, guanine, adenine and thymine) is added one at the time, followed by a washing step and addition of the next specific nucleotide. SOLiD performs sequencing by ligation and in contrast to other technologies, uses DNA ligase instead of DNA polymerase. The number of hybridisation and ligation cycles with specific two-base-encoded probes enables each nucleotide in the template to be read twice, thereby improving accuracy. Every cycle is followed by four-colour imaging and chemical cleavage. Illumina sequencing involves synthesis and reversible terminator-based chemistry with fluorescently labelled nucleotides and modified DNA polymerase. The main features of this chemistry include incorporation of a single modified nucleotide and momentary termination of DNA synthesis, followed by four-colour fluorescent imaging. The cycle terminates with a cleavage of the terminating group and the fluorophore from the incorporated nucleotide.

In the early 2010s third-generation sequencing technologies emerged. Platforms included in this category are Ion Torrent (Thermo Fisher Scientific) and PacBio RS (Pacific Bioscience). Pacific Biosciences is known as a single-molecule, real-time (SMRT) sequencing technology (Eid *et al.* 2009). The proprietary instrument detects fluorescent light that comes from DNA polymerase incorporation of dye-labelled nucleotides onto a DNA template in real time. The library comes from a single DNA molecule and does not require PCR amplification, which differentiates this method from the second generation technologies. The template is polymerase-immobilised on a solid support called a zero-mode waveguide detector. On the other hand, Ion Torrent is based on semiconductors; instead of using light detection, an ion-sensitive field measures the pH change caused by hydrogen ion release upon nucleotide incorporation. The library is constructed by emulsion PCR, DNA capture onto proprietary sphere particles, after which synthesised fragments are loaded onto chip wells. This technology differs from the others by the absence of fluorescence or chemiluminescence for detection.

Newer sequencing methods have also been recently developed, i.e. nanopore sequencing. This technology involves forcing the DNA molecule through a nano-scale pore and measuring the resulting alteration in electrical current through the pore. The nanopores used in sequencing can be either protein pores suspended in lipid bilayers (Tobkes *et al.* 1985) or artificial nanopores made of solid state materials (Li *et al.* 2001), e.g. grapheme (Merchant *et al.* 2010) and titanium nitride (Luan *et al.* 2010). A commercially available nanopore sequencing instrument has been introduced by Oxford Nanopore Technologies (Clarke *et al.* 2009). It uses the protein pore and measures changes in conductivity across a protein pore as DNA passes through. The nature of the current change is used to detect the nucleotides in the template.

Besides the differences in underlying chemistry between the sequencing platforms, they also vary in quality of the data, read length, error type and rate, throughput, speed, cost and applications. The main drawback of second generation sequencing methods, compared to conventional Sanger sequencing, is the read length (string of nucleotides sequenced at a time). Sanger methods reach up to about 1 Kb in length, whereas read lengths for Illumina are about 100-150 bp and for SOLiD are approximately 60 bp. Roche 454 and Ion Torrent produce longer reads, up to 1 Kb and up to 400 bp, respectively; however, the read quality is poorer. Read length has been improved significantly with the introduction of PacBio and Oxford Nanopore with have average read lengths 10-15 Kb and 5 Kb in turn; nonetheless the quality is considerably lower. Sanger is still considered by the research community as a gold standard for providing very high accuracy data.

1.4 Genome assembly

Genome assembly is a process of reconstructing an original genome sequence from many small DNA fragments derived by a sequencing instrument. Two different types of genome assembly can be distinguished: *de novo* and reference-based assembly. The latter process consists of mapping sequencing reads to a reference genome and calling a consensus based on that genome. The reference genome comes from the same or closely related species. Reference-based assembly is mainly used in identifying and classifying genetic variations within the same species or genera and they are widely used in re-sequencing projects. A *de novo* process reconstructs the sequence without access to a reference; as a result this approach is extensively used for sequencing of non-model organisms.

The two main paradigms used in the *de novo* genome assembly are greedy and graph-based approaches (Miller *et al.* 2010). The greedy algorithm is one of the simplest strategies which involves merging together the sequencing reads based on their overlaps and their quality (Pop & Salzberg 2008). However, it represents only a locally optimum solution to the assembly problem and tends to collapse the repeats. It was widely used in the early assemblers, e.g. phrap, which was the main assembler in the Human Genome Project (International Human Genome Sequencing Consortium 2001).

Graph-based approaches represent the sequencing data as graph structures where the reads form nodes connected by edges (relationships between the reads). Some types of graphs used in genome assemblies are overlap-layout-consensus (OLC), de Bruijn and string graphs. OLC graph is the simplest model with reads represented as nodes and overlaps between them as edges (Kececioglu & Myers). In this strategy the assembly follows three main steps: overlapping of the reads and forming initial continuous sequences (contigs), layout, i.e. ordering and orienteering the contigs, and consensus generation by multiple alignments. It is suitable for long reads but inappropriate for large amounts of short read data. The approach was used in another human genome sequencing project (Venter *et al.* 2001). Another type of graph-based genome assembly is the de Bruijn graph (Pevzner *et al.* 2001). Briefly, each sequencing read is split into k-mers of a specific length prior to graph creation. The k-mers are extracted from reads and represented as nodes; the adjacent k-mers are also extracted and represented as edges. The solution to the assembly problem in this case consists of walking through the graph and finding a path that visits every edge only once, which is called the Eulerian path. This is the most widely used approach in the genome assembly of short read data and has been applied in numerous genome assemblies. The main downside of the de Bruijn graph approach is that the repeats are also collapsed in the final assembly. String graphs are another type of graph, which are conceptually very similar to the de Bruijn graph but it is not necessary to break the reads into k-mers (Myers 2005).

De novo assembly of a genome remains very challenging. Firstly, because of technological issues associated with the sequencing platforms, i.e. errors in the reads, underrepresentation of long homopolymer stretches of DNA and or GC-/AT-rich regions in the raw data; secondly, because of computational issues, i.e. processing of the data, available software and algorithms, computational resources and access to them; thirdly, due to the nature of the organism sequence itself, i.e. size and complexity of the genome.

Plant genomes are particularly challenging to sequence. Many plant genomes are much larger than animals; they have high ploidy levels, extensive repetitive sequence content, duplications (whole genome, segments and tandems) and significant heterozygosity rates. Also, they exhibit the presence of large gene families, high numbers of genes and pseudogenes that arose as a result of whole genome duplications and transposon movements (Schnable *et al.* 2009). All these characteristics negatively affect the assembly process making it more troublesome, especially with regard to repeat numbers and lengths (Nagarajan & Pop 2009). As a consequence, draft plant genome assemblies tend to be very fragmented, with a high number of contigs and collapsed repeats.

Overall, short read technologies alone are not suitable for highly polyploid and repetitive plant genomes as they prevent the construction of high quality assemblies. In order to overcome some of the assembly constraints, additional strategies and or data can be utilised. One of the most common solutions is to use paired end or mate pair reads. These reads come from both ends of a DNA fragment that is too long to be sequenced entirely. Knowing the size of the fragment can inform how far the sequenced ends are from one another. This extra information aids with orienting shorter contigs into longer scaffolds. In one assembly project several libraries with different insert sizes can be sequenced, combining the strategies of paired end and mate pair reads to improve overall assembly quality.

Another improvement can be accomplished by hybrid assemblies, i.e. combining data from a number of different platforms or completely replacing short reads by third generation long read technologies, i.e. PacBio or Oxford Nanopore, as longer reads can resolve most of the repeats. These platforms not only extend the read length to the point where they easily span repeats, but also by numerous technological improvements they eliminate the issues with homopolymers and GC-/AT-rich parts. However, the weakness of these instruments is a high error rate. This can be overcome by sequencing to high coverage, which increases the overall cost of sequencing. Unfortunately, higher coverage can only improve the assembly up to a certain point, after which further increases will not affect the assembly contiguity.

Further improvements, which do not relate to sequencing, can be accomplished by creating genome maps. These can be either genetic or physical maps. They both represent a set of molecular markers with known position and distance on the genome,

either genetic or physical. Subsequently, a map can be used to resolve misassemblies and anchor *de novo* genome assemblies by means of orienting and ordering the contigs (Fierst 2015).

Typically, genome assembly is only the first step in a whole genome sequence project; functional annotations of protein and non-protein coding regions, comparative genomics, phylogenetic analysis, variation determination are only few examples of possible subsequent studies.

None of the genomes reported until now are fully completed. There are always gaps and errors present in the assemblies. Completely finishing a genome is an expensive, time- and labour-consuming task and impossible to accomplish with the current sequencing and assembly technologies. However, technologies in both of these areas are advancing very rapidly and methods outlined above are likely to be further developed and significantly improved in the near future. Moreover, new technologies are actively under development that can revolutionize the genomics field once more, e.g. sequencing based on microfluidic channels and nanowires (Zheng *et al.* 2005), electron microscopy (Bell *et al.* 2012), reversible terminator chemistry using photocleavable dyes and terminating groups (Wu *et al.* 2007), nanopore arrays (McNally *et al.* 2010), and sequencing by expansion (Kokoris & McRuer 2008).

1.5 Phylogenetic inference

Advances in DNA sequencing have delivered vast amounts of sequence data that is also used in the field of phylogenetics. Phylogenetic studies provide insight into evolutionary history and relationships between organisms. In these studies, species are defined in terms of their DNA, with or without supporting synapomorphic morphological characters (shared by two or more taxonomic groups of interest). A typical workflow includes sequence assembly, homologous region identification, sequence alignment, and inference of gene and species trees (Guang *et al.* 2016).

Commonly used analytical methods to infer phylogenies can be classified as distance- or character-based (Whelan *et al.*). In distance approaches, first the pairwise distance (number of differences between every pair of sequences) is calculated and represented as a matrix. Subsequently, a phylogenetic tree is reconstructed using a clustering algorithm (Van de Peer *et al.* 2009). Examples of those methods include neighbour-joining [NJ; Saitou & Nei (1987)] and unweighted pair group methods using

arithmetic averages [UPGMA; Michener & Sokal (1956)]. Computational speed and efficiency make the methods suitable for initial examination of the relationships between species in an alignment. However, they have several limitations. Distance-based methods can ignore some of the evolutionary information within the dataset by calculating pairwise distances only and disregarding character states and their relatedness. Furthermore, they provide only one tree as opposed to character-based methods where the final tree is a consensus from numerous candidate trees.

In character-based approaches each character from each sequence of interest is compared (in the sequence alignment) and scored; at the end a phylogenetic tree is reconstructed. Examples of character-based methods include maximum parsimony [MP; Sober (1988)], maximum likelihood [ML; Cavalli-Sforza & Edwards (1967)] and Bayesian inference [BI; Rannala & Yang (1996)]. These methods are more accurate than distance-based methods but considerably slower.

The MP approach seeks a tree with the lowest number of evolutionary changes among sequences and adopts a very simple evolutionary model in which all probable nucleotide substitutions are possible to the same extent. Maximum parsimony may report multiple similarly parsimonious trees that pose an additional challenge of finding the consensus tree. On the other hand, ML is a statistical concept which estimates a set of parameters with a probability under an explicit evolutionary model (Brocchieri 2001); the consensus tree represents the tree topology with maximum probability. Finding the ML tree is a mathematically complex and computationally demanding process. Both MP and ML utilise numerous search methods, i.e. exhaustive, exact and approximate. The exhaustive methods exploit all possible tree topologies and output the best topology; nevertheless are impractical in use. Exact methods (e.g. branch-and-bound) reduce the number of trees that must be examined and are mostly applied to small datasets. For large datasets, heuristic (approximate) methods, such as nearest-neighbour interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection reconnection (TBR), are used. Although they are extremely fast, a globally optimal tree cannot be guaranteed since heuristic methods report highly possible trees but not necessarily the one with maximum likelihood. However, they make the computational task feasible.

Bayesian inference also involves an evolutionary model selection and it is based on the posterior probabilities of phylogenetic trees that are calculated using Bayes theorem. The final tree is the one with the highest posterior probability. The strengths of BI include

the possibility of incorporating prior information, if such is available, and computational effectiveness (Yang & Rannala 2012).

1.5.1 Chloroplast-based phylogenies

The chloroplast has a maternally inherited haploid genome that is present in high copy number in plant cells. Most plants have an extremely conserved chloroplast genome with conserved gene order but occasional sequence rearrangements in some species (Wolfe *et al.* 1987). The chloroplast genome also does not undergo recombination.

In the last decade a number of full chloroplast genomes have been reported using various NGS platforms and have been proven effective for inferring genetic relationships in plants (Parks *et al.* 2009) as well as for plant and food identification. Advances in NGS data analysis also eliminated the need for chloroplast isolation and amplification prior to sequencing (Nock *et al.* 2011). Owing to their size [120-160 Kb; Green (2011)], chloroplast genomes are large enough to collect a sufficient number of differences in order to distinguish between and within species. Therefore, whole chloroplast genomes are used as a single locus for plant barcoding. Some very important phylogenetic trees have been resolved using plastid genomes, e.g. basal angiosperms (Moore *et al.* 2007) and eudicots (Moore *et al.* 2010).

The main drawback of using chloroplast DNA for phylogenetic reconstruction is that it does not capture biological events that may have happened throughout the life history of an organism, such as incomplete lineage sorting, interspecific hybridisation and coalescent processes (Comes & Abbott 2001; Hudson 1990; Mason-Gamer *et al.* 1995).

1.5.2 Whole genome-based phylogenies

The ever increasing number of whole genome sequences has opened possibilities of using greater amounts of data to infer phylogenetic relationships. To determine an evolutionary tree using whole genome data, a single or low copy set of orthologous sequences (genes descending from a common ancestor) is used. This multilocus analysis from a number of genomes offers more characters (information) than chloroplast genomes and delivers more accurate molecular taxonomy.

A phylogenetic tree reconstructed from DNA sequences from a group of genes is called a gene tree, while a tree that represents the evolutionary relationship between groups of species is called a species tree. These trees can be contrasting if the evolution of species and genes has been distinct. If the genes have their own descent patterns

and/or incomplete lineage sorting has occurred, the calculated gene tree topology is different from a species tree. Usually, the gene sequences are short and do not provide enough data to distinguish between or within species. However, it is important to create a species tree rather than gene tree and it can be accomplished by increasing the number of genes used in the reconstruction.

There are several strategies to analyse data from high numbers of genes or genomic regions. The first is to concatenate all sequences together into one supersequence (supermatrix) and analyse it as a single sequence/gene [e.g. Murphy *et al.* (2001)]. A disadvantage of this approach is that, because of the concatenation, it ignores the possibility of different evolutionary history for each of the genes, as well as different substitution rates. Additionally, a supermatrix may lead to incorrect topologies because of incomplete lineage sorting, interspecies hybridisation or gene flow (Maddison 1997). Nevertheless, this strategy is probably the most widely used among all approaches. The second method is to investigate each gene separately, infer a phylogeny and subsequently combine individual trees together into one supertree. A drawback of this approach arises at the combining step where the commonly used heuristic methods do not include statistical properties and ignore individual estimates from subtrees.

Views of phylogeneticists differ as to which approach is better; some researchers support the supermatrix approach [e.g. de Queirozand Gatesy (2007)] and others prefer the supertree [e.g. Bininda-Emonds (2004)].

1.5.3 Molecular clock

A molecular clock may be used to estimate the divergence time between organisms based on molecular data and the mutation rate of their DNA. The hypothesis states that the DNA sequence evolves at a reliable rate that is constant over time and among distinct species (Ho 2008), and based on this theory, a number of molecular methods and models have been developed to estimate times of speciation.

The first steps of the process consist of selection of an adequate method of analysis, model of nucleotide substitution and model of the molecular clock (Ho & Duchêne 2014). Main methods of analysis include distance-based, maximum likelihood and Bayesian inference. Strict and relaxed molecular clocks are the two most widely used models, where the former assumes homogeneous rates among sites in the data and has one parameter, i.e. the rate of substitutions, and in the latter a distinct evolutionary rate

can be assigned to each branch. In yet another model, the multi-rate clock, a restricted number of rates is used and, as a consequence, some branches exhibit the same substitution rates. Relaxed molecular clocks can be further divided into autocorrelated and uncorrelated clocks, where the former presumes that adjacent branches exhibit comparable rates, whereas in the latter this assumption is ignored.

Calibration is by far the most important step in the molecular clock procedure. It is usually done by constraining the age to one of the nodes or specifying a prior age distribution for a number of nodes in the phylogenetic tree (in BI). It is typically based on fossil records (the earliest fossil evidence) and has huge impact on the estimates. Other methods of calibration may include geological events (continental drift, colonisations), demographic events, and sequence data (age of the sample, used mainly for viruses). Multiple calibrations are also possible and if the substitution rate is known, it can be used for calibration as well.

1.6 Project objectives

The overall objective of this study was to investigate Australian wild rice species with an emphasis on two new perennial populations recently discovered in northern Queensland, referred here to as Taxon A and Taxon B.

The first aim was to determine the phylogenetic relationship between Australian wild rice relatives (Taxon A and Taxon B) and two closely related *Oryza* species, namely *O. rufipogon* and *O. meridionalis*. A further objective was to resolve the relationship between these rices and other AA genome *Oryza* relatives based on available full chloroplast sequences. Additionally, functional annotation of the Taxon A chloroplast genome was undertaken and comparative genomics of the species was explored. In order to perform these analyses a high quality chloroplast genome of Taxon A was required. Therefore, another aim of the project was to compare sequencing strategies, as well as analysis tools for accurate and cost effective whole chloroplast barcoding.

The central goal of the study was to apply advanced sequencing and genomic approaches to generate nuclear genome sequences for the Australian perennial wild rice relatives, both Taxon A and Taxon B. The target was to obtain the most complete genome possible taking into account the most advanced sequencing and assembly technologies, as well as the cost and timing. Having Taxon A and Taxon B genome sequences available, the study aimed to carry out a genome wide analysis of the wild rice and other

available rice genomes. This analysis included phylogenetic reconstruction with high resolution, divergence time estimation, description of the genomes, annotations of protein coding sequences, RNAs and non-coding segments of the genomes (repetitive sequences).

1.7 Thesis outline

Chapter 1 provides a general introduction to the thesis project including thesis objectives and outline. A brief, comprehensive review of recent and relevant literature concerning general information about the genus *Oryza*, its diversification, genomics, and sequencing technologies used in the field, as well as a review of Australian *Oryza* status is described in Chapter 2.

The following chapters (Chapters 3-5) cover the main research undertaken during the project. Chapter 3 aims to compare sequencing platforms and analysis tools for whole chloroplast barcoding, whereas Chapter 4 examines the resulting chloroplast genome of one of the novel wild rice populations identified in northern Australia (Taxon A). Chapter 5 provides strategies for sequencing and *de novo* assembly for generating Australian wild rice draft genomes. This chapter also includes basic assembly statistics and genome characteristics along with results from functional annotation, repeat content exploration and genome-wide studies. The results of whole-genome phylogenetic inference between Australian taxa and other *Oryza* species are presented herein as well. These include phylogenetic relationships, divergence time estimation and evidence for possible introgression events between those species.

In Chapter 6, the main results of the thesis are summarised and discussed. The final chapter, Chapter 7, lists the references used throughout the thesis.

Chapter 2

Literature Review

2.1 Genus *Oryza*

Genus *Oryza* belongs to the subfamily Ehrhartoideae of the family Poaceae. It is a small-sized genus including approximately twenty-three wild species and two domesticated species, *O. glaberrima* and *O. sativa* L. (Khush 1997). The exact number of species depends on the taxonomy adopted. *Oryza* species have an extensive habitat distribution throughout Asia, Australia, Africa, South and Central America. The Asian cultivated rice, *O. sativa*, is found all over the world and consists of two major sub-species, namely *O. sativa* sub-species *japonica* and sub-species *indica*. *Indica* can be further divided into *aus* and *indica*, and *japonica* into *tropical japonica*, *temperate japonica* and *aromatic* groups (Garris *et al.* 2005). *Oryza glaberrima*, the African cultivar, is grown on a limited scale in West Africa. Wild rice species are distributed in tropical and temperate regions, between 55° N and 36° S latitudes.

Both diploid (Uozu *et al.* 1997) and allotetraploid (Ge *et al.* 1999; Martinez *et al.* 1994) *Oryza* species have been reported (Table 2.1). The majority of recognized rice species are diploids (2n = 2x = 24) with the haploid chromosome number of 12. Nine species, which account for nearly one-third of the species number, were found to be tetraploids (2n = 4x = 48). Ten genome types were characterized, specifically AA, BB, CC, EE, FF, and GG for diploids, and BBCC, CCDD, HHJJ and HHKK for tetraploids (Ammiraju *et al.* 2010b; Ge *et al.* 1999; Lu *et al.* 2009). Both rice cultigens are diploids and belong to A-genome complexes along with six other wild relatives.

Table 2.1 *Oryza* species, their chromosome numbers, genome groups and distribution.

Species	Chromosome number (2n for x=12)	Genome group	Distribution
O. alta Swallen	48	CCDD	Central and South America
O. australiensis Domin.	24	EE	Northern Australia
O. barthii A. Chev.	24	AA	West and Central Africa
O. brachyantha A. Chev. et Roehr	24	FF	West and Central Africa
O. coarctata Tateoka	48	KKLL	Asian Coastal Area
O. eichingeri A. Peter	24	CC	Central Africa, Sri Lanka
O. glaberrima Steud.	24	AA	West Africa
O. glumaepatula Steud.	24	AA	South and Central America
O. grandiglumis (Doell) Prod.	48	CCDD	South America
O. granulata Nees et Arn. ex Watt	24	GG	South and Southeast Asia
O. latifolia Desv.	48	CCDD	Central and South America
O. longiglumis Jansen	48	HHJJ	Indonesia, Papua New Guinea
O. longistaminata A. Chev. et Roehr	24	AA	Africa
O. malampuzhaensis Krish. et Chand.[b]	48	BBCC	Southern India
O. meridionalis Ng	24	AA	Northern Australia, Papua New Guinea
O. meyeriana (Zoll. et Mor. ex Steud.) Baill.	24	GG	Southeast Asia
O. minuta J. S. Pesl. ex Steud.	48	BBCC	Southeast Asia
O. neocaledonica Morat	24	GG	New Caledonia
O. nivara Sharma et Shastry[c]	24	AA	South and Southeast Asia
O. officinalis Wall ex Watt	24	CC	South and Southeast Asia, Papua New Guinea, Northern Australia
O. punctata Kotschy ex Steud.	24, 48	BB, BBCC[a]	Africa
O. rhizomatis Vaughan	24	CC	Sri Lanka
O. ridleyi Hook. F.	48	HHJJ	Southeast Asia, Papua New Guinea
O. rufipogon Griff.	24	AA	South and Southeast Asia, Northern Australia
O. sativa L.	24	AA	Worldwide
O. schlechteri Pilger	48	HHKK	Papua New Guinea

[a] A tetraploid form of *O. punctata* is also known as *O. schweinfurthiana* Prod.

[b] *Oryza malampuzhaensis* Krish. et Chand. is also considered as tetraploid form of *O. officinalis*.

[c] *Oryza nivara* is also known as annual form of *O. rufipogon*.

2.2 Diversity in rice

There is an enormous diversity in rice species, cultigens, land races and wild relatives. The diversity is present at morphological, physiological and ecological levels and is due to different climates, adaptation processes, soils, water availability, seasonal changes, and agricultural practices (Garg *et al.* 2014; Giuliani *et al.* 2013; He *et al.* 2014; Iwamatsu *et al.* 2008; Scafaro *et al.* 2012).

On the morphological level wild rices are characterized, among other features, by long awns (bristles growing from the spikelets), mostly open panicles, low number of small dark grains, as well as shattering as a means of seed dispersal, higher dormancy levels, and partial outcrossing (Sweeney & McCouch 2007). In contrast, the typical characteristics of cultivated varieties are short awns, grains differing in size but mostly large and with no pigmentation (in Asian cultivars), compact panicles and a high number of seeds. Moreover, they are harvested more easily due to reduced shattering and uniform germination. These varieties are also almost entirely inbreeding.

Great diversity in rice can also be observed at the genomic level and in the genome sizes of *Oryza* species. *Oryza australiensis* (genome E) has the largest genome size [960 Mb; Figure 2.1; Ammiraju *et al.* (2006)] among all diploid species, which was found to be due to the high number of retrotransposon copies (8×10^4 per haploid genome) and repetitive sequences [more than 10% of the genome; Zao *et al.* (1989)]. The smallest diploid genome is of *O. brachyantha* [genome F; 261 Mb; Chen *et al.* (2013)] which is suspected to be a consequence of a decrease in the number of non-specific genome repetitive regions and retrotransposons or loss of the ability to amplify and spread throughout the genome. In the case of allotetraploid species, *O. ridleyi* was found to have the greatest genome size [1,283 Mb; Ammiraju *et al.* (2006)]. The genome sizes of cultivated species, *O. sativa* and *O. glaberrima,* are 390 Mb and 354 Mb, respectively (Kawahara *et al.* 2013; Wang *et al.* 2014b).

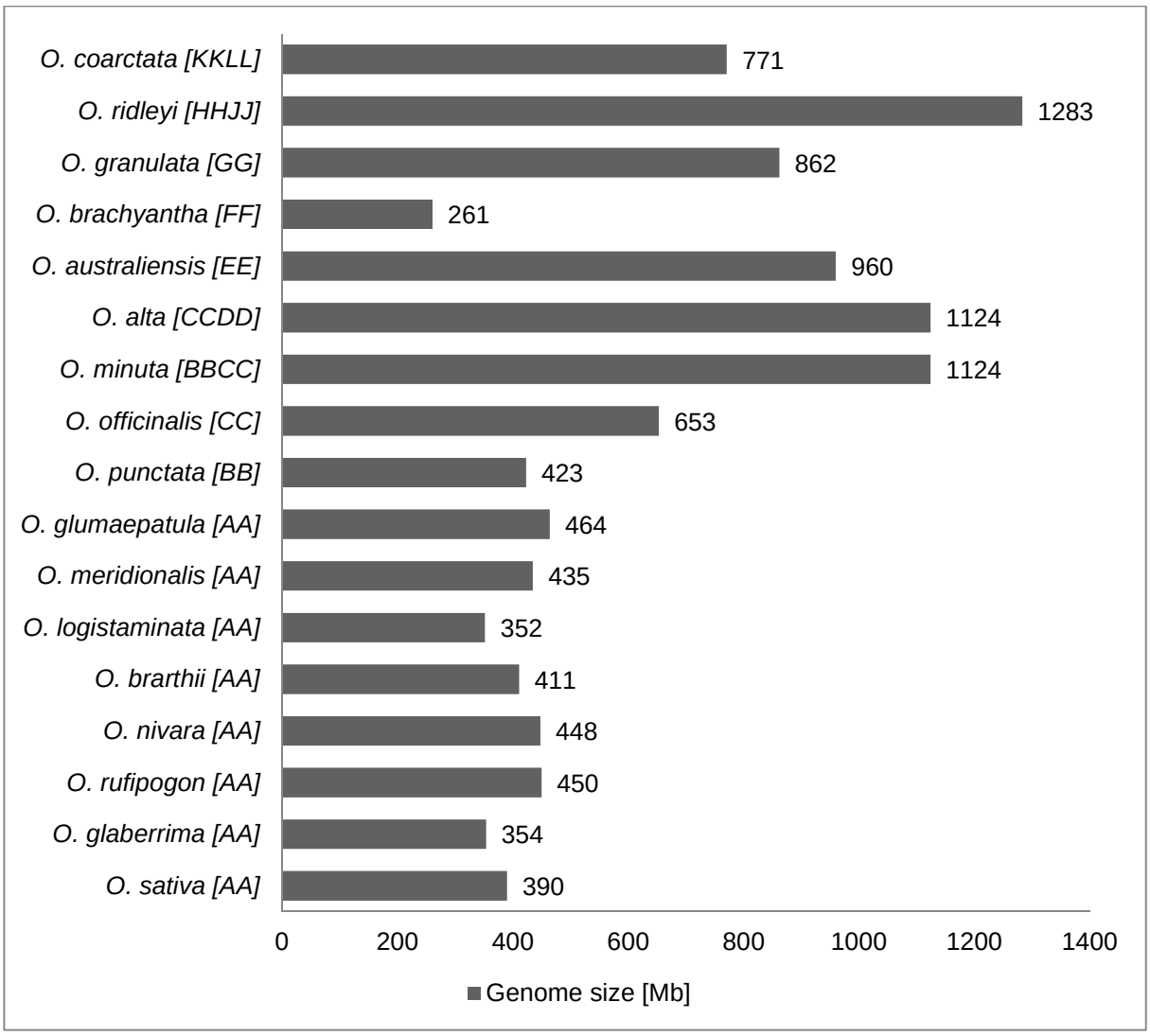


Figure 2.1 The *Oryza* species genome sizes. Data from Ammiraju *et al.* (2006), Kim *et al.* (2008) and Stein *et al.* (personal communication).

2.2.1 Origin of genus *Oryza*

The origin of the genus *Oryza* is not well established yet and two theories exist that try to explain the time and place of the origin of rice.

One hypothesis of *Oryza* origin suggests that the genus originated from the Gondwana continent and, after the break-up of the supercontinent, it spread widely across the world to Africa, South and Southeast Asia, South America and North Australia (Chang 1976; Second 1985). The divergence time was estimated based on isozyme electrophoresis data (Second 1985), which indicated that a split among *Oryza* species took place ~15 million years ago (mya) and separation of the African A-genome from Asian A-genome took place ~7 mya (Miocene epoch).

Another hypothesis states that today's *Oryza* distribution is a result of migration of humans, animals, and birds (Vaughan *et al.* 2005). It also recognises Asia as the origin and the centre of diversity of *Oryza*. A recent molecular phylogenetic study of the Oryzeae tribe also suggested that the genus *Oryza* arose in Asia and successively spread out to America, Africa and Australia, where it gave rise to other species across these regions (Tang *et al.* 2010). Moreover, the split of the *Oryza* genus (from other genera in Oryzeae) and first divergence inside the genus was calculated at ~17.5 and ~15 mya, respectively. These estimations differ from previously reported divergence times, ~14 mya and ~9 mya for *Oryza* split and time of the crown node of *Oryza*, respectively (Guo & Ge 2005). The study proposed that *Oryza* diversification was achieved by both dispersal and vicariance. Moreover, results suggested that there were multiple dispersal events from Asia to Africa and Australia and America. In this study, the Gondwana hypothesis has been rejected, as indications are that the breakup of the supercontinent took place long before the origin of the genus *Oryza* and the importance of long-distance dispersal, including oceanic dispersal, of plant species has been highlighted.

The latest study on diversification in the grasses family (Poaceae) included new fossil evidence from central India related to the rice tribe, Oryzeae, and gave a new light on the origin of this tribe and all grasses (Prasad *et al.* 2011). According to these new findings, Oryzeae divergence occurred earlier in the past than was previously reported. It was estimated to be 62.6 ± 9.9 mya, instead of the former assessment of 34.5 ± 6.8 mya (Vicentini *et al.* 2008). As a consequence, the time of origin of all grasses was pushed back as a whole, including the genus *Oryza*.

There are also two hypotheses of geographical origin of *O. glaberrima* in Africa. The first states that the heart of the origin was the inland Delta of the Upper Niger River and then the species spread across other regions (Porteres 1970). This was also confirmed in a recent multilocus DNA sequence study (Li *et al.* 2011). The second hypothesis considers several different locations of rice divergence (Richards 1996).

2.2.2 Domestication and cultivation of rice

Asian and African cultivated rice species were domesticated independently in Asia and Africa (Chang 1976). The process of the evolution of the two domesticated cultigens can be presented as: "wild perennial → wild annual → cultivated annual". In Asia the perennial wild species corresponds to *O. rufipogon* and the annual to *O. nivara* (also referred to as

an annual ecotype of *O. rufipogon*). In Africa these species are *O. longistaminata* and *O. barthii*, respectively.

Asian rice, *Oryza sativa*, is believed to be domesticated ~8,000-9,000 years ago in Yangtze Valley in China (Higman & Lu 1998; Liu *et al.* 2007). The oldest archaeological evidence of rice in Asia was found at sites of Xianrendong and Diaotonghuan (Yangzi River Valley) and was dated to 11,000-12,000 BC (Zhao 1998). It is not clear whether those were cultivated or wild rice plants. The oldest discovered rice paddy fields were dated to 4,000 BC which delivered a proof of rice cultivation in those times (Cao *et al.* 2006). In India, rice consumption was dated to 7,000-5,000 BC by archaeological works in Ganges Valley (Fuller 2006).

Domestication of African rice took place much later than Asian rice but before the introduction of *O. sativa* to Africa (Sweeney & McCouch 2007). Archaeological findings of rice grains in Africa are dated from 1800 BC to 800 BC (it is not known if those were domesticated or wild rices). The latest studies indicate that the African rice originated about 1500 BC in the Niger River Delta (Linares 2002). Evidence of the oldest domesticated plants comes from 300 BC-200 BC from the Inland Niger Delta (McIntosh 1995).

The domestication of *O. glaberrima* has been found to be due to a single domestication event (Li *et al.* 2011), whereas the precise origin of cultivated rice in Asia is an ongoing debate. Two main models support either single or multiple origins of *O. sativa* subspecies: *indica* and *japonica*. The first model states that rice was domesticated from wild rice species and the differentiation of *indica* and *japonica* occurred after domestication. This is supported by various studies of individual genes (Li *et al.* 2006; Sweeney *et al.* 2006), genetic loci (Tan *et al.* 2008), microsatellite data (Gao & Innan 2008) and genome-wide SNP assays of both subspecies (Huang *et al.* 2012a; Molina *et al.* 2011). The multiple-origin model proposes that both subspecies were independently domesticated from different wild rice populations. This hypothesis was proposed based on molecular phylogenetic analyses (Londo *et al.* 2006; Tang *et al.* 2006; Zhu & Ge 2005), archaeological evidence (Fuller *et al.* 2010; Higman & Lu 1998; Liu *et al.* 2007) and population genetics analyses of whole genome data (He *et al.* 2011; Xu *et al.* 2012). A recent study carried out on cultivated and wild rice accessions revealed a specific geographical and genetic origin of rice (Huang *et al.* 2012b). This study proposed than *japonica* was domesticated in southern China (Yangtze valley) and by crossing with

O. rufipogon (annual form, also called as *O. nivara*) gave origin to *indica* rice in Indochina and the Brahmaputra valley. However, a follow up study (Civáň *et al.* 2015) proposed not two but three independent rice domestication events in distinct regions in Asia. Besides *japonica* and *indica* varieties, they further described an origin of *aus* group (a variety of *indica*) to be from central India to Bangladesh.

Because of continuous gene flow and introgressive hybridisation between cultivated rice and the wild populations (Chang 1976; Song *et al.* 2006), it is challenging to study the origin of cultivated rice. Moreover, the presence of two genetically divergent cultigens, *indica* and *japonica*, makes it even more complicated.

2.2.3 Phylogeny of genus *Oryza*

A good understanding of the phylogenetic relationships between rice species is essential for improving the use of rice genomic and genetic resources. It is important for effective utilisation of wild rice as germplasm in rice improvement and production. Moreover, it provides a basis for breeding programmes and comparative genomics studies. Also, it plays a significant role in rice conservation.

Rice phylogeny has been widely analysed. The methods that have been used include analysis of morphological traits (Morishima & Oka 1960), isozymes (1996) and numerous molecular markers, i.e. Inter-simple Sequence Repeat [ISSR; Joshi *et al.* (2000)], SSR (Ren *et al.* 2003), Amplified Fragment Length Polymorphism [AFLP; Aggarwal *et al.* (1999)], Restriction Fragment Length Polymorphism [RLFP; Sun *et al.* (2001)] and Random Amplified Polymorphic DNA [RAPD; Martin *et al.* (1997)].

Recent advances in DNA sequencing have improved the phylogenetic reconstruction of the genus *Oryza* and other plant species. Formerly, phylogenetic analysis in rice was based on single chloroplast genes [e.g. *matK* gene; Ge *et al.* (2002)], combined chloroplast, mitochondria and nuclear genes (Guo & Ge 2005) or pooled chloroplast fragments from different regions [coding genes, intergenic regions and introns; Tang *et al.* (2010)]. Additionally, DNA fragments of chloroplast and nuclear origins were used to investigate the phylogeny of tetraploid genome species (Bao & Ge 2004). In other study, in order to completely resolve phylogenetic relationships between diploid species of *Oryza*, 142 single copy nuclear genes were used (Zou *et al.* 2008). The findings showed that tree reconstruction can be enhanced by adding more data and using appropriate

phylogenetic approaches. Phylogeny of *Oryza* genomes is currently well determined (Figure 2.2).

However, the phylogenetic relationship between AA genome species is more challenging. In one approach introns from genomic genes were sequenced and used for phylogeny reconstruction (Zhu & Ge 2005). In another, DNA fragments from the chloroplast, mitochondria and nucleus were used, which improved the resolution of the phylogenetic tree but did not unambiguously resolve the topology (Duan *et al.* 2007). Nevertheless, it showed the limitation of using only a small number of DNA fragments for phylogenetic reconstruction.

All these approaches have limitations and are not able to fully resolve tree topologies between closely related species. The main constraint is the relatively low number of DNA sequences used, which therefore limits the number of discovered genetic variations, resulting in low genetic resolution and great disparity between results (Duan *et al.* 2007).

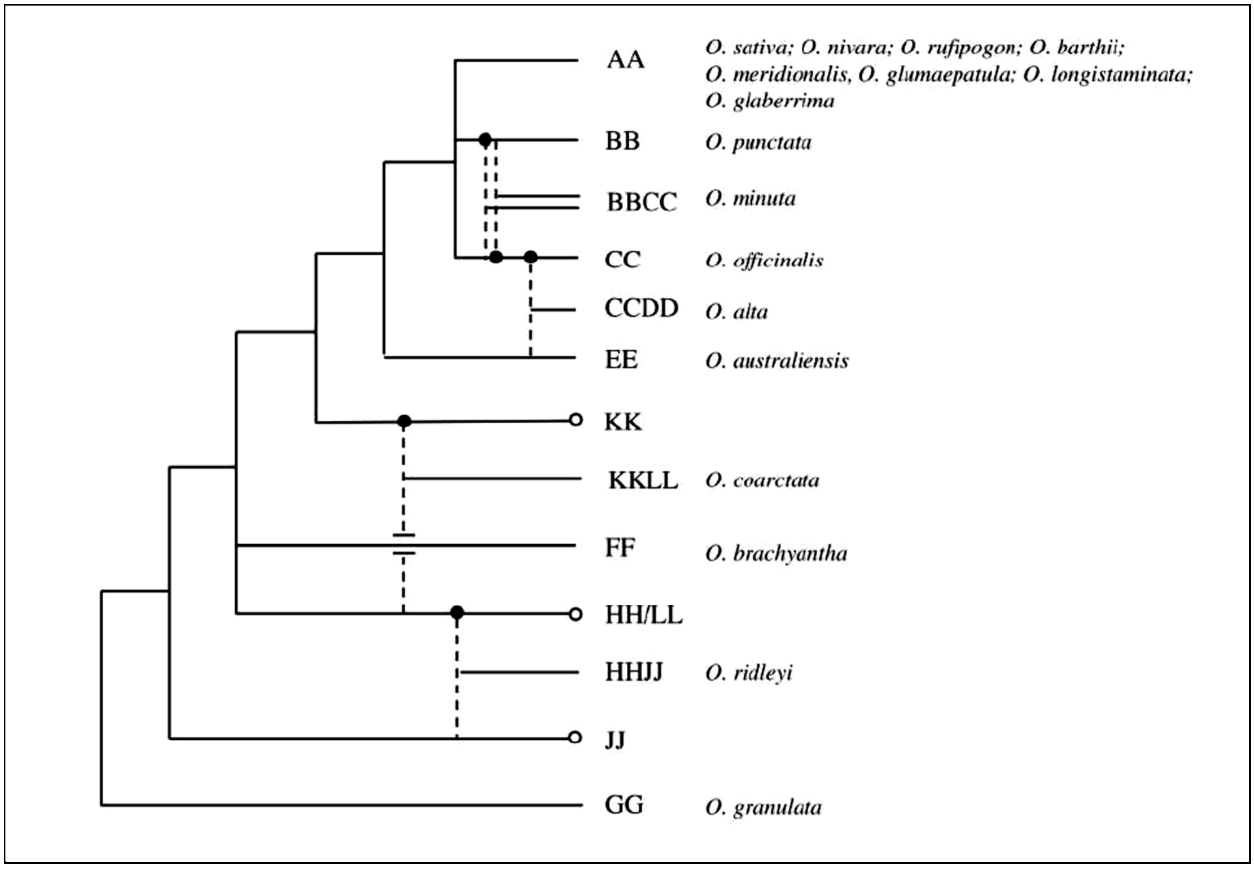


Figure 2.2 Phylogenetic relationship of *Oryza* genomes. Dashes lines (- - -) specify putative origins of polyploid species; black circles (●) specify putative maternal parents and white circles (○) specify unidentified genomes. Adapted from Goicoechea *et al.* 2010.

One of the most recent approaches was to use the full chloroplast sequence from several *Oryza* AA genome species (from Australia and Asia) to improve genetic resolution. Results also confirmed the distinctness of Asian and Australian wild rice relatives (Nock *et al.* 2011; Waters *et al.* 2012). This approach proved to be valid for phylogeny of closely related species and was subsequently used to resolve the entire AA genome rice species tree with addition of African and South American species (Wambugu *et al.* 2015).

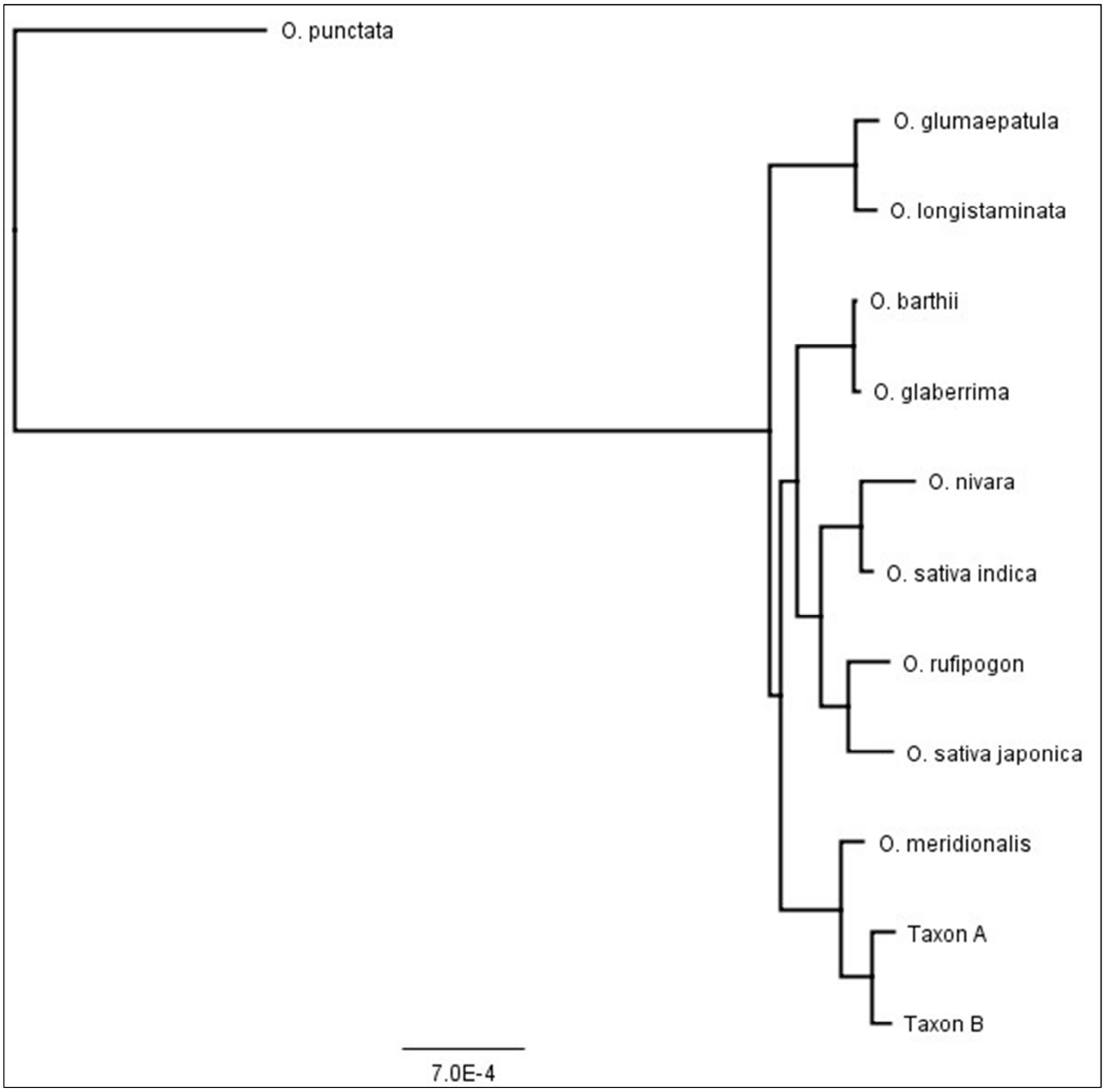


Figure 2.3 Relationship between chloroplast genomes of *Oryza* AA genome species. Modified from Wambugu *et al.* (2015).

In yet another study, 10 *Oryza* reference genomes (including eight from AA genome group) were used to reconstruct the phylogeny of wild rice species [Figure 2.4; Stein *et al.* (personal communication)]. However, the tree topology differed from the one inferred using chloroplast sequences and it did not include the newly characterised Australian wild rice populations (Taxon A and Taxon B).

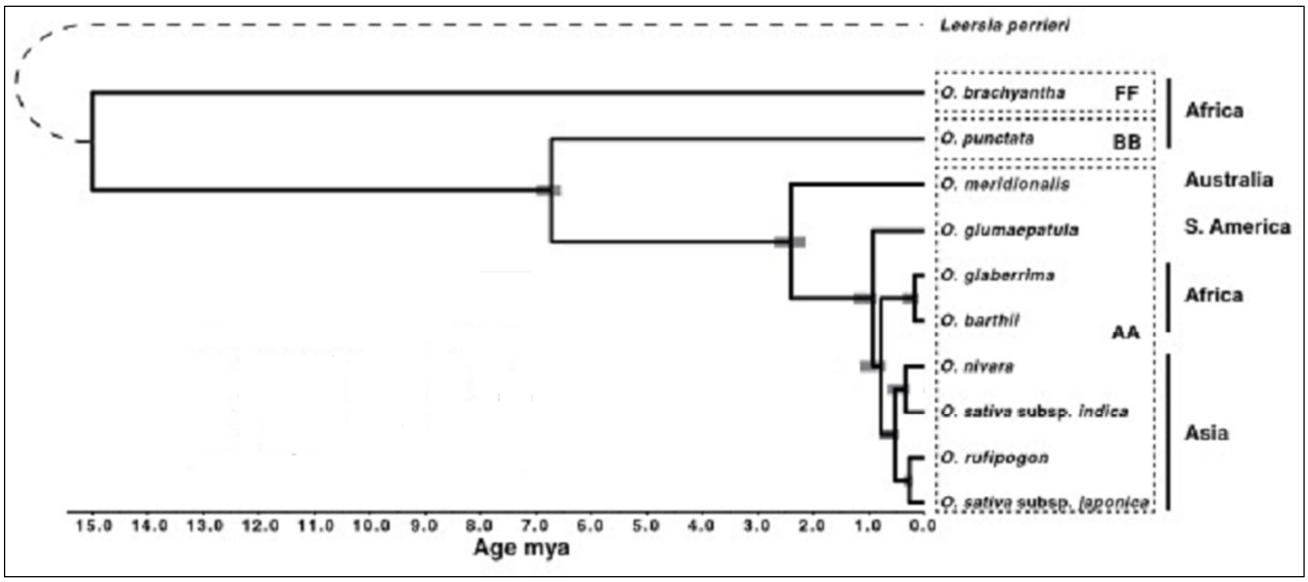


Figure 2.4 Relationship between genomes of *Oryza* AA genome species and the outgroup species *L. perrieri* (Stein *et al.,* personal communication).

2.3 Genomics and sequencing-based studies in rice

Draft genomes of *O. sativa* spp. *japonica* and *O. sativa* spp. *indica* were published in 2002 (Goff *et al.* 2002; Yu *et al.* 2002). The completed *O. sativa* spp. *japonica* var. Nipponbare reference genome was released in 2004 (International Rice Genome Sequencing Project 2005) and this high quality map-based rice genome revolutionised the field of rice research. It has been important for the crop's basic biology, as well as its improvement. The reference rice genome has had advantages not only for rice genomics but also for other major cereal crops, i.e. maize, wheat, barley and sorghum. It has facilitated identification of variations in different rice species and the search for important agronomic traits [e.g. Doi *et al.* (2008); Yamamoto *et al.* (2009)]. The reference genome has become a tool for investigating rice evolution, the domestication process and genome organisation

(Wing *et al.* 2005). It also has allowed SNPs screening in different rice cultivars to analyse genotypic and phenotypic diversity (McNally *et al.* 2009). Moreover, it has started GWAS in rice that focuses on identifying genetic variants and their implications for agronomic traits (Huang *et al.* 2010). Additionally, having a reference genome allows comparative analysis between rice cultivars and wild relatives. Also it provides the basis for discovery of new genes, above all those that are of agronomic significance. Since the release of the rice reference genome there have been various improvements in the sequence (Ouyang *et al.* 2007; Zhou *et al.* 2007) and in the latest *O. sativa* assembly [Os-Nipponbare-Reference-IRGSP-1.0 (IRGSP_MSU.v7)] a total of 37,872 loci were annotated, including 33,276 protein coding and 2,191 non-protein coding loci (Kawahara *et al.* 2013). Additionally, 2,405 loci were predicted *ab initio.* This reference genome is considered to be of very high quality, along with the *Arabidopsis thaliana* genome, with only a few gaps left at the centromere regions and even fewer in the euchromatin.

Some of the primary features described in the rice genome were (1) a high number of repetitive elements, mostly transposable elements, (2) paralogous sequences which are present probable due to ancestral genome duplication and (3) high diversity in sequence between varieties and highly polymorphic regions in the genomes.

Also the organelle DNA sequence has been reported form various rice species, both chloroplast (Hiratsuka *et al.* 1989; Masood *et al.* 2004; Tang *et al.* 2004; Waters *et al.* 2012) and mitochondria sequences (Fujii *et al.* 2010; Notsu *et al.* 2002; Tian *et al.* 2006).

2.3.1 Sequencing of wild rice species

Rice is the first crop for which two cultivated subspecies were sequenced, as well as numerous wild relatives and the sequencing of other species is in progress. The development and improvement of next-generation sequencing has allowed the sequencing, re-sequencing and assembly of other rice genomes at lower cost, higher depth and high throughput.

The *Oryza* Map Alignment Project was established in 2003 (Wing *et al.* 2005) with objectives to (1) generate reference sequences for all AA genome rice species and for one species of other genome groups, (2) create and analyse advanced backcross (ABC), chromosome segment substitution lines (CSSL) and recombinant inbred lines (RIL) populations for species with AA genome group for breeding purposes and (3) identify new collections of wild *Oryza* relatives for studies on development, population, preservation

and diversity within the entire genus. Owing to its great importance for modern rice breeding, it has become an international initiative called I-OMAP – International *Oryza* Map Alignment Project (Jacquemin *et al.* 2013). The species sequenced by I-OMAP are shown in Table 2.2.

New reference rice genomes give a great opportunity for comparative genomics within the genus *Oryza*. Because of great diversity among rice species, one reference genome for the genus *Oryza* is insufficient. The main objectives of having more rice reference sequences are (1) facilitating MAS in breeding programmes, (2) finding new valuable wild alleles with agronomic significance and (3) studying the evolution of the entire genus.

Table 2.2 Wild *Oryza* sequencing status as of July 2016.

Species	Genome group	Assembly size (Mb)	Status
O. australiensis	EE	-	in progress
O. barthii	AA	308	completed, unpublished
O. brachyantha	FF	261	completed
O. eichingeri	CC	-	in progress
O. glaberrima	AA	285	completed
O. glumaepatula	AA	373	completed, unpublished
O. granulata	GG	-	in progress
O. longistaminata	AA	-	in progress
O. meridionalis	AA	336	completed, unpublished
O. nivara	AA	338	completed, unpublished
O. officinalis	CC	-	in progress
O. punctata	BB	394	completed, unpublished
O. rhizomatis	CC	-	in progress
O. rufipogon	AA	338	completed, unpublished

2.4 Australian *Oryza*

2.4.1 Introduction

Four wild rice species of *Oryza* have been found in Australia: *O. australiensis*, *O. meridionalis* and *O. rufipogon* that are native to Northern Australia, and *O. officinalis* that has been observed in only two places in Australia (Table 2.3). Very recently two distinct groups of wild rice have been discovered in northern Queensland (Sotowa *et al.*

2013). Both of them exhibit a perennial life history and belong to the AA genome group. Here they are referred to as Taxon A and Taxon B (Table 2.3). They could be a new Australian wild rice species and represent a novel gene reservoir in rice.

Australian wild rice relatives are a potentially important source of valuable traits for improvement of domesticated rice. In Asia, cultivated rice, *Oryza sativa* L., has grown in close proximity to wild species for a long time with great potential for gene flow and continuous hybridization between them (Sweeney & McCouch 2007). On the contrary, rice seeds were brought to Australia around 1850 and the first record of rice cultivation is from 1906 in South Eastern Australia (Ricegrowers' Association of Australia). Nowadays, it is grown in the Murrumbidgee, Coleambally and Murray Valley areas in Victoria and southern New South Wales (Figure 2.5), whereas wild rice populations are almost exclusively found in tropical regions of Australia, and are thereby geographically (and genetically) isolated from the cultigens. Additionally, there was no cultivation and domestication pressure for selection and dispersal of tropical species. As a consequence, it is suspected that Australian *Oryza* species may possibly provide valuable information about the biology and genetics of rice preceding its domestication (Henry *et al.* 2010).

Table 2.3 Australian *Oryza* species, their life histories, genome groups and distributions.

Species	Life history	Genome group	Distribution
O. australiensis Domin.	perennial	EE	Northern Queensland, Northern Territory and Western Australia
O. meridionalis Ng	annual	AA	Northern Queensland, Northern Territory and Western Australia, Papua New Guinea
O. officinalis Wall ex Watt	perennial	CC	South and Southeast Asia, Papua New Guinea, Moa Island and Gove Peninsula (Australia)
O. rufipogon Griff.	perennial	AA	South and Southeast Asia, Northern Queensland, Northern Territory and Western Australia
Taxon A	perennial	AA	Northern Queensland
Taxon B	perennial	AA	Northern Queensland

However, these populations have not been well studied yet and only limited collections of Australian rice species are available. Neither of these species has been used extensively in rice breeding programmes. There is an exigent need for better understanding, characterisation and exploitation of these wild relatives, as well as conservation of wild rice populations and their habitats.

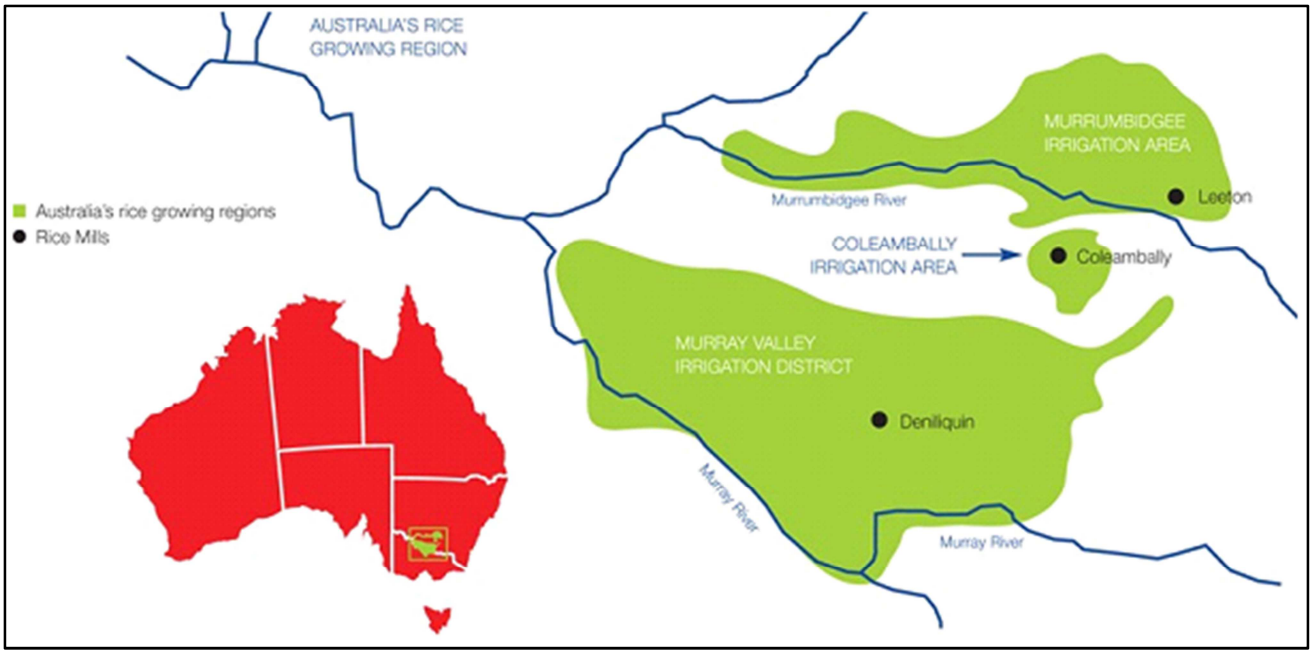


Figure 2.5 Rice growing regions in Australia (adopted from Ricegrowers' Association of Australia).

2.4.2 Revision of *Oryza* species in Australia

Oryza australiensis

Oryza australiensis Dom. is a perennial tall plant (>2 m) growing in tufts (Duistermaat 1987). Panicles are open; spikelets are 6-9 mm long and 2-3.2 mm wide; awns are 0.6-4.4 mm long. Anthers are 3.5-5.5 mm long and brown, and stigmas are white or yellow. It is a diploid species belonging to the E genome group (EE) with chromosome number 2x = 2n = 24.

This species has been reported only from Australia (northern Queensland, Northern Territory and Western Australia) and is endemic to these regions (Figure 2.6). O*. australiensis* grows in black and red loamy soil and clay, usually in open habitats. It is found in low undulating plains of *Eucalyptus* and *Leptochloa* or box woodland with a grassy ground layer. *O. australiensis* also found in wet places such as ponds, the edges of lagoons, seasonally dry swamps, muddy streams, or behind river levees. It survives dry season as rhizomes or seeds.

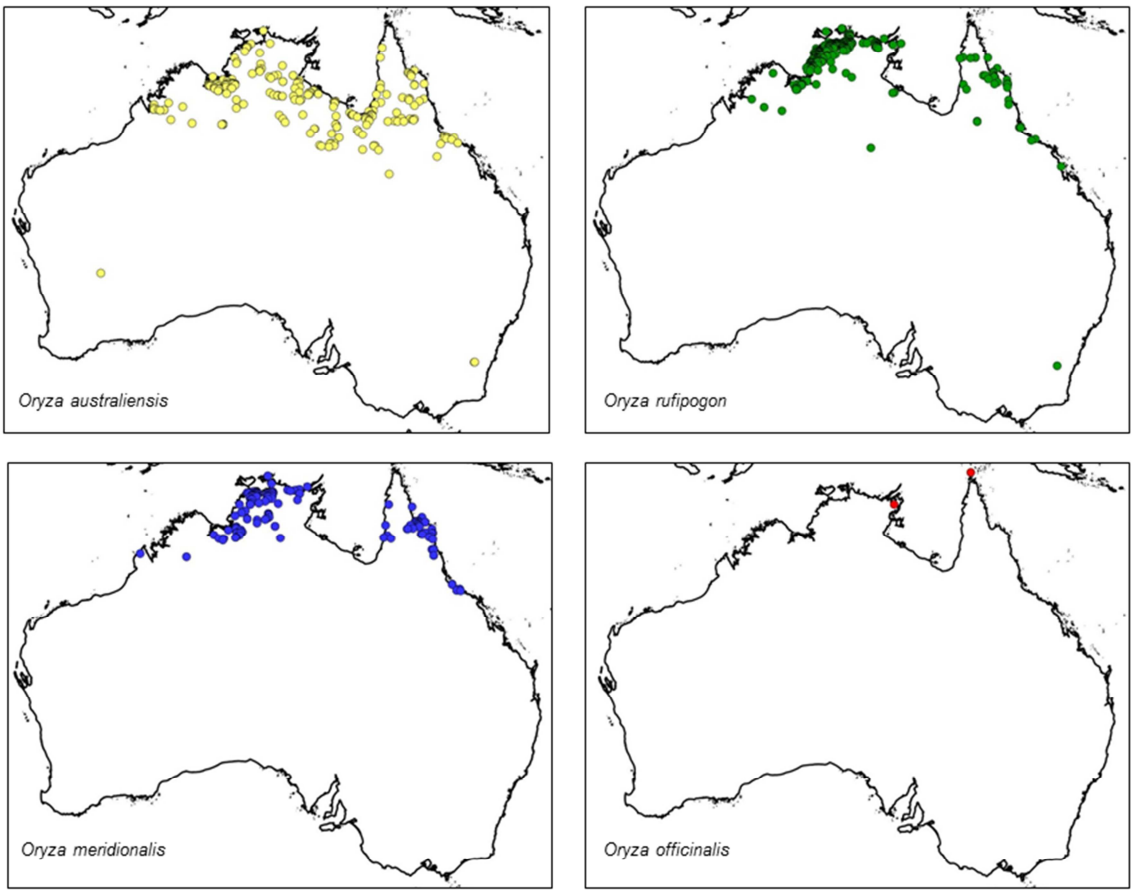


Figure 2.6 Distribution of *Oryza* species in Australia. Maps adopted from National Australian Herbarium.

Oryza rufipogon

Oryza rufipogon Griff. is a perennial plant growing in dense clumps (Duistermaat 1987). Plant height varies depending on the depth of the water and can reach between 1-5 m. Panicles are open; spikelets are usually 4.5-10.6 mm long and 1.6-3.5 mm wide and hairy. Anthers are 3.5-7.4 mm long, brown or yellow and stigmas are brown or dark purple. Awns are also very variable with lengths from 40 up to 100 mm long. It is a diploid species belonging to A genome group (AA) with chromosome number 2x = 2n = 24.

Besides its distribution in Australia (northern Queensland, Northern Territory and Western Australia; Figure 2.6), *O. rufipogon* is found in south and southeast Asia, and Papua New Guinea. It grows in moist clay, loam and black soil, usually in full sun. This species can be found in swampy habitats like marshes, open ditches, swamps, grasslands, ponds, billabongs, pools along river banks, at the edges of lakes, and in or at the margins of rice fields. The water depth preferences are between 0.2-4 m (Duistermaat 1987).

Oryza meridionalis

Oryza meridionalis Ng is an annual plant sometimes growing in small tufts (Duistermaat 1987). Plant height is usually 1-2 m. Panicles are compact; spikelets are typically 7.2-9.6 mm long and 2-2.8 mm wide. Anthers are 1.5-2.5 mm long and yellow and stigmas are dark purple. This species has strong awns about 70-130 mm in length. It is a diploid species belonging to the A genome group (AA) with chromosome number 2x = 2n = 24.

Oryza meridionalis is distributed mainly in Australia (northern Queensland, the Northern Territory and Western Australia; Figure 2.6) and Papua New Guinea. It grows in black and clay soils on moist ground, usually in open habitats. *Oryza meridionalis* can be found at the edges of swamps, temporary ponds and pools (Figure 2.7), lagoons, and rivers in shallow water up to 20 cm deep. It survives during the dry season as a seed.

Oryza officinalis

Oryza officinalis Wall ex Watt is a perennial plant, erect and usually of variable height. Panicles are open; spikelets 4.3-8.8 mm long and 2-3.3 mm wide; anthers around 1.5-4.5 mm long. It is a diploid species belonging to C genome group (CC) with chromosome number 2x = 2n = 24.

This species is widely distributed in South and Southeast Asia and Papua New Guinea, whereas in Australia it has so far been found only in two locations (Moa Island and Gove Peninsula; Figure 2.6). It grows in muddy silt soils or sandy soils on limestone in full sun or partial shade. It is found in seasonally wet areas, ditches, swampy and marsh places, near small ponds, and at the edges of lakes, streams, or rivers. It is also found in forests and on the edge of cultivated rice grounds or fruit and plantain fields.

Figure 2.7 *Oryza meridionalis* growing in a seasonally dry swamp in Northern Queensland.

Taxon A and Taxon B

Identification of potentially new species was carried out by field observation and molecular analysis and distinction was made based on molecular and morphological traits (Sotowa *et al.* 2013). Taxon A was characterised as perennial wild rice, morphologically similar to Asian *O. rufipogon* and classified as *rufipogon*-type (r-type) by Sotowa *et al.* (2013). It was found in a small swamp in Abattoir Swamp Environmental Park in northern Queensland [Jpn1 site in Sotowa *et al.* (2013)]. It is a diploid plant belonging to the AA rice genome group. It was found to carry organellar genomes (chloroplast and mitochondria) of *O. meridionalis*-type and a nuclear genome of *O. rufipogon*-type. Anthers of Taxon A are long (5.02 mm) and are similar in length to *O. rufipogon*. Awns are shorter however (7.60 ± 0.86 mm).

Taxon B was characterised as perennial wild rice morphologically similar to Australian annual *O. meridionalis* and classified as *meridionalis*-type (m-type) by Sotowa

et al. (2013). It was collected from a small pond near Laura in Northern Queensland [Jpn2 site in Sotowa *et al.* (2013)]. It is a diploid plant belonging to the AA rice genome group. It was found to carry both organellar genomes (chloroplast and mitochondria) and nuclear genome of *O. meridionalis*-type. The characteristic features of Taxon B are large spikelet size, lemma size (9.14 ± 0.53 mm), large awn size (10.23 ± 0.55 mm), and a higher number of spikelets per panicle (70 ± 14.4) than in Taxon A. The anther is shorter than in Taxon A. In Asia, the annual form of perennial *O. rufipogon* has been well known and is referred to as *O. nivara* (Sharma 2003). However, this form has not been observed in Australia and the unique annual *Oryza* species found is *O. meridionalis*. According to Sotowa *et al.* (2013), Taxon B is a possible perennial from of *O. meridionalis* that is also distinct from *O. rufipogon*.

Phylogenetic studies carried on new Australian perennial accessions placed Taxon A closer to *O. rufipogon* and Taxon B to *O. meridionalis* (Sotowa *et al.* 2013). However, this analysis was based on seven SSR loci from nuclear genomes of the two species. In order to increase the phylogenetic resolution, a higher number of DNA fragments and longer sequences should be used. Phylogenetic reconstruction based on full chloroplast genomes was demonstrated to improve accuracy of plant identification (Nock *et al.* 2011), including the genus *Oryza* (Waters *et al.* 2012). Moreover, this approach can also be used to confirm the origin of the plastids in both species. Similarly, comparison on the whole genome level between Taxon A and Taxon B, as well as other *Oryza* species gives more insight into their origin, diversification and evolution.

2.4.3 Genomics of Australian *Oryza* species

The International *Oryza* Map Alignment Project [I-OMAP; Jacquemin *et al.* (2013)], which was established with a goal of characterising the rice genomes, includes four species found in Australia, namely *O. australiensis*, *O. officinalis*, *O. rufipogon* and *O. meridionalis* (Table 2.4).

Although four of *Oryza* species found in Australia were included in I-OMAP only two of them have been sourced from Australia (Table 2.4) and only one rice species (*O. meridionalis*) is within the AA genome group. The recent identification of two new wild rice relatives (Sotowa *et al.* 2013) shows the need to incorporate more Australian rice accessions into sequencing projects. It is likely that *O. rufipogon* sequenced by I-OMAP will not be of much use as a reference genome sequence for Australian perennial rice populations.

Table 2.4 Reference genome sequencing project of *Oryza* species found in Australia (as of July 2016).

Species	Country of origin	Genome group	Assembly size (Mb)	Status
O. australiensis	Australia	EE	-	in progress
O. meridionalis	Australia	AA	336	completed, unpublished
O. officinalis	Thailand	CC	-	in progress
O. rufipogon	Vietnam	AA	338	completed, unpublished

2.4.4 Utilization

Abiotic and biotic resistance traits have been reported in Australian *Oryza* species. Resistance genes for bacterial blight and brown plant hopper were described in *O. australiensis*, such as bp*h10* and bp*h18* (Ishii *et al.* 1994; Jena *et al.* 2006). Moreover the latter was introduced into cultivated rice (Jena *et al.* 2006). *Oryza australiensis* also contains the blast resistance gene (*Pi40*) which was introgressed into breeding lines (Jeung *et al.* 2007; Suh *et al.* 2009). Among abiotic stresses, drought avoidance and tolerance to heat and drought were recognised in this species, as well as in *O. meridionalis* (Sundaramoorthi *et al.* 2009; Zhang & Wing (eds.) 2013). Chromosomal segment introgression lines have been established for *O. meridionalis* (Yoshimura *et al.* 2010), allowing introduction of useful genes and creating plants with different ploidy and genome compositions (Khush 1997). *Oryza officinalis* has been found to be resistant to bacterial blight, white-backed plant hopper and brown plant hopper (Brar & Khush 1997), as well as to thrips, green and white leaf hoppers, and stem rots (Zhang & Wing (eds.) 2013). Moreover, tolerance to heat has been demonstrated in this species. Numerous agronomically important loci have been revealed in *O. rufipogon* including resistance to bacterial blight, blast, brown plant hopper and tungro virus (Brar & Khush 1997). *Oryza rufipogon* populations have been also found to be moderately tolerant to sheath blight as well as abiotic stresses such as acid sulphate soils, high aluminium levels, phosphorous deficiency and have been shown to have increased elongation under deep water (Nguyen *et al.* 2003; Ram *et al.* 2007; Zhang & Wing (eds.) 2013). This species is also a source of cytoplasmic male sterility. Breeders are particularly interested in *O. rufipogon* because of its yield-enhancing traits. Nearly isogenic lines (NILs) derived from a cross of *O. rufipogon* and *O. sativa* and high-resolution physical mapping revealed a cluster of yield-enhancing QTLs derived from *O. rufipogon* introgressions (Xie *et al.* 2008).

Whole grains of Australian wild rice varieties may be consumed (Kasem *et al.* 2010). *O. meridionalis*, *O. australiensis* and *O. rufipogon* have great commercial potential. *Oryza officinalis* has moderate commercial potential in terms of grain size, shape and colour. The morphology of Australian wild rice grains are similar to cultivated *O. sativa*. Both Taxon A and Taxon B grains exhibit physical characteristics that are in a similar scope as those of cultivated rice (Tikapunya *et al.* 2016). There is still a need for complete physiochemical analysis of grains to reveal their nutritional potential as a whole grain food.

New Australian wild rice populations (Taxon A and Taxon B) may have great potential as a source of novel allelic variation and genetic diversity in rice. There is also a high probability that more new wild rice species may be found in tropical Australia. Therefore, it is crucial for future rice improvement programmes to better understand their potential as a source of desirable genes and traits, as well as creating novel collections of rice genetic resources.

Preface to Chapter 3

The following chapter was published as: Brozynska M, Furtado A and Henry RJ (2014) Direct Chloroplast Sequencing: Comparison of Sequencing Platforms and Analysis Tools for Whole Chloroplast Barcoding. *PLoS ONE* **9**:e110387.

My contribution to this study was as follows: performance of the experiments (100%), data analysis (70%), writing and editing the manuscript (50%) and designing the experiments (40%).

The sequencing data for *Oryza sativa* ssp. *japonica* var. Nipponbare was provided by Dr Catherine Nock from Southern Cross Plant Science, Southern Cross University, Lismore, NSW, Australia. These data consisted of Illumina reads (36 bp paired end) generated on the GAIIx sequencer.

Chapter 3

Direct chloroplast sequencing: comparison of sequencing platforms and analysis tools for whole chloroplast barcoding

3.1 Abstract

Direct sequencing of total plant DNA using next generation sequencing technologies generates a whole chloroplast genome sequence that has the potential to provide a barcode for use in plant and food identification. Advances in DNA sequencing platforms may make this an attractive approach for routine plant identification. The HiSeq (Illumina) and Ion Torrent (Life Technology) sequencing platforms were used to sequence total DNA from rice to identify polymorphisms in the whole chloroplast genome sequence of a wild rice plant relative to cultivated rice (var. Nipponbare). Consensus chloroplast sequences were produced by mapping sequence reads to the reference rice chloroplast genome or by *de novo* assembly and mapping of the resulting contigs to the reference sequence. A total of 122 polymorphisms (SNPs and indels) between the wild and cultivated rice chloroplasts were predicted by these different sequencing and analysis methods. Of these, a total of 102 polymorphisms including 90 SNPs were predicted by both platforms. Indels were more variable with different sequencing methods, with almost all discrepancies found in homopolymers. The Ion Torrent platform gave no apparent false SNP but was less reliable for indels. The methods should be suitable for routine barcoding using appropriate combinations of sequencing platform and data analysis.

3.2 Introduction

A universal plant barcoding method based upon sequencing a specific gene or combination of specific genes has proven elusive. Analysis of different plant groups has required the use of different genetic loci. The emergence of next generation sequencing (NGS) technology allowing whole chloroplast genomic DNA sequencing has provided the

opportunity to use the whole chloroplast genome sequence as a barcode (Straub *et al.* 2012). The chloroplast has unique features found in all green plants, conserved sufficiently to be readily aligned for comparison of different samples and large enough to contain variation, which allow species or sub-species distinction across the seed plants (Parks *et al.* 2009). Analysis of complex mixtures of plants has applications in many ecological studies (Taberlet *et al.* 2012).

A generic method for rapid and cost effective DNA-based identification of plants at this level (species or sub-species) will find wide application in industry and research (Henry 2001; Henry 2013). Industrial applications will include the identification of plant components in whole and processed foods and the management of food processing to ensure food safety and authenticity of labelling (Galimberti *et al.* 2013). This will complement the use of next generation sequencing to screen for the presence of pathogenic microbes (Diaz-Sanchez *et al.* 2013). Protection of intellectual property rights associated with plant varieties will also be simplified by a standard approach to plant identification. These tools will also support protection of biosecurity and management of wild plant populations of rare or threatened plants.

Early analysis relied on either specific amplification of the chloroplast genome or separation of chloroplasts from nuclear and mitochondrial DNA before chloroplast DNA isolation and sequencing the amplified or cloned product derived from PCR amplification (Jansen *et al.* 2005). However, the universal amplification of chloroplast sequences from all species has proven difficult and chloroplast purification is laborious and not perfect (Cronn *et al.* 2008; Lang & Burger 2007). Recently a simple approach involving NGS of a total genomic DNA preparation was proposed (Nock *et al.* 2011). This method relies on successful extraction of chloroplast genome sequence reads from total genomic DNA reads and their assembly to form a consensus sequence for the sample that can be used as a barcode. Reference-guided mapping of sequence reads and *de novo* assembly have been used to generate these whole chloroplast barcodes (McPherson *et al.* 2013).

DNA sequencing platforms differ in their throughput and accuracy (Loman *et al.* 2012). Accuracy of a DNA sequence assembled using NGS reads is dependent on the read length, sequencing depth, sequence coverage or width and evenness of coverage but also on the accuracy of the sequencing platform and the assembly and analysis pipeline (Sims *et al.* 2014). We compared two different sequencing technologies, the Illumina platform that uses sequencing-by-syntheses (SBS) chemistry and the Ion Torrent

that is based on semiconductor sequencing technology. Sequencing on both platforms commences with library preparation, which involves fragmentation of total genomic DNA, purifying to uniform and desired fragment size and ligation to sequencing adapters specific to the platform. In Illumina systems, the fragments are subsequently denatured and fixed on the surface of a proprietary glass flowcell, followed by solid-phase amplification (bridge amplification). As a result, clusters are created that contain clonally amplified DNA templates. Next, the clusters are sequenced in parallel using four fluorescently labelled nucleotides with reversible dye terminators. After every sequencing cycle and base incorporation to the template the polymerization terminates, a charge-coupled device (CCD) captures the fluorescent signal and identifies the base. After the base call the dye is cleaved and the sequencing process continues. This technology was first implemented in an instrument called Genome Analyzer (GA) with subsequent releases of improved series up to GAIIx. The next major system launched by Illumina was the HiSeq2000 with improved output and read length. Data from both GAIIx and HiSeq was used in the current study.

In Ion Torrent systems, the prepared libraries are immobilized to beads and amplified using emulsion PCR that takes place within microdroplets of aqueous solution and oil (Rothberg *et al.* 2011). During the sequencing, individual bases are incorporated by the action of a DNA polymerase. As a result of this reaction, a proton is released and the resulting change in pH is measured. The reaction occurs in a proprietary chip, which acts as a pH meter. Unlike Illumina, this system does not use fluorescent dyes and light detection, shortening the sequencing time. The Ion Torrent instrument used in the study was a Personal Genome Machine (PGM).

We examined the reliability of these two different sequencing platforms coupled with different data analysis approaches when applied to obtaining the chloroplast consensus sequences of a wild rice genotype and a cultivated rice genotype. Although the Ion Torrent platform is widely used for sequencing and limited published data is available on its application for transcript (Lee *et al.* 2013) and amplicon (Hartwig *et al.* 2012) sequencing in plants, it has not previously been applied to plant genomic DNA analysis. The most common applications of the Ion Torrent system have been to small genomes (bacterial samples) and targeted sequencing of amplicons or transcripts, whereas the Illumina has been widely used for applications including whole human genome re-sequencing, *de novo* plant and animal genome sequencing, exome, transcriptome (RNA-Seq) and metagenomics investigations. The chloroplast genome represents a specific sequencing

challenge with the presence of an inverted repeat. This study compares total DNA sequencing with the Illumina and Ion Torrent platforms and options for analysis of the sequences generated to produce a whole chloroplast genome sequence as a barcode for plant identification. Whole genome NGS was carried out on two rice genotypes: *Oryza sativa* spp. *japonica* var. Nipponbare referred to here as the reference-rice-genotype (R-rice-genotype) as its chloroplast sequence (reference-sequence) is publicly available and an Australian wild rice genotype sample (W-rice-genotype).

3.3 Materials and Methods

3.3.1 Plant materials

Seeds of rice (*Oryza sativa* var. Nipponbare), referred to as reference-rice-genotype (R-rice-genotype), were germinated and entire seedlings were used for DNA extraction. Leaf tissue from a wild rice plant, referred to as wild-rice-genotype (W-rice-genotype) collected from a field containing a population of perennial wild rice (*Oryza rufipogon*-like) at Abattoir Creek in North Queensland, was used for DNA extraction (Sotowa *et al.* 2013).

3.3.2 DNA extraction

DNA from leaf tissue of the W-rice-genotype was extracted using modification of the CTAB method (Furtado 2014). DNA from seedlings of the R-rice-genotype was extracted as described for the W-rice-genotype but with further purification as described by (Nock *et al.* 2011).

3.3.3 Sequencing

The Illumina and the Ion Torrent sequencing platforms were used for the shotgun sequencing of total genomic DNA of both rice samples. The wild rice sample was sequenced on the Ion Torrent (200 bp reads) and the Illumina HiSeq platform (100 bp paired end reads). The cultivated rice genotype was subjected to sequencing using the Ion Torrent (200 bp reads) platform, while Illumina reads (36 bp paired end), generated on the GAIIx, were sourced from available archived data (Nock *et al.* 2011). Sequencing on the Illumina HiSeq 2000 (Illumina, San Diego, CA, USA), to generate 100 bp paired end reads with an average library insert size of 500-600 bp, was outsourced to the Australian Genome Research Facility (AGRF, Melbourne, Australia). For sequencing using the Ion Torrent platform, genomic DNA was sheared using the Covaris S220 instrument (www.covarisinc.com) and used for preparing sequencing libraries according to the

standard Ion Torrent PGM protocol. The resulting individual DNA libraries were quality checked and quantified on the Agilent 2100 Bioanalyzer using the High Sensitivity DNA kit (Agilent). Following template amplification and enrichment on the Ion OneTouch2 (Ion OneTouch 200 Template Kit v2; #4478316) and OneTouch2 ES instruments, each sample was loaded onto one PGM #318 chip and sequenced using Ion PGM Sequencing 200 Kit v.2 (#4482006) according to manufacturer's protocol.

3.3.4 Data analysis – consensus

Raw reads from both sequencing platforms, PGM Ion Torrent and Illumina, were imported to CLC Genomics Workbench 6.0 (CLC-GW; www.clcbio.com) and read statistics assessed using sequencing data quality control, followed by read trimming for quality, length and presence of ambiguous bases. Ion Torrent reads were trimmed with the quality score limit set to 0.05 (which corresponded to Phred quality value > 15) and a minimum read length of 30 bp. Reads from both Illumina platforms, GAIIx and HiSeq, were trimmed with a quality score limit of 0.01 (which corresponded to Phred > 22) and the same minimum read length of 30 bp. Ion Torrent data was trimmed at a slightly lower quality value due to the lower average quality compared to the Illumina reads (Table 3.1; avg. Phred for Ion Torrent reads: 25 and 26, and avg. Phred for Illumina: 29 and 32). The Illumina HiSeq platform generates sequencing data fifty to sixty times more in order of magnitude as compared to the Ion Torrent data. Thus, in order to compare Ion Torrent and Illumina HiSeq platforms at similar coverage, a subset of HiSeq Illumina reads for the wild rice sample and matching the Ion Torrent read numbers for that same sample, was randomly extracted from the whole reads set. Ion Torrent raw reads were also alternatively trimmed using the Torrent Suite Software version 3.6 using default analysis settings with the following modifications: the quality-based trimming of the sequencing reads was adjusted to predicted-Phred-score 17 over the sliding window of 20 bp (default values 15, and 30, respectively) and the minimal length filter for reported reads was set to 20 (default 8).

Consensus sequence generation using the mapping tool was carried out as follows. The CLC-GW was used to map trimmed reads to the chloroplast genome sequence of *Oryza sativa* spp. *japonica* var. Nipponbare (GenBank accession: GU592207) used as the reference-sequence. The mapping process involved the following parameters: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.9, similarity fraction 0.9, and global alignment setting. Non-specific matches were also mapped randomly to the

reference genome. Consensus sequences were extracted from all mapping runs giving five sequences in total; two consensus sequences for *O. sativa* Nipponbare each derived from the Illumina GAIIx reads and the Ion Torrent reads, and three consensus sequences for the wild rice sample each derived from the Illumina HiSeq sequence reads, Illumina HiSeq sequence subset-reads and the Ion Torrent reads. Conflicts between reads were resolved by voting for the majority of the reads at the given position. Consensus sequence were aligned to the reference sequence and analysed for SNPs and indels.

Consensus sequence generation using the assembly tool was carried out as follows using both the CLC-GW and the Torrent Suite Software. The CLC trimmed reads from both platforms were assembled in CLC-GW using the *de novo* assembly tool with the following parameters: automatic word and bubble sizes with 200 bp for minimum contig length. Moreover the reads were mapped back to contigs (mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.9, and similarity fraction 0.9) and contigs were updated based on the results. The trimmed Ion Torrent reads were also assembled in the Torrent Suite Software using the Assembler plugin v.3.4.2 utilizing the MIRA assembly algorithm (http://sourceforge.net/projects/mira-assembler/). The assembly was performed in the reference-assisted mode using the reference-sequence. All contigs generated were aligned to the reference-sequence using NUCmer alignment generator from MUMmer package (Kurtz *et al.* 2004). Owing to the chloroplast genome structure (two long inverted repeats) the option to use all matches regardless of the uniqueness was enabled in the script. All other parameters were left as default. Based on the MUMmer output the longest contigs with the best mapping results (query coverage, precent identity and alignment length) were chosen as scaffolds and manually joined together by overlapping the remaining short contigs. In cases when there were not enough contigs to reconstruct the chloroplast genome sequence, *de novo* assembly was repeated with different parameter for word size in the CLC-GW and subjected again to the alignment tool. Consensus sequences from contigs mapping and joining were also screened for SNPs and indels based on the reference-sequence.

3.3.5 Data analysis – variant detection

The consensus sequences generated by both mapping and *de novo* assembly of the reads were imported into Geneious 6.1.6 (www.geneious.com) and each of them was separately mapped to reference-sequence. To generate the alignments the Geneious Map

to Reference tool was used with the default parameters. Subsequently the alignments were screened to find variations/ SNPs using the tool available in Geneious package.

3.3.6 Validation of the most probable variant

The wild rice chloroplast consensus sequences derived from reads mapping from Illumina and Ion Torrent platforms were aligned and scanned for SNPs and indels (as described above). If a variant was found in one consensus but was not present in the other it was reported as a discrepancy and closely studied. To confirm which of the disagreeing variants was more likely to be true the following procedure was adopted: (1) visual investigation of mapping results in CLC-GW and (2) aligning the conflicting regions from mapping derived consensus with contig sequences originating from *de novo* assemblies.

3.3.7 Data deposition

The chloroplast genome for Australian wild rice genotype sample (W-rice-genotype) has been deposited in GenBank with the sequence accession number KF428978.

3.4 Results

3.4.1 Mapping and *de novo* assembly of sequence reads

The CLC Genomics Workbench was used to trim sequence reads from both the Illumina and Ion Torrent platforms and then processed, using the mapping and *de novo* assembly mode, to generate a mapping-consensus or *de novo* consensus chloroplast sequences, respectively. The Torrent Suite Software, a dedicated analysis tool for Ion Torrent reads, was used to trim sequence reads from the Ion Torrent platform and then processed using the reference-guided *de novo* assembly mode (the only available mode) to generate a reference-guided *de novo* consensus chloroplast sequence. The published chloroplast sequence of the R-rice-genotype (*Oryza sativa* spp. *japonica* var. Nipponbare) was used as a reference chloroplast genome sequence (reference-sequence) for mapping of reads or contigs. Sequencing data of the two rice genotypes, generated by both sequencing platforms, was of good quality with average Phred scores ranging above 25 (Table 3.1).

Table 3.1 Summary statistics of raw reads obtained from PGM Ion Torrent and Illumina platforms for rice (*Oryza sativa* spp. *japonica* var. Nipponbare) and wild rice (*Oryza rufipogon*-like plant from Australia). Quality distribution is represented as Phred scores.

	Num. of reads	Approach	Length distribution	Length avg.	GC avg. (%)	Quality distribution	Quality avg.
Oryza sativa spp. *japonica* var. Nipponbare							
PGM Ion Torrent	6,655,114	Single reads	8-376	204	41.2	8-35	26
GAIIx Illumina (Nock *et al.* 2011)	9,689,084 (4,844,542 each pair)	Paired-end reads	36	36	46.3	2-34	29
Wild rice *(Oryza rufipogon*-like*)*							
PGM Ion Torrent	5,732,260	Single reads	8-362	189	41.3	6-36	25
HiSeq Illumina	390,365,066 (195,182,533 each pair)	Paired-end reads	100	100	41.5	2-40	32

In case of the R-rice-genotype, CLC trimmed-reads from both sequencing platforms mapped to the entire reference-sequence indicating complete reference-sequence coverage. However, the mapping-consensus chloroplast sequences varied in length based on sequencing platform reads. Similar results were obtained for the W-rice-genotype (Table 3.2).

Table 3.2 Mapping statistics of CLC trimmed reads from Illumina and PGM Ion Torrent platforms.

	***Oryza sativa* spp. *japonica* var. Nipponbare**		**Wild rice *(Oryza rufipogon*-like*)***		
	Illumina GAIIx	**Ion Torrent PGM**	**Illumina HiSeq**	**Illumina HiSeq subset**	**Ion Torrent PGM**
Mapped reads (Forward/ Reverse)	525,452 (262,567/ 262,885)	223,605 (112,154/ 111,451)	12,424,965 (6,211,497/ 6,213,468)	327,403 (164,024/ 163,898)	169,665 (85,317/ 84,348)
% mapped reads	5.42	3.39	3.43	3.46	3.17
Non-specific matches (% non-specific matches)	204,877 (39.9)	64,560 (28.9)	3,601,138 (29.0)	93,502 (28.5)	50,521 (29.8)

Non-perfect matches (% non-perfect matches)	18,385 (3.5)	164,176 (73.4)	1,745,820 (14.1)	47,444 (14.5)	130,855 (77.1)
Min coverage	7	96	1,203	75	69
Max coverage	688	534	13,231	360	431
Average coverage	127.7	333.4	8,143.2	218.4	243.6
% GC	39	39	39	39	39
Consensus length	134,551	134,520	134,531	134,529	134,525

Using the *de novo* mode in the CLC-GW and reads from both sequencing platforms, a number of contigs were generated with a subset of these mapping to the reference-sequence (Table 3.3). Similar results were obtained with the Ion Torrent reads when using the reference-assisted *de novo* mode on the Torrent Suite Software (Table 3.3). For both rice genotypes, contigs, which mapped to the reference-sequence, were checked if contiguous/overlapping to generate a *de novo* consensus sequence. Using the *de novo* mode in the CLC-GW, respective *de novo* consensus chloroplast sequences were generated for the reference-rice-genotype and the wild-rice-genotype but only from their respective Ion Torrent reads and not their Illumina reads. Using the reference-assisted *de novo* mode, reference-guided *de novo* consensus chloroplast sequences were generated for both genotypes.

Thus, all contigs generated from Ion Torrent reads using both assembly algorithms (in the CLC-GW and Torrent Suite Software) could be joined together based on their position and overlapping information creating full length consensus sequences for the chloroplast genome (Table 3.3). Consensus sequences could not be created from contigs generated in the CLC-GW from Illumina reads for either of the rice genotypes, possibly because of the limitations of read length (Table 3.3). Although these contigs were well mapped to the reference, there were gaps between some of them that prevented their joining.

Table 3.3 Assembly results performed on CLC Genomic Workbench and Torrent Suite Software of reads from Illumina and Ion Torrent platforms and mapping results of generated contigs to chloroplast reference-sequence.

Sequencing platform	Assembly method	Assembly tool	Overall number of contigs	Contigs mapped to chloroplast genome[b]	Contigs chosen to create consensus sequence
Oryza sativa spp. *japonica* var. Nipponbare					
GAIIx Illumina	*de novo*	CLC Genomic Workbench	961	25	NA
PGM Ion Torrent	*de novo*	CLC Genomic Workbench	268,869	116	25
PGM Ion Torrent	reference-assisted	Torrent Suite Software	213*	34	2
Wild rice *(Oryza rufipogon*-like*)*					
HiSeq Illumina	*de novo*	CLC Genomic Workbench	224,756	94	NA
PGM Ion Torrent	*de novo*	CLC Genomic Workbench	227,610	99	19
PGM Ion Torrent	reference-assisted	Torrent Suite Software	222[a]	31	2

NA – not applicable, corresponds to method that could not be used to create consensus sequence because of the gaps between contigs and reference genome, detailed description in the text.

[a] In Torrent Suite Software the assembly was run in the reference-assisted mode using chloroplast reference and the number of contigs given corresponds to contigs assembled based on the chloroplast sequence not the entire genome.

[b] Numbers correspond to all contigs mapped to the reference genome by means of MUMmer with default parameters and all matches regardless of their uniqueness, duplicated contigs are not reported.

3.4.2 Analysis of reference-rice-genotype

The mapping-consensus chloroplast genome sequence (134,551 bp) generated by mapping the Illumina reads from Nock *et al.* (2011) to the reference-sequence in the CLC-GW was, as expected, identical to the reference-sequence (Table 3.4). The mapping-consensus sequence generated by mapping of the Ion Torrent reads to the reference-sequence in the CLC-GW was shorter than the reference-sequence by 31 nucleotides with no base mismatches, but diverged at 31 positions comprising of 30 deletions and 1 insertion and where all but one of the polymorphisms were found in homopolymer regions of length from 2 to 9 bases. The *de novo* consensus sequence built by mapping and joining the contigs from *de novo* assembly of Ion Torrent reads (performed in CLC-GW) was shorter than the reference-sequence by 48 bp with no base mismatches, but diverged at 54 positions comprising of 48 deletions and 6 insertion and where all but two of the polymorphisms (indels) were found in homopolymer regions of length from 2 to 17 identical bases.

Table 3.4 Comparison of chloroplast consensus sequences of the cultivated reference rice genotype (*Oryza sativa* Nipponbare).

Sequencing platform	Source	Sequence length (bp)	Variants	Deletions	Insertions	Mismatches
Used available reference	GenBank (GU592207)	134,551	-	-	-	-
Illumina (Nock *et al.* 2011)	mapping-consensus	134,551	0	0	0	0
Ion Torrent	mapping-consensus	134,520	31	30	1	0
	de novo-consensus (CLC-GW)	134,503	54	48	6	0
	reference-assisted *de novo*-consensus (Torrent Suite)	134,581	50	9	41	0

The reference-guided *de novo* consensus chloroplast sequence obtained from contigs generated in Torrent Suite Software using Ion Torrent reads was longer than the reference-sequence by 30 bases with no base mismatches, but diverged at 50 positions comprising of 9 deletions and 41 insertions and where all polymorphisms (indels) were found in homopolymer regions of length from 2 to 12 identical bases.

The variants in the mapping-derived consensus sequences generated by the CLC-GW and using the Illumina or Ion Torrent reads were predominantly deletions, which is in contrast to the reference-guided *de novo* consensus chloroplast sequence generated by the Torrent Suite Software using Ion Torrent reads, where the variants were predominantly insertions. Overall, 135 variants were detected using these different approaches. Among them only 8 were common for all of the methods (the same position, length and type). The rest of the variants were either present in two of the approaches or were unique to only one of them.

The analysis of regions in the chloroplast genome where the indels were found in the Ion Torrent data (both mapping and assembly consensus) revealed biases in terms of polymorphism nature, nucleotide type and location. Data from all approaches were merged and duplicated regions were removed. The number of indels that involved C/G bases was slightly higher than A/T bases (53 and 44, respectively). It is worth noting that the GC content of the chloroplast genome is 39%. Also, more deletions than insertions were observed (except for the consensus from the Torrent Suite Software assembly) and the majority were of G/C nucleotides in short homopolymeric regions (2-5 bp long with the vast

majority of these variants occurring in 3 bp long homopolymer) (Figure 3.1). On the other hand, insertions were found mostly in longer homopolymers (5-17 bp long) and involved almost only A/T nucleotides.

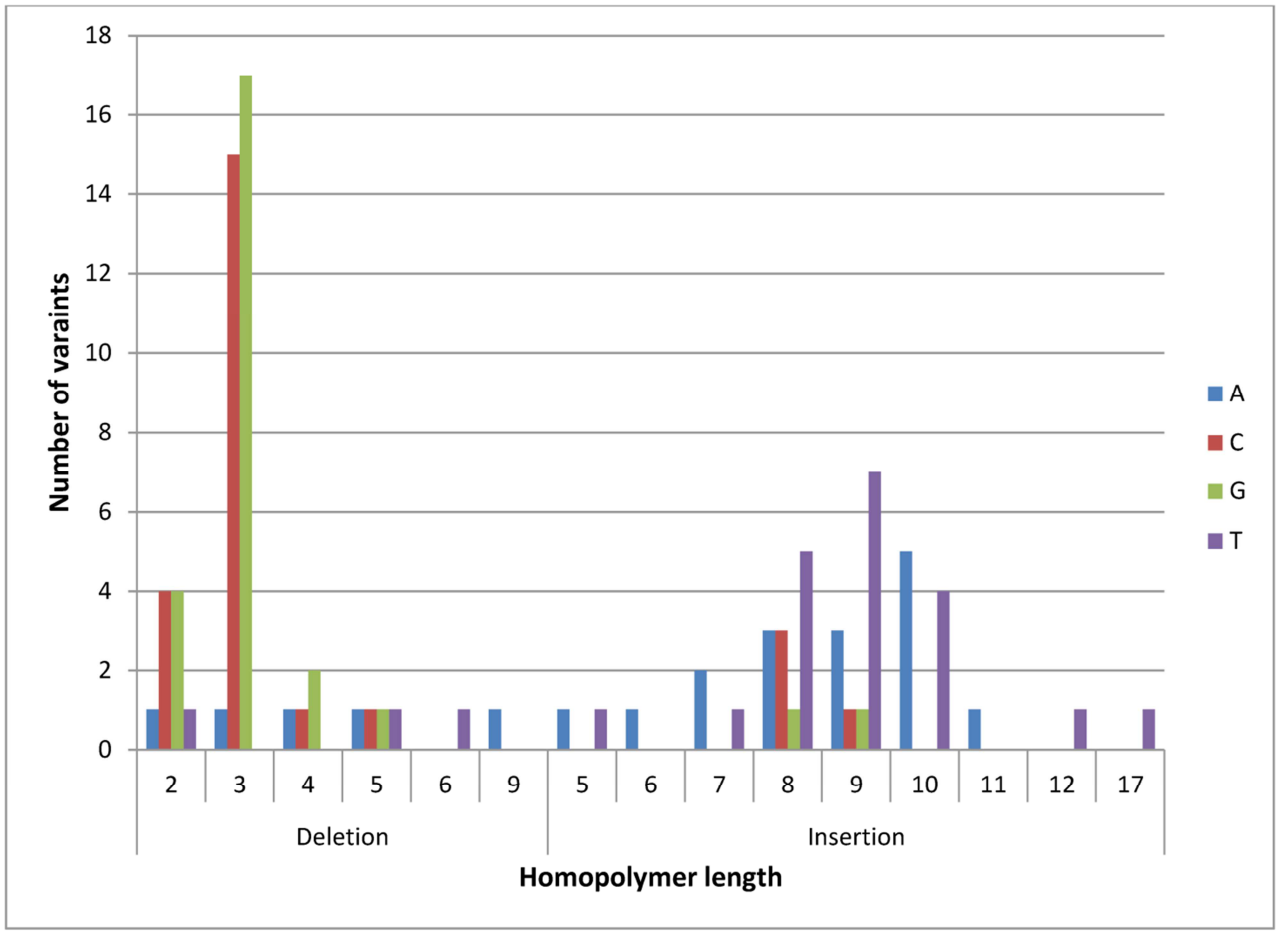


Figure 3.1 Variants in indels in cultivated (var. Nipponbare) rice chloroplast consensus. Sequences generated by mapping and assembly of Ion Torrent reads to the available chloroplast sequence in GenBank for this genotype. The number of variants is shown with respect to its type (deletion or insertion) and position (the length of homopolymer region where the variants were found).

3.4.3 Analysis of wild-rice-genotype

The reference-sequence was used to generate consensus chloroplast sequences using mapping or *de novo* approaches. For variant analysis, consensus sequences were compared to the reference-sequence unless indicated otherwise. The mapping-consensus sequence generated by mapping the Illumina reads of the wild rice plant using the CLC-GW was shorter than the reference-sequence by 20 bases (Table 3.5). In comparison, there were 128 variants found between the two sequences, which included 18 deletions, 13 insertions and 97 mismatches. Among mismatches, 92 were single-nucleotide variants

(SNPs) (48 transitions and 44 transversions) and 5 were multi-nucleotide variants (MNVs). The mapping-consensus obtained by mapping a subset of Illumina reads differed from the previous sequence by one deletion of two bases (Table 3.5); although, this variation was not found in the reference-sequence.

Table 3.5 Comparison of chloroplast consensus sequences of the wild rice (*Oryza rufipogon*-like).

Sequencing platform	Source	Sequence length (bp)	Variants	Deletions	Insertions	Mismatches
Used available reference	GenBank (GU592207)	134,551	-	-	-	-
Illumina	mapping-consensus	134,531	128	18	13	97
	mapping-consensus (subset of reads)	134,529	129	19	13	97
Ion Torrent	mapping-consensus	134,525	139	30	14	95
	de novo-consensus (CLC-GW)	134,521	155	43	17	95
	reference-assisted *de novo*-consensus (Torrent Suite)	134,554	147	23	28	96

The consensus sequence generated by mapping Ion Torrent reads of the wild-rice-genotype using the CLC-GW was shorter than the reference-sequence by 26 bases and shorter by 6 bases than the consensus sequence from Illumina sequencing. The two sequences differed by 139 variants, which included 30 deletions, 14 insertions and 95 mismatches. Among the mismatches 90 were SNPs (47 transitions and 43 transversions) and 5 were MNVs.

As was observed for the R-rice-genotype, the *de novo* consensus sequence differed from the reference-sequence at more positions than the mapping-consensus sequence (Table 3.4 and Table 3.5). In addition, a higher number of deletions were observed in consensus sequences obtained from the CLC-GW analysis as compared to the Torrent Suite Software analysis where more insertions were observed (Table 3.4 and Table 3.5).

3.4.4 Comparison of sequencing platforms

The mapping-consensus chloroplast sequences of the wild-rice-genotype generated in the CLC-GW using the Illumina and Ion Torrent reads were used to compare sequence

platforms. When comparing variations found in the wild-rice-genotype by both sequencing technologies, 20 discrepancies (variant in one consensus not present in the other) were observed (Table 3.6). However, in the Ion Torrent consensus sequence there were fifteen variants and in the Illumina consensus there were only three. Moreover, two variants found at the same position differed in nucleotide composition in both consensus sequences.

Table 3.6 Inconsistent variations found in wild rice chloroplast mapping-consensus sequences and their validation.

Variations		**Reference-sequence**		**Mapping-consensus sequence**		**The most probable variant**
No	**Type**	**Position**	**Allele**	**HiSeq**	**Ion Torrent**	
#1	MNV	57,036	TT	TT	AA	TT
#2	ins	65,465^65,466	–	–	TCCTATTTAATA	TTCCTATTTAATA
#3	MNV	66,897	CGAT	TAGA	CGAT	TAGAAATAAAAAATTCTAA
#4	SNP	66,902	C	A	C	TAGAAATAAAAAATTCTAA
#5	SNP	17,366	T	A	T	T
#6	SNP/del	17,368	C	A	–	–
#7	ins	3,545^3,546	–	AA	A	–
#8	del	21,808	C	C	–	C
#9	del	57,027	T	T	–	T
#10	del	81,342	G	G	–	G
#11	del	91,427	C	C	–	C
#12	del	91,589	C	C	–	C
#13	del	97,135	G	G	–	G
#14	del	111,639	G	G	–	G
#15	del	116,139	C	C	–	C
#16	del	118,025	C	C	–	C
#17	del	119,245	G	G	–	G
#18	del	122,914	C	C	–	C
#19	del	123,568	G	G	–	G
#20	del	133,816	C	C	–	C

Variations derived by Illumina and Ion Torrent sequencing.

SNP – single-nucleotide variant, MNV – multi-nucleotide variant, ins – insertion, del – deletion.

Read mapping files derived from both mapping-consensus sequences were visually inspected to check for reads mapped to all 20 positions to determine sequence read error or for mapping error. The region in the Ion Torrent mapping-consensus sequence with the substitution of TT → AA (Table 3.6, variant #1) had several mapped reads with many

mismatches. Reads from this region when extracted and blasted against the nucleotide sequence collection (nr/nt) at NCBI website (http://blast.ncbi.nlm.nih.gov) gave hits to the chloroplast but also to the nuclear DNA of *Oryza* species. In addition, this variant was not observed in the corresponding region of the contig sequences obtained from *de novo* assembly (from both analysis software), indicating that the variant was not due to sequencing error but due to mapping artefact. Similarly, mapping error and not sequence read error was the reason for variants detected at several positions in one or the other mapping-consensus sequences (Table 3.6: for #2 see Figure 3.2, for #3 and #4 see Figure 3.3), as these variants were not observed in the corresponding region of the contig sequences obtained from *de novo* assembly (from both analysis software). One of the variants (#7), an insertion of an A in the Ion Torrent consensus and AA in the Illumina consensus, was found in a long homopolymer stretch of 10 A's. The location of the variation suggested that in both cases it could be an error. Variations at this position in contigs sequences varied from 2 to 3 A's insertions. This polymorphism was not called in the consensus from the subset of Illumina data. Comparison of other chloroplast genomes, known to have been sequenced on the Illumina platform [GAIIx; Nock *et al.* (2011)], showed some to have this insertion, namely Australian *Oryza rufipogon* (GenBank accession: JN005833), Asian *Oryza rufipogon* (GenBank accession: JN005832) and *Oryza meridionalis* (GenBank accession: JN005831). Interestingly, the insertion was not present in the chloroplast genome sequence of *Oryza sativa* spp. *japonica* var. Nipponbare (reference-sequence), *Oryza sativa* ssp. *indica* isolate 93-11 (GenBank accession: AY522329) and *Oryza nivara* (GenBank accession: AP006728) that were sequenced by Sanger technology. Thus for this variant, we cannot conclude with certainty if this discrepancy is due to a mapping artefact or to a read error. However, all other errors found in Ion Torrent sequence (Table 3.6, #8 to #20) were not due to mapping errors but read errors from either deletions or insertions in homopolymer regions.

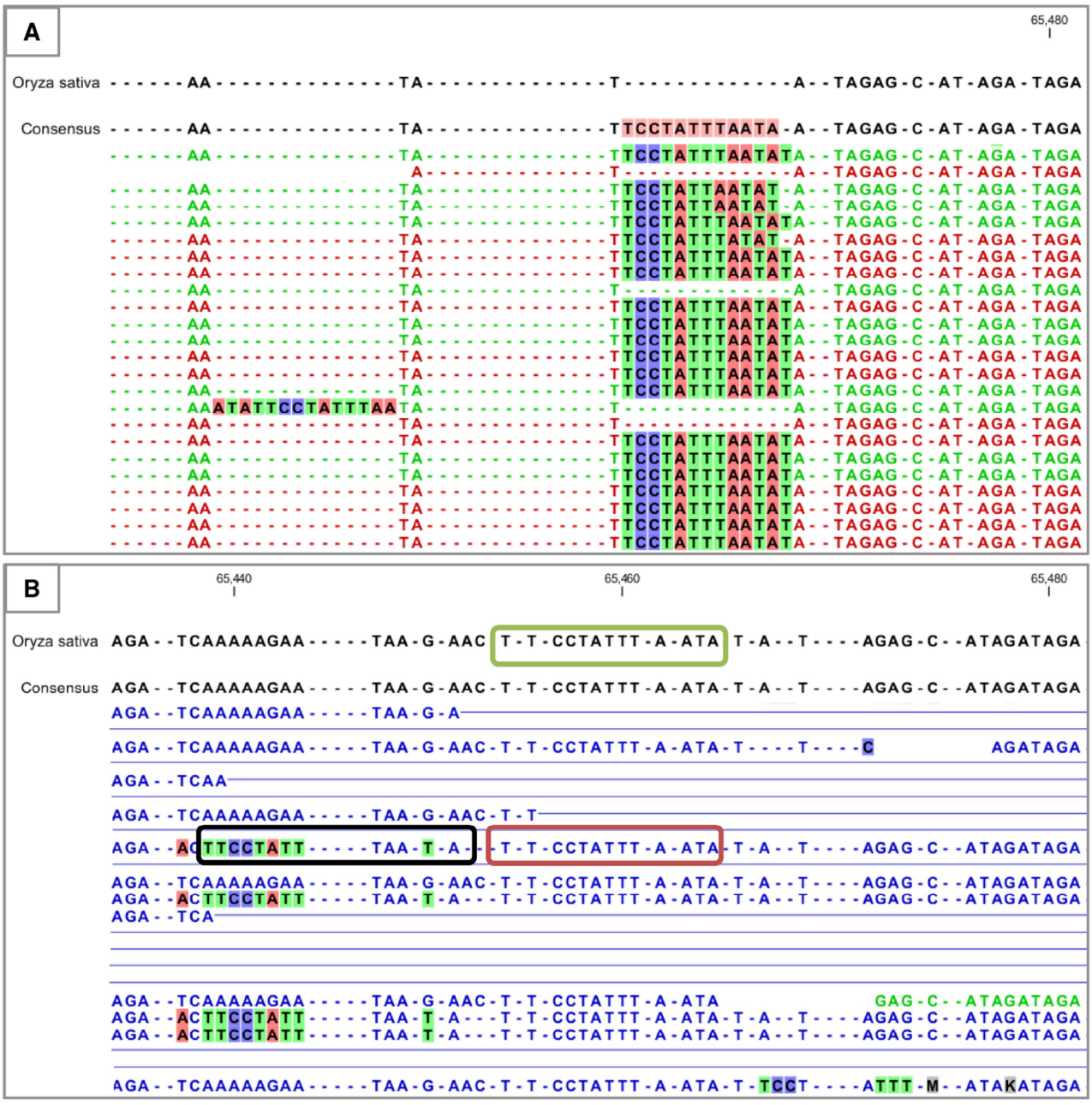


Figure 3.2 Snapshot of mapping results of wild rice Ion Torrent (A) and Illumina (B) reads. Reads were mapped to the chloroplast reference of *Oryza sativa* var. Nipponbare. In the mapping of Ion Torrent reads there was a long insertion (TCCTATTTAATA) reported in the consensus sequence of wild rice chloroplast. (A), marked with orange background colour). This insertion was missed in the mapping of Illumina reads, although it was present in the reads. (B), example of the read sequence marked in black rectangle). The nucleotides in the insertion were duplicated in wild rice (sequence marked in red rectangle), and not in the reference genome where only one copy of these nucleotides was present (marked in green rectangle). The duplicated region was a probable cause of the misalignment of reads. *Oryza* sativa – fragment of chloroplast sequence of *Oryza sativa* spp. *japonica* var. Nipponbare; Consensus – consensus sequence of wild rice chloroplast sequence derived by mapping reads from Illumina (A) and Ion Torrent (B) platforms to the reference. Nucleotides with background colours represent the mismatches between reads and the reference sequence; paired end reads are shown in blue; single reads are shown in green and red (in forward and reverse orientation, respectively).

Based on the analysis and all the findings using various approaches outlined above, an amended wild-rice-genotype chloroplast consensus sequence was created incorporating the most probable variants from Table 3.6. To assess the correctness of this

amended chloroplast sequence, paired-end reads from Illumina sequencing platform were re-mapped to this new wild rice chloroplast genome. Illumina reads as against Ion Torrent reads of the W-rice-genotype were selected for remapping as even short Illumina reads (36 bases) of the R-rice-genotype provided a consensus matching the reference-sequence [Table 3.4; Nock *et al.* (2011)], as against an inaccurate consensus when Ion Torrent reads were used (Table 3.4) due to indel-associated errors. The new mapping results for the wild rice show only one discrepancy described earlier (Table 3.6, #7) in a long homopolymer region. All other sites were identical giving an additional assurance of the correctness of the final consensus.

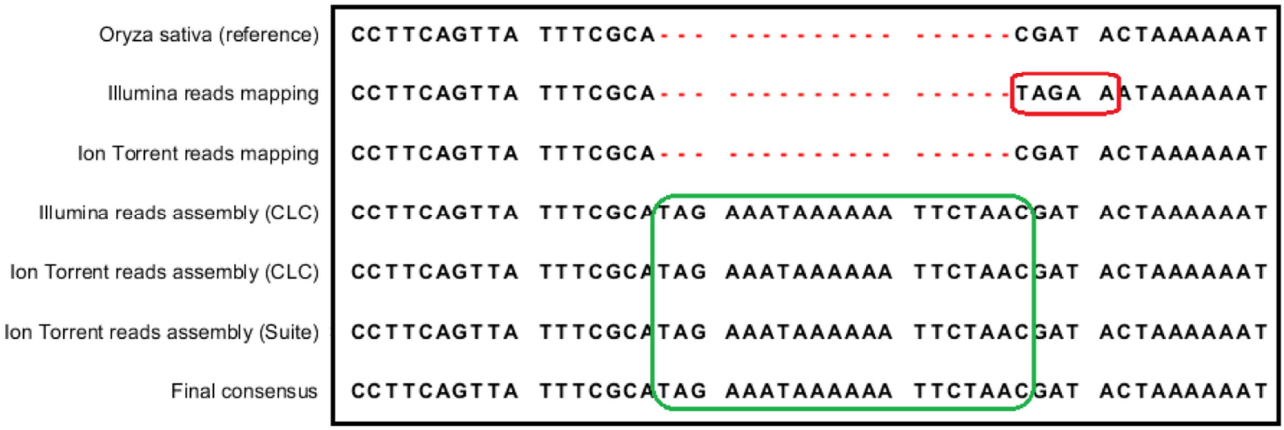


Figure 3.3 Alignment of regions #3 and #4 from Table 3.6 showing discrepancies in consensus sequences. The fragment circled in red shows false called SNPs (#3 and #4, Table 3.6, Illumina consensus); these SNPs were incorrect because of the long insertion present in wild rice sequence but not in the reference. The fragments circled in green illustrate this long insertion found in wild rice chloroplast genome by means of reads assembly from both platforms and both assembly tools. Final sequence was created based on this information. *Oryza* sativa (reference) – region 66860..66940 from chloroplast sequence of *Oryza sativa* spp. *japonica* var. Nipponbare; Illumina reads mapping and Ion Torrent reads mapping – regions from consensus sequence generated by mapping wild rice Illumina and Ion Torrent reads, respectively, to the reference sequence; Illumina reads assembly and Ion Torrent reads assembly – regions from contigs generated by assembly of reads from Illumina and Ion Torrent platforms, respectively; CLC – assembly performed in CLC Genomic Workbench; Suite – assembly performed in Torrent Suite Software; Final consensus – final wild rice chloroplast genome sequence (GenBank accession: KF428978).

3.5 Discussion

Various NGS technologies are now available for the rapid sequencing of whole genomes (Ferrarini *et al.* 2013; Rothberg *et al.* 2011; Shendure & Ji 2008). The choice of selecting one or more NGS technologies depends on the yield of data required, read length required, cost per data point and accuracy of the sequence data (Quail *et al.* 2012). Systematic as well as random errors can compromise the use of sequence data (Bragg *et*

al. 2013; Meacham *et al.* 2011). In this study, we analysed NGS reads of two rice genotypes obtained from the Ion Torrent platform and compared it for accuracy to those obtained from the Illumina platform. A comparison of the three sequencing systems is presented in Table 3.7.

Table 3.7 The comparison between the three sequencing systems utilised in the study.

	Illumina GAIIx	**Illumina HiSeq2000**	**Ion Torrent PGM – 318 chip**
Sequencing method	Synthesis (light detection)	Synthesis (light detection)	Synthesis (proton detection)
Amplification method	Bridge PCR	Bridge PCR	Emulsion PCR
Read length[a]	Up to 2 x 150 bp	1 x 50 bp, 2 x 50 bp, 2 x 100 bp	~ 200 bp, ~ 400 bp
Paired reads	yes	yes	yes
Insert size	Up to 700 bp	Up to 700 bp	Up to 250 bp
Output data/run	30 Gb	600 Gb	Up to 2 Gb
Time/run[b]	10-14 days	8-11 days	4-7 hours
Cost/Gb[c]	$148	$41	$1000
Instrument cost	$256 K	$654 K	$80 K
Accuracy	> 99.9%	> 99.9%	99%
Error rate[d]	≤0.1%	≤0.1%	~1%
Primary errors	Substitutions	Substitutions	Indels
DNA requirements	0.05-1 µg	0.05-1 µg	0.1-1 µg

Gb – gigabase, bp – base pair, K – thousand, µg – microgram.

[a] Annotation '2 x' refers to paired end reads and '1 x' to single reads.

[b] Run time from minimum to maximum read lengths.

[c] Includes one sample and one sequencing kit per run.

[d] Percentage of errors per base in single read.

The Ion PGM is characterised by long read length, very short run times and inexpensive consumables (Liu *et al.* 2012). The drawbacks include low output, high cost per gigabase (Gb) of sequence (Quail *et al.* 2012) and the longer time required for the library preparation relative to that required for the sequencing run. Moreover, the PGM produces biased sequences from high homopolymer regions, which result in indel errors (Bragg *et al.* 2013). However, this is a developing platform with high potential for improvement in accuracy, read length and cost per read and per Gb. As the machine is the chip, it can be easily upgraded by the release of new chips. The PGM was the first next-

generation sequencer with a price that dropped below $100 K. At the time of the study the PGM cost was around $80 K, which included the instrument, temple preparation system, enrichment system and server. The main advantages of HiSeq2000 system are the high throughput, large output data volumes per run and low cost of reagents per Gb (Glenn 2011). The shortcomings of the platform embrace short read length and the need for advanced computational resources to process and store the enormous data volumes created from each run. Furthermore, the run time is considerably longer than that for the PGM, and the Illumina has a much higher initial capital cost. The specification of the GAIIx system was very similar to the HiSeq2000 with smaller output per run, higher cost per Gb and significantly lower instrument cost. The sequencing accuracy on both Illumina platforms is high with the most commonly encountered errors being substitutions. A wide range of Illumina instruments are now available with differing data volumes, running times and costs.

We selected the cultivated rice genotype (var. Nipponbare) as one of the samples as its chloroplast sequence, derived by Sanger sequencing and considered as accurate, was available and could be used as reference sequence to check for errors in consensus sequences derived from NGS sequencing on both platforms. The aim was to generate both mapping and *de novo* consensus sequences to determine read accuracy of both platforms.

We were unable to obtain *de novo* consensus sequences for both genotypes with Illumina reads when processed by CLC-GW as the contigs generated were non-contiguous when mapped to the reference-sequence. In the case of reference-genotype, read coverage depth and read length contributed to non-contiguous contigs as the read depth in regions corresponding to gaps was between 7 and 30, when examined for base calling and read mapping depth to the mapping-consensus, which is well below the average mapping read coverage depth of 127.7 (Table 3.2). In the case of wild-rice-genotype with longer read length (100 bases), read length but not read coverage depth may have contributed to non-contiguous contigs as the read depth in regions corresponding to the gaps was between 7,000 to 11,000, when examined for base calling and read mapping depth to the in the mapping-consensus, which is well above the average mapping read coverage depth of 8,143.2 (Table 3.2). Interestingly, *de novo* consensus, although not accurate (Table 3.4) was generated for the reference-genotype using CLC-GW and the longer (200 bases) Ion Torrent reads. We can thus conclude that

read lengths of more than 100 bases would be required to successfully obtain a *de novo* consensus of the chloroplast genome.

In case of the reference-genotype, mapping-consensus sequences were generated using reads from both platforms, with no mismatches. However, mapping-consensus sequences showed several indels due to inaccuracies in the Ion Torrent reads mainly in homopolymer regions. In case of the short Illumina reads (36 bases) of the R-rice-genotype, neither the read length (36 bases) nor the sequence yield (minimum reference-sequence read mapping coverage of 7) was a limitation in generating an accurate mapping-consensus when processed by the CLC-GW, indicating the accuracy of the Illumina sequencing platform. In case of the longer Ion Torrent reads (200 bases), when processed by the CLC-GW and the Torrent Suite Software, consensus-sequences with no mismatches were generated but with several variants all of which were indels and predominantly in homopolymer regions (Table 3.4). These results indicate the actual presence of these anomalous indels in the Ion Torrent reads that contributed to the inaccuracy in the consensus sequence generated and to the unreliability of any indels calls made when using these reads. It is known that the Ion Torrent sequencing platform generates anomalous indels (Bragg *et al.* 2013). Interestingly, the preference of the CLC-GW and the Torrent Suite analysis tool in filtering out predominantly either insertions or deletions, respectively, as observed in both rice genotypes (Table 3.4 and Table 3.5), indicates the possibility of these analysis tools to be amended to deal with anomalous indels in the Ion Torrent reads. We also observed a bias for indels in homopolymer regions, where deletions occurred in shorter regions while insertions occurred in longer regions. These findings are inconsistent with a recent study on Ion Torrent sequencing bias where it was reported that an increased deletion rate was positively correlated with increased lengths of homopolymer regions and that insertions were independent of homopolymer length (Bragg *et al.* 2013; Ross *et al.* 2013). This is probably because of the different genomes used or that the chloroplast genome is AT-rich and the longest C/G stretches found are 9 bp long (2 regions), 8 bp (4 regions) and 7 bp (9 regions).

The homopolymer repeat numbers and lengths found in the chloroplasts are likely to be responsible for the systematic pattern of bias in the single base indels in these repeats (Figure 3.1), i.e. specific types of nucleotides deleted or inserted from long and short repeats. Specific types However, systematic errors reported in the Ion Torrent data in earlier studies (Bragg *et al.* 2013) were not found in this analysis of the chloroplast possibly because we used a more recent version of the technology. We can conclude that

both sequencing platforms produce high quality data with relatively low rates of discrepancy in the calling of polymorphisms, especially SNPs, but the Ion Torrent data would be less reliable for indel calls.

We compared the two sequencing platforms in generating an accurate consensus chloroplast sequence by comparing the mapping-consensus sequences generated for both genotypes. For the R-rice genotype, no mismatches were observed indicating the two platforms to be comparable for SNP calls as has been reported for microbial genomes (Quail *et al.* 2012). When comparing the mapping-consensus sequences of the W-rice-genotype we identified twenty discrepancies. It must be noted that some of the discrepancies were caused by misalignment of reads and as has been previously recognized (Albers *et al.* 2011), there is an important issue with mismatches in close proximity of indel events. Some read mapping tools can have problems with these regions resulting in false SNPs calling. However, for both the rice genotypes we observed erroneous indel calls with Ion Torrent reads mainly in homopolymer regions.

The importance of obtaining a *de novo* sequence was clear from our study as discrepancies observed in the mapping sequences of the W-rice-genotype were curated using the *de novo* contigs. The *de novo* assembly of chloroplast genomes using the tools applied in this study was easier with the longer Ion Torrent data than with the Illumina data. Although increasing number of reference genome sequences are being generated to support chloroplast genome analysis, the ability to generate whole chloroplast genome sequences *de novo* will find wide application. The first report of total plant DNA analysis for chloroplast sequencing (Nock *et al.* 2011) used short Illumina reads (36 bp) and relied on a reference genome for successful assembly. More recent studies based upon longer reads (Ferrarini *et al.* 2013; McPherson *et al.* 2013) have reported more success with *de novo* assembly.

This study demonstrates that *de novo* assembly of an accurate whole chloroplast genome sequence will be possible for routine plant barcoding. Analysis of plants based upon appropriate and careful analysis of shotgun sequencing of total DNA promises to provide a barcode that will have wide application. Having a well curated and reliable consensus chloroplast sequence for a plant sample provides greater certainty of obtaining reliable results in many critical studies such as plant identification, purity assessment, phylogenetic analysis and heteroplasmy analysis. Continuing improvements in sequencing platforms and analysis tools will make this method more reliable and cost effective for a

wide range of research and industrial applications. The significant advances in these competing sequencing platforms that have been foreshadowed by the manufacturers promise dramatic reductions in cost in the near future. This would make barcoding by sequencing the whole chloroplast the preferred option for plant identification in many research and industrial applications. A comparison of the platforms at the same level of coverage is useful in providing a guide to likely comparative performance as the number of reads on these platforms increases in the future. Costs will change for each platform as read length and read volume are increased.

However, cost is not the only advantage of this approach. The use of total DNA from the sample without amplification simplifies the analysis but also provides greater opportunity to measure in an unbiased way the contribution of different genotypes in complex mixtures. The risks of preferential amplification or enrichment of chloroplasts from a specific genotype that complicated earlier analyses will be avoided.

Preface to Chapter 4

The following chapter was published as: Brozynska M, Omar ES, Furtado A, Crayn D, Simon B, Ishikawa R and Henry RJ (2014) Chloroplast Genome of Novel Rice Germplasm Identified in Northern Australia. Trop Plant Biology **7**:111-120.

My contribution to this study was as follows: performance of the experiments (100%), data analysis (70%), writing and editing the manuscript (50%), designing the experiments (40%) and supporting filed collections (30%).

Chapter 4

Chloroplast genome of novel rice germplasm identified in northern Australia

4.1 Abstract

Rice (*Oryza sativa* L.) was probably domesticated from *O. rufipogon* in Asia in the last 10,000 years. Relatives of cultivated rice (A genome species of *Oryza*) are found in South America, Africa, Australia and Asia. These A genome species are the close relatives of cultivated rice and represent the effective gene pool for rice improvement. Members of this group in northern Australia include, an annual species, *O. meridionalis*, and two recently distinguished perennial taxa, to one of which the name *O. rufipogon* has been applied and the other a perennial form of *O. meridionalis*. Comparison of whole chloroplast genome sequences of these taxa has now been used to determine the relationships between the wild taxa and cultivated rice. The chloroplast genomes of the perennials were both found to be distinguished from *O. rufipogon* from Asia by 124 or 125 variations and were distinguished from each other by 53 variations. These populations have remained isolated from the overwhelming genetic impact of the large domesticated rice populations in Asia and may be unique descendants of the gene pool from which domesticated rice arose. The conservation of this wild genetic resource may be critical for global food security.

4.2 Introduction

The evolution of rice within the grasses has been the subject of considerable research effort. The grasses are considered by some to be of Gondwanan origin (Bouchenak-Khelladi *et al.* 2010), although there is still a body of thought that attributes a later origin and their distribution to long-distance dispersal. Kellogg (2009) uses this line of reasoning to date the rice tribe back to the start of the Miocene at 20.5 mya. Earlier publications (Clifford & Simon 1981; Simon & Jacobs 1990) have proposed a Gondwanan origin of the

grasses, particularly with reference to Australia. Recent fossil evidence, based on examination of cuticles with silica bodies (phytoliths) suggests that the Oryzeae tribe may have been distinct as early as the Late Cretaceous at 65 mya (Prasad *et al.* 2011), but Stevens (2014) questions the correct identity of these fossils and the major conflict with the data assembled from chloroplast and nuclear data. The phylogenetic relationships between *Oryza* species have been widely analysed and the origin and phylogenetic tree for rice genomes types, both diploids (AA, BB, CC, EE, FF and GG) and allotetraploids (BBCC, CCDD and HHJJ, HHKK) is well determined (Ammiraju *et al.* 2010a; Ge *et al.* 1999; Lu *et al.* 2009). However, the more recent origin of the *Oryza* genus and the A genome clade of *Oryza* species from which rice was domesticated has not been well defined. Despite numerous attempts to resolve diversification of the A genome species [e.g. Zhu & Ge (2005), Kwon *et al.* (2006), Duan *et al.* (2007), Zou *et al.* (2008)], no conclusive phylogenetic relationship of these species have been generated from those studies, nor has the method or timing of these events been defined. Regarding distribution, if the Gondwanan explanation is invoked, the A genome species have had their present distribution for a much longer time than if long-distance dispersal is used to explain how very similar species occur in four continents.

The greatest concentration of diversity in *Oryza* is found in South East Asia and northern Australia (Vaughan 1989). Both annual and perennial species of wild rice, belonging to the A genome type, have been identified (Vaughan *et al.* 2008). In Asia, the perennial species is referred to as *O. rufipogon* while the annual form is referred to as a distinct species *O. nivara* (Oka 1974; Sharma 2003). Four *Oryza* species are currently recognized from tropical Australia with two of these, one an annual and the other a perennial, belonging to the A genome clade (Henry *et al.* 2010). The perennial wild rice in Australia is referred to as *O. rufipogon*, while the annual species is known as *O. meridionalis*. Waters *et al.* (2012) used whole chloroplast genome sequencing to show that the Australian and Asian wild rices, including both the annual and perennials, were distinct from each other. However, only one accession of the Australian perennial wild rice, referred to as *O. rufipogon*, was analysed by Waters *et al.* (2012). Recent studies by Sotowa *et al.* (2013) identified two distinct types within the Australian perennial wild rice populations, suggesting that the Australian populations that had been identified as *O. rufipogon* consist of two different types of perennial wild rices that may be distinct species. One of the perennial populations was referred to as the m-type as it is morphologically similar to *O. meridionalis* (closed panicles and short anthers) and shared

two loci in the chloroplast genome. This population, here referred to as Taxon B, may have a common ancestor with *O. meridionalis*. Chloroplast gene sequence data (Sotowa *et al.* 2013) suggested that this was the type included in the study by Waters *et al.* (2012) then referred to as Australian *O. rufipogon*. Sotowa *et al.* (2013) combined molecular and morphological evidence to suggest that this perennial was a new taxon distinct from both the annual *O. meridionalis* and from the Asian *O. rufipogon*. The other Australian perennial type wild rice population, referred to as the r-type (Sotowa *et al.* 2013), was morphologically similar to *O. rufipogon* suggesting a possible common origin (Figure 4.1). We now report the complete chloroplast genome sequence for an accession of this second *O. rufipogon*-like Australian wild rice (Taxon A) and comparison with the complete chloroplast sequence of other wild and domesticated rices. This study aims to further clarify the relationships between the Australian and Asian wild *Oryza* from the A genome clade by comparing the whole chloroplast genomes of all of the known taxa for the first time.

Figure 4.1 Australian perennial wild rice (Taxon A) growing in its natural habitat in northern Queensland, Australia. Open panicles and anther length (>3-7.4 mm) affirm its morphological resemblance to *Oryza rufipogon*.

4.3 Materials and Methods

4.3.1 Plant material

The wild rice plant used in the study (Taxon A) was collected from a population of perennial wild rice at Abattoir Swamp Environmental Park in North Queensland, Australia (S16.38.085, E145.19.366). This group of wild rice plants, illustrated in Figure 4.1, was described by Sotowa *et al.* (2013) as *Oryza rufipogon*-like (r-type) and was found at the Jpn1 site. In this study it is referred as “wild rice Taxon A”.

4.3.2 DNA extraction, sequencing and data analysis

The chloroplast genome of a wild rice Taxon A collected from northern Australia was sequenced using next generation sequencing of total DNA. DNA from leaf tissue of a wild rice plant was extracted using a modification (Carroll *et al.* 1995) of the CTAB method (Bernatzky & Tanksley 1986) as described by (Furtado 2014). Illumina HiSeq2000 (Illumina, San Diego, CA, USA) and Ion Torrent (PGM, #318 chip) sequencing platforms were used for shotgun sequencing of total genomic DNA from the sample. A consensus sequence of the wild rice Taxon A chloroplast genome was created and validated using the following approaches: (1) Illumina and Ion Torrent read mapping to the reference chloroplast sequence of *Oryza sativa* ssp. *japonica* var. Nipponbare, (2) *de novo* Illumina reads assembly and subsequent contigs mapping to the reference, and (3) reference-assisted reads assembly from Ion Torrent platform followed by ordering the contigs based on the *O. sativa* reference sequence. All the Illumina analyses were performed on CLC Genomics Workbench 6.0 (www.clcbio.com) and Ion Torrent reads were alternatively imported to Torrent Suite Software version 3.6. Default programmes from both software platforms were used for quality analysis, read trimming, read mapping and *de novo* or reference-based read assembly.

4.3.3 Functional annotation

Wild rice Taxon A chloroplast genome annotation was performed using the Bacterial Genome Annotation system BG7 (Pareja-Tobes *et al.* 2012) followed by manual curation of exon/intron boundaries and start/stop codons.

The full circular chloroplast gene map was visualised using the OrganellarGenomeDRAW tool (Lohse *et al.* 2013).

4.3.4 Phylogenetic analysis

Eight fully sequenced chloroplast genomes of diploid rice species were used to perform a phylogenetic study (Table 4.1). All of the species, but one, were A genome rice relatives. *Oryza australiensis*, belonging to EE rice genome group, was included as an outgroup as this species is less closely related to the other *Oryza* genus species analysed in this study.

The consensus chloroplast sequence of the wild rice Taxon A was aligned with other publicly available chloroplast genomes of species from the genus *Oryza* (GenBank, www.ncbi.nlm.nih.gov). The following sequences were used in the study: *O. sativa* spp. *japonica* var. Nipponbare, *O. sativa indica* isolate 93-11, Australian perennial sample used by Waters *et al.* (2012) and referred therein as *O. rufipogon* [referred in the present study as "wild rice Taxon B" due to its similarity to populations relating to the *O. meridionalis*-like perennials from the Jpn2 site in Sotowa *et al.* (2013)], Asian *O. rufipogon*, *O. meridionalis*, *O. nivara*, and *O. australiensis*. A summary of the genome groups, sequence lengths and GenBank accession numbers are shown in Table 4.1. The multiple genome alignment was conducted using Mauve 2.3.1 software and the progressive Mauve algorithm (Darling *et al.* 2010) with default parameters. Prior to the analysis, one copy of the inverted repeat (IR) sequence was deleted from the chloroplast sequence of each of the genomes. The length of the IR ranged between 20,792 and 20,803 bp depending on the species.

Table 4.1 Summary of rice chloroplast sequences used in phylogenetic analysis and comparative study: their genome group, full chloroplast length, length without one IR sequence and GenBank accession number.

Rice species	Genome group	Full length chloroplast sequence (bp)	Sequence length without IR (bp)	Accession number
Taxon A	AA	134,557	113,754	KF428978
O. sativa japonica	AA	134,551	113,749	GU592207
O. sativa indica	AA	134,496	113,698	AY522329
O. meridionalis	AA	134,558	113,755	JN005831
Taxon B	AA	134,557	113,754	JN005833
O. rufipogon (Asia)	AA	134,544	113,743	JN005832
O. nivara	AA	134,494	113,692	AP006728
O. australiensis	EE	134,549	113,749	GU592209

IR – inverted repeat.

The phylogenetic tree reconstruction for chloroplast genome alignment was performed using three distinct methods: maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Selection of the best-fit model of nucleotide substitution was conducted using jModelTest2 (Darriba *et al.* 2012) software and Akaike information criterion. The model chosen for maximum likelihood calculation was 012010+I+G+F (I = 0.8990). MP and ML analyses were completed in PAUP* 4.0 software package (Swofford 2003) choosing a heuristic search for finding the optimal tree, the random stepwise addition procedure for obtaining a starting tree and with the tree bisection-reconnection (TBR) algorithm for branch-swapping. 200 random-addition sequence replications were performed, as well as 2,000 bootstrap pseudoreplications to measure group support (frequency of occurrence). Gaps were considered as missing data. All characters were treated as unordered and weighted equally.

Bayesian analysis was performed using MrBayes version 3.2 (Ronquist *et al.* 2012) through the Geneious 6.1.2 software platform (www. geneious.com). The evolutionary model used was the General Time Reversible Model with gamma-shaped among-site rate variation with an estimated proportion of invariable sites (GTR+I+G; I = 0.8990). The branch length prior was set to exponential with parameter 10.0. Two independent and simultaneous analyses starting from diverse random trees were performed. Monte Carlo Markov Chains (MCMC) were run for 1×10^6 generations, with chains sampled every 200 generations, followed by burn-in of 1×10^5 MCMC. Three heated chains and one cold chain were used with the heating coefficient of 0.2 (by default). Consensus nodal support was assessed by posterior probability distribution. All trees were rooted using the outgroup method.

4.3.5 Distances between rice chloroplast genomes and comparative chloroplast genomics

Each of the full length rice chloroplast genomes was imported into the Geneious 6.1.2 software platform and the sequences were aligned using the alignment tool available in the package, followed by variant analysis using the variant/SNP detection tool. In order to create a distance matrix for chloroplast genomes all the variants, as well as the differences (i.e. number of non-identical bases) were counted. The variants between *Oryza* Taxon A and other rice chloroplast were counted once, regardless of their length.

4.4 Results

4.4.1 Chloroplast genome features of wild rice Taxon A

The plastid genome of Taxon A consists of the four typical components found in angiosperms, i.e. two inverted repeats regions, IRA and IRB (both 20,803 bp in length), a large single copy (LSC; 80,604 bp) and a small single copy (SSC) region (12,347 bp; Figure 4.2). Overall 162 genes were found in the chloroplast genome, including 114 protein-coding genes, 40 transfer RNAs and eight ribosomal RNAs. Twenty CDS and eight tRNAs were duplicated owing to the location in both of the inverted repeats. rRNAs genes were located in the IR forming two operons of 23S, 4.5S and 5S, and two separate 16S subunits. A total of 97 genes were single copy in the wild rice Taxon A chloroplast genome (75 protein-coding genes and 24 tRNAs). A functional description and gene visualisation is presented in Figure 4.2. Eleven genes contained one intron: six tRNA genes (*tRNA-Lys*, *tRNA-Gly*, *tRNA-Leu*, *tRNA-Val*, *tRNA-Ile* and *tRNA-Ala*) and five protein-coding genes (*rps16*, *atpF*, *rpl2*, *ndhB* and *ndhA*). The genome is AT-rich with an AT content of 61%. Coding sequences occupy 58.29% of the genome, comprising protein coding regions (49.24%), rRNA (6.83%) and tRNA (2.22%). Non-coding sequences contain intergenic regions and introns which represent 33.01% and 8.70% of the chloroplast genome, respectively.

4.4.2 Phylogenetic analysis

The multiple genome alignment of rice chloroplast sequences was 113,960 bp in length. One of the inverted repeats was excluded prior to phylogenetic analysis due to the identical sequence of these repeats. This avoids any parsimony informative sites in the inverted repeat regions being weighted with twice the value relative to other informative sites in the alignment. Of the total number of bases which were subjected to MP analysis, 112,969 were constant, 903 were variable and parsimony uninformative and the number of parsimony informative characters was 88. The most parsimonious tree was obtained after 798 rearrangements during a heuristic search and was 1,014 in length with consistency index CI = 0.98 and retention index RI = 0.89. The CI excluding uninformative characters was 0.84. The score for the best tree (-lnL) under ML criterion was 162,388.46454 after 142 rearrangements. Bootstrap nodal support calculated from both MP and ML was strong (>99.2%) except for the node of Australian wild rice Taxon B and *O. meridionalis* where the support was slightly lower (83% in ML and 85% in MP). Posterior probabilities of the

tree resulted from BI were all 100%. The optimal trees found by all three phylogenetic methods were consistent and the final tree is presented in Figure 4.3.

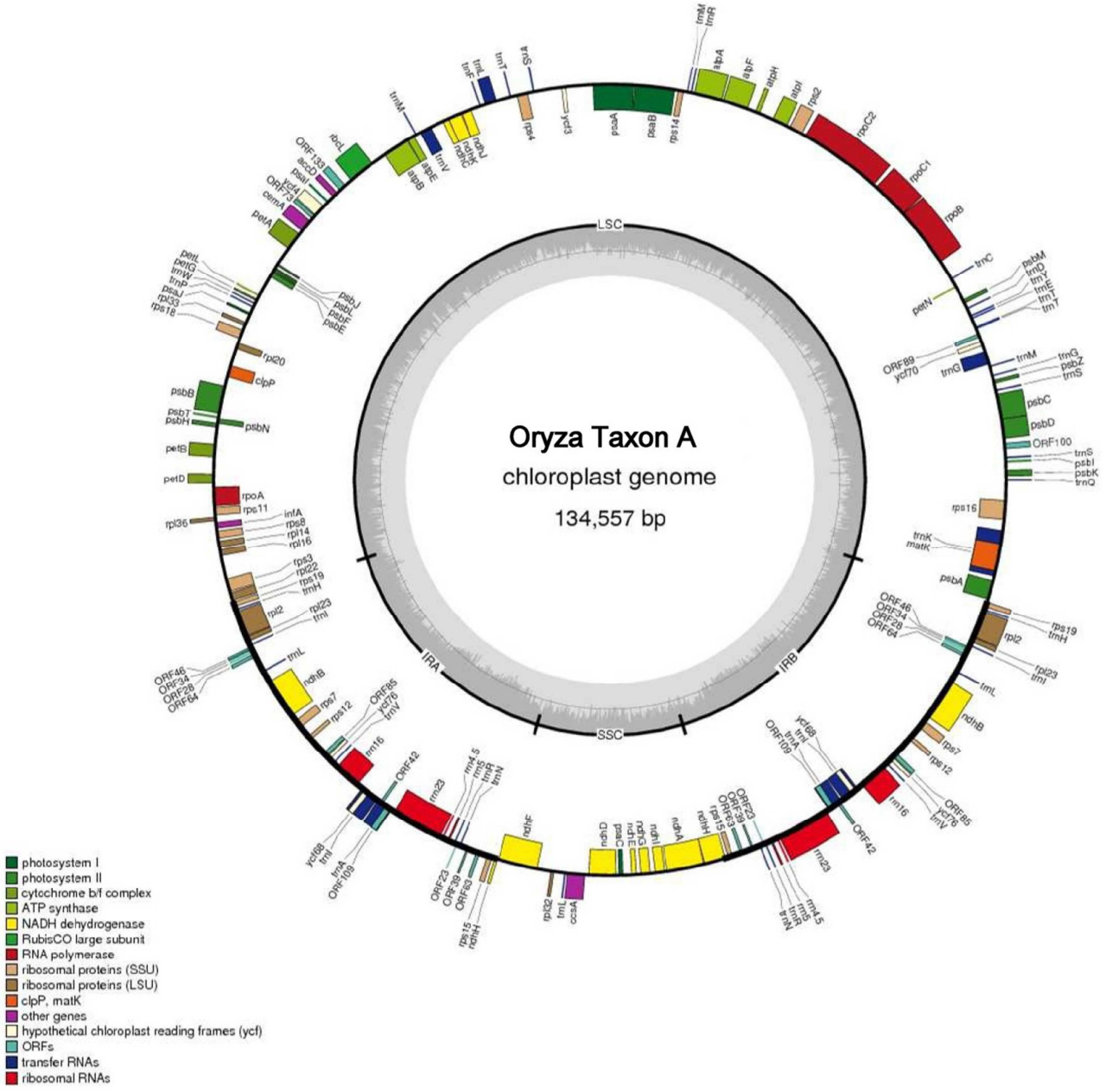


Figure 4.2 A chloroplast genome map of an Australian perennial wild rice (Taxon A). The inner circle indicates the boundaries between LSC, IRA and IRB, and SSC. The dashed area shows the GC content of the genome and marks the 50% threshold with additional grey line inside. Genes are transcribed either clockwise (those found outside the outer circle) or counter clockwise (inside the outer circle). Features which belong to diverse functional groups are marked with different colours.

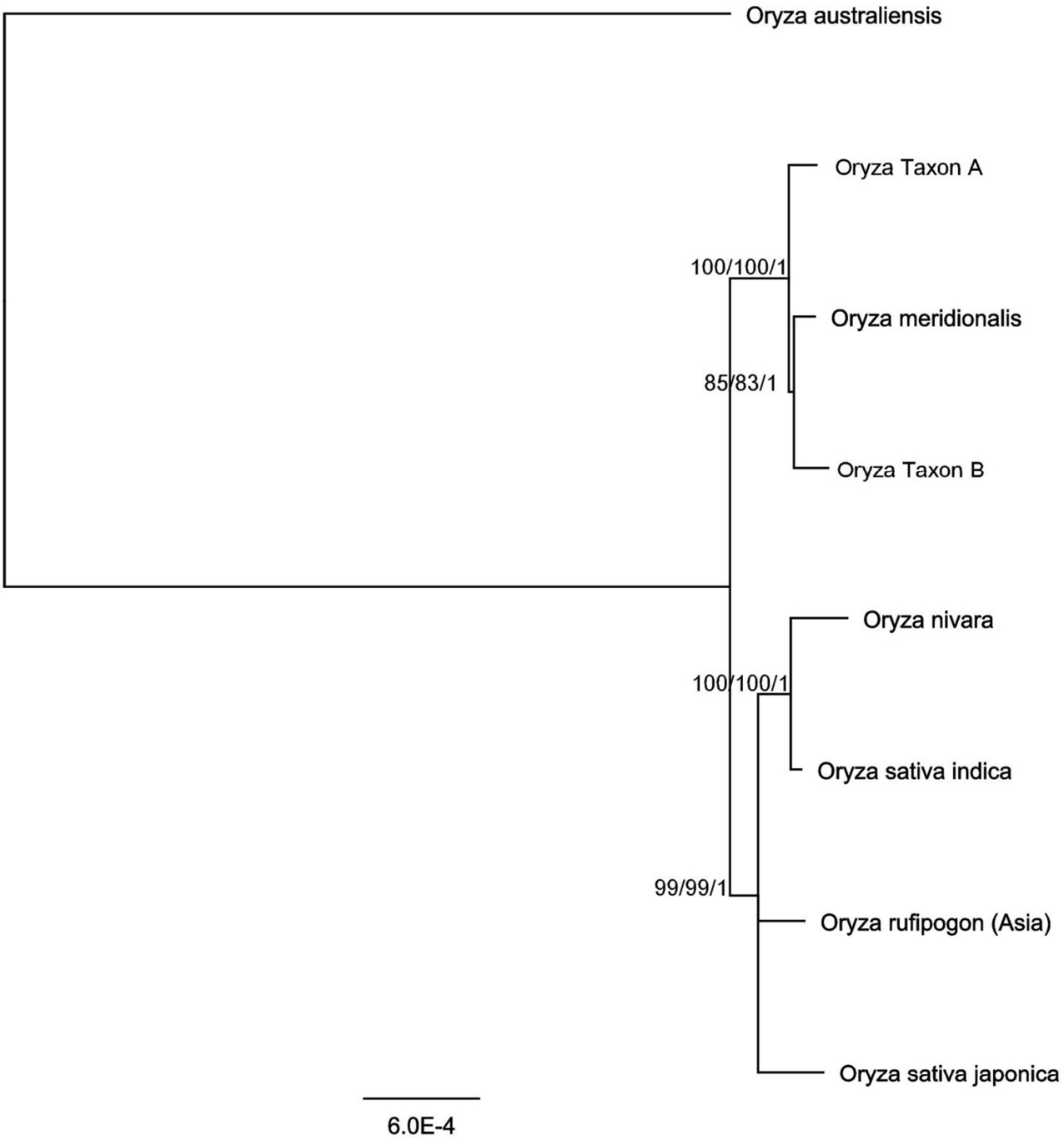


Figure 4.3 Phylogenetic relationships between chloroplast genomes of *Oryza* species. Nodal support is presented as precent bootstrap in MP/precent bootstrap in ML/BI probability. Scale bar is the number of substitutions per site.

4.4.3 Distances between rice chloroplast genomes

Distances between the individual chloroplast genomes from all of the species studied were represented as the number of variants between each of them (variant represented a SNP or MNV regardless of its length). Additionally the nucleotide differences in the sequences (i.e. number of bases which are not identical) were also determined (Table 4.2).

Table 4.2 Distance matrix of rice species chloroplast genomes.

	O. australiensis	*O. sativa japonica*	*O. sativa indica*	*O. rufipogon* (Asia)	*O. nivara*	*O. meridionalis*	Taxon B	Taxon A
O. australiensis		838	845	846	876	860	870	861
O. sativa japonica	854		76	80	104	122	135	125
O. sativa indica	1028	249		62	57	110	122	122
O. rufipogon (Asia)	896	118	232		95	111	124	125
O. nivara	1054	272	138	259		141	152	143
O. meridionalis	908	159	290	148	317		34	36
Taxon B	914	168	310	177	335	41		53
Taxon A	964	222	350	217	381	124	122	

The top panel represents number of variants between species (a variant corresponds to a single nucleotide- or a multi nucleotide-variant at any given position). The bottom panel shows number of nucleotide differences in the sequences (numbers of non-identical bases). The heatmap corresponds to individual values in a matrix as colours: red cells indicate close distance; orange, yellow and green cells show gradually greater distances.

The lowest number of variants within the Australian clade, and also within all the species studied, was found between wild rice Taxon B and *O. meridionalis* (34 variants). Interestingly, there were more variants between wild rice Taxon B and wild rice Taxon A (53 variants) than between *O. meridionalis* and each of these perennial species (34 and 36 variants, respectively). The Australian wild rice Taxon A analysed in the present study had a considerably lower number of differences in comparison with the other Australian A genome species than with the Asian species. Within the Asian clade all of the species' chloroplast genomes had the lowest number of polymorphisms with *O. sativa* spp. *indica* in the following order: *O. nivara* (57), *O. rufipogon* (62) and *O. sativa japonica* (76). As

expected, considerably greater distances were present between *O. australiensis* and all other rice species.

4.4.4 Comparative chloroplast genomics

Table 4.3 shows the number and types of variants in the chloroplast genomes of the Australian wild rice and other rice species used in this study. The most abundant variation types in all species were SNPs. The chloroplast genome is mainly composed of coding region followed by intergenic region while a small component is comprised of introns (Figure 4.4A). However the distribution of variants as a proportion is mainly in the intergenic regions (66-75% variants) followed by coding region (17-30% variants) and a small component in introns (5-8% variants; Figure 4.4B). The majority of polymorphisms were found in the intergenic regions which occupy only 33.01% of the chloroplast genome, almost two times less than the coding sequences (58.29%).

Table 4.3 Total number and type of sequence variants in rice species chloroplast genomes in comparison with Taxon A.

Rice species	Variant number	Deletions	Insertions	MNV	SNPs (Ts/Tv)
O. meridionalis	36	4	10	3	19 (8/11)
Taxon B	53	6	9	5	33 (16/17)
O. sativa indica	122	18	17	10	77 (39/38)
O. rufipogon (Asia)	125	18	14	7	86 (40/46)
O. sativa japonica	125	14	17	4	90 (47/43)
O. nivara	143	21	20	10	92 (46/46)
O. australiensis	861	14	16	20	811 (518/293)

MNV: multi-nucleotide variant; Ts and Tv – transitions and transversions, respectively.

The SNPs found in coding regions of chloroplast genomes were analysed based on their effects on protein sequences. Synonymous and non-synonymous substitutions, as well as frame shifts were discovered among genomes (Table 4.4). In order to acquire a better understanding of the variation or differences present in protein coding regions between the Australian and Asian *Oryza* clades, as well as the unique occurrences in the Australian wild rice, the differences were grouped together based on their origin (Table 4.5). Sixteen SNPs and two indels were found to be specific to either the Australian or Asian clade. Six out of them resulted in amino acid substitutions. Four SNPs were unique to the wild rice (Taxon A) and one of them caused an amino acid substitution of asparagine to serine in *rpoC1* gene (DNA-directed RNA polymerase subunit beta'). All of

the proteins with variants were blasted against non-redundant protein sequences (nr) on the NCBI database (http://blast.ncbi.nlm.nih.gov/) to investigate the uniqueness of the substitutions in the chloroplasts of Asian and Australian rices, as well as in other plants. Among the substitutions analysed, two were found to be exclusive to the Australian clade, one to the Asian, and one in Taxon A and two were found in numerous plant species (Table 4.6).

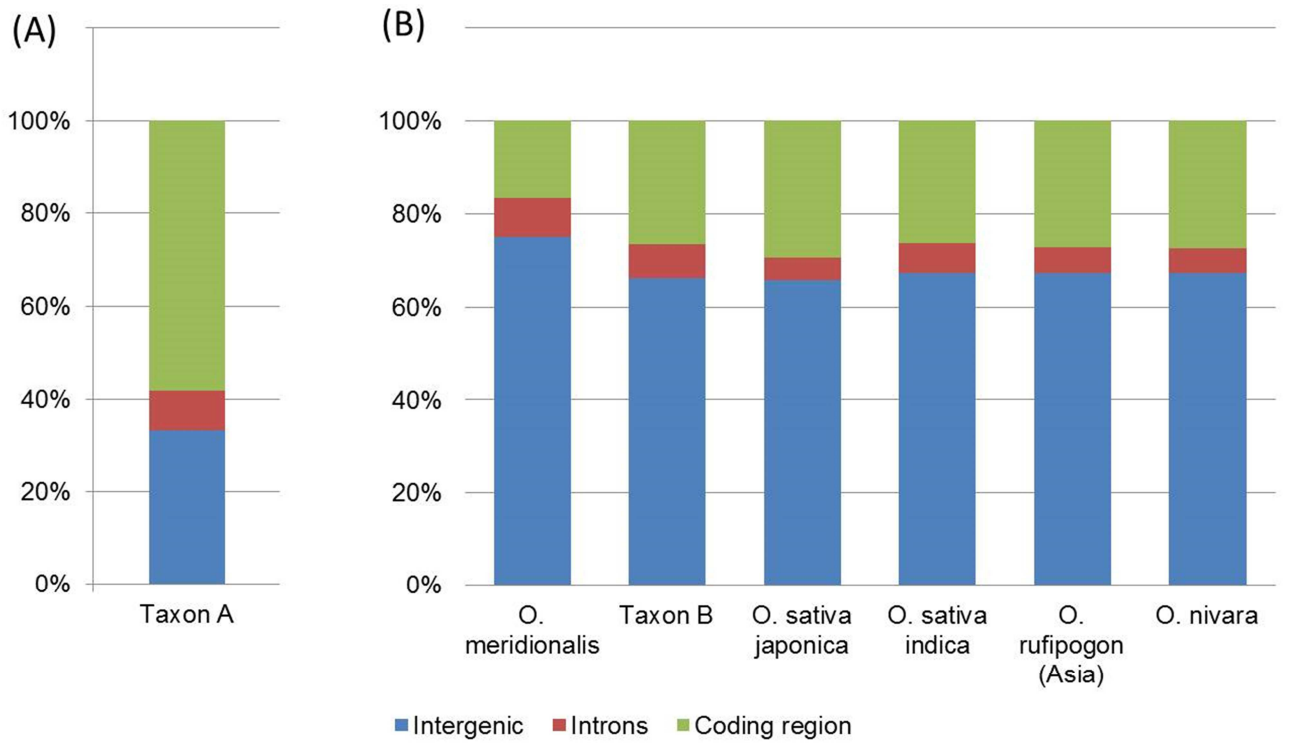


Figure 4.4 Analysis of coding and non-coding regions in chloroplast genomes. (A) Distribution of sequence types in *Oryza* Taxon A and (B) proportion of variants relative to *Oryza* Taxon A found in each of these regions.

Table 4.4 SNPs and frame-shift events in coding region of the chloroplast genomes of rice species in comparison with Taxon A.

Rice species	Synonymous SNPs	Non-synonymous SNPs	Frameshifts
O. meridionalis	5	1	0
Taxon B	6	7	0
O. sativa japonica	22	11	4
O. sativa indica	17	9	4
O. rufipogon (Asia)	16	15	3
O. nivara	22	11	3

SNP – Single Nucleotide Polymorphism.

There was also a frame-shift found in one of the chloroplast encoded proteins of unknown function, Ycf68, which is duplicated in the chloroplast genome due to its location in the inverted repeats (the frameshift was present in both copies as these regions are perfect repeats). The frameshift is produced by a single base insertion in the Australian chloroplast genomes that causes an early stop codon and protein truncation (in Asian genomes it is 133 amino acids in length, in Australian 101 amino acids; data not shown). Despite the premature termination of the gene *ycf68* the functional protein domain (which lies in N-terminal end) is still present in the protein sequence (Uncharacterised protein family Ycf68, InterPro: IPR022546). Also, proteins of similar length were found in other plant species, e.g. *Triticum monococcum* ssp. *aegilopoides* (GenBank accession: AGP51191) and *Setaria italica* (Doust *et al.* 2009).

The chloroplast genome lengths vary in the *Oryza* genus between 134,494 bp (*O. nivara*) and 134,558 bp (*O. meridionalis*). The length of the wild rice Taxon A chloroplast genome was 134,557 bp that corresponds to the usual *Oryza* chloroplast length and to the standard sizes of plastid genomes described so far [120-160 kbp; Green (2011)]. The chloroplast genome of the Australian wild rice Taxon A is AT-rich as reported for other angiosperms (Raubeson *et al.* 2007).

Table 4.5 Polymorphisms in chloroplast coding regions found to be specific to Australian or Asian clades or Taxon A.

Position	Australian clade	Asian clade	Taxon A	Gene	Protein effect
LSC					
448	G	A	G	*psbA*	*
817	G	A	G	*psbA*	*
2,220	G	T	G	*matK*	Ala → Glu
3,067	G	T	G	*matK*	Leu → Ile
4,546	A	T	A	*rps16*	*
9,321	G	G	***A***	*psbD*	*
24,164	A	A	***G***	*rpoC1*	Asn → Ser
24,742	T	G	T	*rpoC2*	His → Gln
25,000	G	A	G	*rpoC2*	*
27,949	C	A	C	*rpoC2*	*
29,885	T	C	T	*rps2*	*
55,325	G	A	G	*rbcL*	*
IR_A					
90,581	G	T	G	*ycf76*	*
93,538	C	del	C	*ycf68*	frameshift
SSC					
104,506	T	T	***C***	*rpl32*	*
105,927	C	A	C	*ccsA*	Ser → Tyr
106,567	G	T	G	*ndhD*	Ala → Glu
106,710	G	T	G	*ndhD*	*
110,849	G	A	G	*ndhA*	*
113,255	G	G	***A***	*ndhH*	*
IR_B					
121,627	G	del	G	*ycf68*	frameshift
124,581	C	A	C	*ycf76*	*

Positions refer to the positions in Taxon A chloroplast genome. Amino acid changes refer to a change with respect to Australian clade and Taxon A with the exception of position 24,164 (Ser is exclusive for Taxon A). Australian clade includes: *O. meridionalis* and Taxon B; Asian clade includes: *O. sativa japonica* and *indica*, *O. rufipogon* (Asian origin) and *O. nivara*. Nucleotides marked in bold and italics are found to be unique to Taxon A. LSC – long single copy, SSC – short single copy, IR_A and IR_B – inverted repeats, del – deletion, * – synonymous SNP.

Table 4.6 Predicted amino acid substitutions in chloroplast encoded proteins between Australian and Asian clades of *Oryza*.

Position	Australian clade	Asian clade	Taxon A	Gene	Uniqueness of substitution
46	Leu	Ile	Leu	*matK*	Leu exclusive for Australian clade
328	Ala	Glu	Ala	*matK*	known
567	Asn	Asn	Ser	*rpoC1*	Ser exclusive for Taxon A
10	His	Gln	His	*rpoC2*	Gln exclusive for Asian clade
224	Ser	Tyr	Ser	*ccsA*	Ser exclusive for Australian clade
452	Ala	Glu	Ala	*ndhD*	known

Positions refer to the amino acids positions in given proteins from Taxon A; known – given substitution was found among other genes from chloroplast genomes of numerous plant species; exclusive for Australian/Asian clade – given substitution was found only in Australian or Asian clades, respectively; exclusive for Taxon A – given substitution was found only in Australian wild rice (Taxon A).

4.5 Discussion

Long indels found between *O. nivara* and *O. sativa indica* and the other rice species (data not shown) affected the distance analysis among rice chloroplast genomes increasing the number of different nucleotides and consequently the distance from other rice relatives (Table 4.2). However, the presence of indels did not affect the phylogeny which placed these two species together within the Asian clade, as has been reported by others (Huang *et al.* 2012b), and indicates that it is very likely that each of those indels was one separate evolutionary event.

The most dominant cultivated rice species (*Oryza sativa*) belongs to the A genome group, and was most likely domesticated from wild populations of *O. rufipogon* in Asia. The A genome group is the most recently diverged group in the *Oryza* genus and is comprised of eight diploid species, with annual and perennial types, and distributed worldwide including Australia (Ge *et al.* 2001; Vaughan 1989). The Australian perennial wild rice of the A genome type and generally referred to as Australian *O. rufipogon* (Henry *et al.* 2010; Vaughan *et al.* 2008), has recently been shown to be comprising of at least two types, the m-type and the r-type based on the analysis of loci in the nuclear genome (Sotowa *et al.* 2013). In addition, Sotowa *et al.* (2013) examined many accessions of these taxa and reported that all wild type perennials found in Australia had *meridionalis*-type plastid genomes on the basis of analysis of two loci within the chloroplast. In our study, the best phylogenetic tree obtained confirms the distinctness of Australian and Asian rice relatives

as reported by Waters *et al.* (2012). Moreover, the phylogenetic analysis placed the perennial wild rice Taxon A within the Australian clade and also supported earlier findings about perennial wild rice Taxon B (Waters *et al.* 2012). Most importantly, this study shows for the first time the distinctness of the Australian Taxon A from *O. rufipogon* and despite superficial morphological resemblance, the chloroplast genome shows 125 variations. This indicates that the genetic difference between Taxon A and *O. rufipogon* (125) is comparable with that between Taxon A and *O. sativa japonica* (125) and that between *O. rufipogon* and *O. sativa japonica* (118). This perspective suggests that the Australian perennial, *Oryza* Taxon A, is a similar genetic distance from both domesticated Asian rice (*O. sativa*) and the wild Asian perennial rice (*O. rufipogon*) and that this is a similar genetic distance to that found between the wild and domesticated Asian species. The divergence between *O. rufipogon* and *O. sativa* was probably driven by human selection in the last 10,000 years while the divergence of *O. rufipogon* in Asia and *Oryza* Taxon A in Australia was likely to have happened over a much longer period and been driven by allopatric evolution.

The Australian wild rice species previously reported (*O. meridionalis* and the Australian *O. rufipogon* referred here as wild rice Taxon B) were found to be more closely related to each other than the other wild relative (Taxon A) studied here. Based on the distance analysis, it can be concluded that rice species within the Australian clade are generally more closely related to each other than the species within the Asian clade which is also reflected in the phylogenetic tree (Figure 4.3). The number of differences in the chloroplast between *O. nivara* and *O. sativa indica* (138), and *O. rufipogon* and *O. sativa japonica* (118) is similar to the distance between *O. meridionalis* and *Oryza* Taxon A (124). Based on chloroplast analysis, the relationship between the two morphologically distinct Australian perennial wild *Oryza* species and *O. meridionalis* was relatively close. Both are more closely related to *O. meridionalis* than the two cultivated subspecies of rice *O. sativa*, ssp. *japonica* and ssp. *indica,* are to one another (Table 4.2). The chloroplast genome sequence of *Oryza* Taxon A confirms that it is distinct from *O. rufipogon*, *O. meridionalis* and *Oryza* Taxon B.

Nuclear genomes not analysed here may provide further information on the relationships between these species. Sotowa *et al.* (2013) analysed INDEL and SSR markers in the nuclear genome of other wild rice perennial accessions with the morphology of Taxon A collected from the same site (Jpn1 site) revealing some individuals to be closer to *O. meridionalis* and others to *O. rufipogon* at the loci tested. Analysis of the

complete nuclear genomes may provide further information on the relationships between these Australian perennial accessions and their relatedness to cultivated rice.

The chloroplast genome sequences used here provide a guide to the relationships between the taxa but analysis of the nuclear genomes could result in greater insights into the evolution of the A genome clade of *Oryza*. Earlier studies explored some possible relationships between these taxa on a morphological basis (Sotowa *et al.* 2013). Despite some superficial morphological similarities, these two Australian wild rice perennial lineages, the r-type (Taxon A) and the m-type (Taxon B), have now been characterised as considerably divergent from both, *O. rufipogon* and *O. meridionalis*, and especially from one another at organellar, nuclear and morphological levels.

This discovery adds further weight to present a case for South East Asia and northern Australia being considered as the centre of origin for the A genome. While the tribe is possibly of Gondwanan origin, the divergence of the A genome species is likely to be a relatively recent event involving long distance dispersal throughout the tropical world. More collections are required to determine the distribution of this species and to define the genetic variation within the population. This becomes more urgent with the prospect of agricultural expansion in tropical Australia and with rice cultivation becoming more likely. The wild rice populations in Australia have to date been isolated from the impact of genetic contamination by variants of cultivated rice as has been reported to have occurred widely in Asia with *O. rufipogon*. Pressure for more rice to satisfy food security could see extensive rice cultivation occur in Tropical Australia. The conservation of this wild genetic resource may require efforts in both *ex situ* and *in situ* conservation. The perennial A genome rice species from northern Australia are expected to be a valuable new genetic resource for rice improvement, with potential to contribute novel disease resistances, environmental stress tolerances and possible nutritional value.

Preface to Chapter 5

The following chapter was submitted to *Plant Biotechnology Journal* as: Brozynska M, Copetti D, Furtado A, Wing RA, Crayn D, Fox G, Ishikawa R and Henry RJ (2016) Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice.

My contribution to this study was as follows: performance of the experiments (100%), data analysis (90%), writing and editing the manuscript (50%) and designing the experiments (40%).

The set of putatively single copy orthologs (6,015 genes) used in the phylogenetic inference between Australian wild rice taxa and other A genome *Oryza* species was provided by Prof. Rod Wing from Arizona Genomics Institute, The University of Arizona, Tucson, AZ, USA. This set was generated as a part of a study under the international initiative I-OMAP – International *Oryza* Map Alignment Project.

Chapter 5

Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice

5.1 Abstract

Rice is the most important food crop in the world. The related A genome species of the *Oryza* genus are the effective gene pool for rice. Here we report draft genomes for two Australian wild rice taxa: *O. rufipogon*-like population, referred to as Taxon A and *O. meridionalis*-like population, referred to as Taxon B. These two taxa were sequenced and assembled by integration of two distinct next generation sequencing (NGS) data. This study creates a genomic platform for investigating a new wider gene pool and agriculturally important traits in Australian taxa. Recently, based upon chloroplast genome sequences, the Australian A genome species were shown to be a sister clade to the group that includes domesticated rice. Here we report that, despite the distinct chloroplast genome, the nuclear genome of the Australian Taxon A, has a sequence that is much closer to that of domesticated rice (*O. sativa*) than to the other Australian wild populations. Analysis of 4,643 genes in the A genome clade showed that the Australian annual, *O. meridionalis*, and related perennial taxa have the most divergent (around 3 million years) genome sequences relative to domesticated rice. A test for admixture showed possible introgression into the Australian Taxa A (diverged around 1.6 million years ago) especially from the wild *indica*/*O. nivara* clade in Asia. These results demonstrate that northern Australia may be the centre of diversity of the A genome *Oryza* and suggest the possibility that this might also be the centre of origin of this group.

5.2 Introduction

Rice is a pantropical crop that is a staple food consumed by over half of the world's population This crop has a long history of domestication and its cultivation dates back

around 10,000 years in Asia (*O. sativa*) and over 3,000 years in Africa (*O. glaberrima*). The *Oryza* genus diversified into six diploid (A-C and E-G) and five tetraploid (BC, CD, HJ, HK, and KL) genome groups. The phylogeny of the *Oryza* genome groups has been widely studied and is now well known (Ammiraju *et al.* 2010a; Ge *et al.* 1999; Lu *et al.* 2009). The relationships between the most recently diverged A genome diploids, that include domesticated rice, has been more challenging and only lately has their phylogeny been more fully described, using both chloroplast (Wambugu *et al.* 2015) and nuclear genomes (Stein *et al.*, personal communication).

Rice food security requires continued increases in rice productivity and relies on ongoing genetic improvement. Climate change adds to the difficulty of achieving the necessary rates of genetic gain (Abberton *et al.* 2016). The wild relatives of rice provide a gene pool that allows for the expansion of diversity (Krishnan *et al.* 2014) in domesticated rice for the creation of new high yielding genotypes, with new nutritional and functional traits (Kharabian-Masouleh *et al.* 2012) and adaptation to new environments (Brozynska *et al.* 2016). The A genome species of *Oryza*, which include the species that are readily inter-fertile with rice, represent the effective primary gene pool for rice. Recent investigations of large and widespread wild populations in tropical Australia (Henry *et al.* 2010) suggest the presence of two distinct and possibly novel perennial wild A genome taxa [Figure 5.1; Brozynska *et al.* (2014), Sotowa *et al.* (2013), Waters *et al.* (2012)]. Of these, Taxon A has plant and seed morphology similar to that of *O. rufipogon* and Taxon B appears to be similar to the annual *O. meridionalis*. Here we report draft genomes for these two Australian wild rice taxa: Taxon A and Taxon B, which are likely to be novel and different species. The two taxa were sequenced and assembled by integration of two distinct NGS data, namely Illumina and Pacific Biosciences. The draft nuclear genome sequences of 384.8 Mb (Taxon A) and 354.9 Mb (Taxon B) were placed on 12 pseudochromosomes based on available rice reference sequences. Taken together, this study creates a new genomic platform for investigating the gene pool and agriculturally important traits potentially present in Australian taxa.

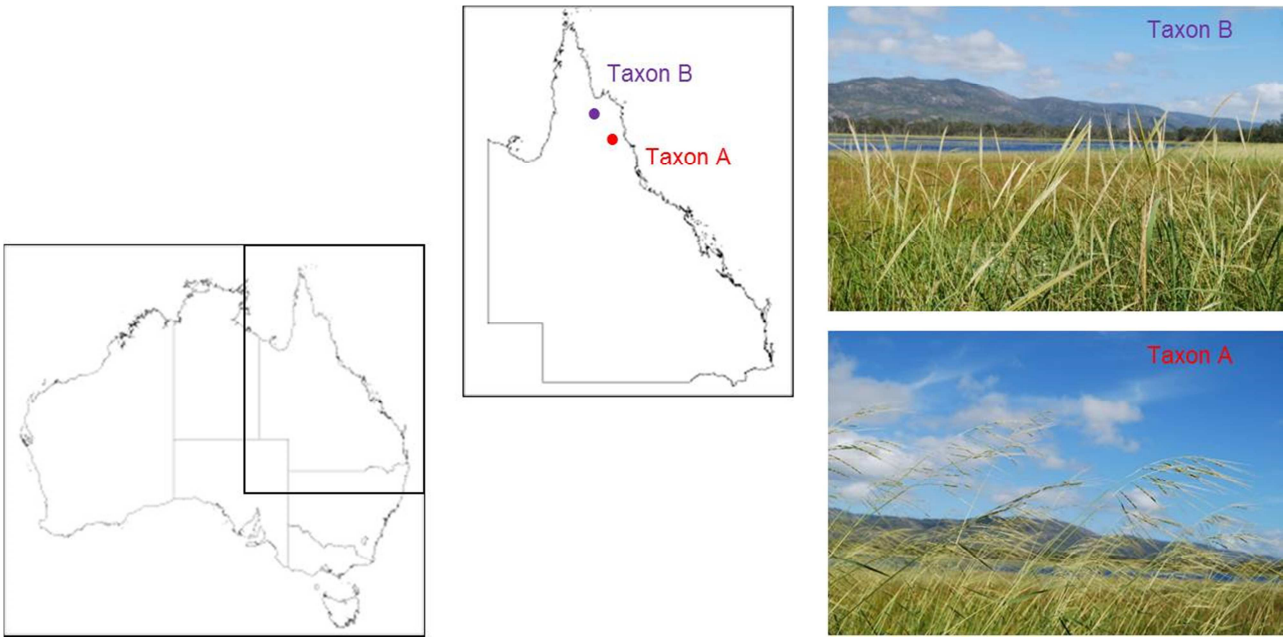


Figure 5.1 Australian perennial A genome taxa from northern Australia. Taxon A is characterised by open panicles, while Taxon B has closed panicles.

The Australian wild rices have been isolated from the impact of gene transfer from domesticated rice that may complicate interpretation of the genetics of wild rice populations in Asia where rice has been cultivated on a large scale for thousands of years. An understanding of genetic relationships and diversity between and within these Asian and Australian populations will guide the effective use of wild genetic resources for global rice improvement. The phylogenetic relationships between all of the A genome taxa have recently been estimated using whole chloroplast genome sequences (Brozynska *et al.* 2014; Wambugu *et al.* 2015). In this phylogeny the Australian A genome taxa form a distinct clade, which is a sister to the Asian domesticated rice clade (Figure 5.2A). We now report a phylogenetic analysis of the corresponding nuclear genomes.

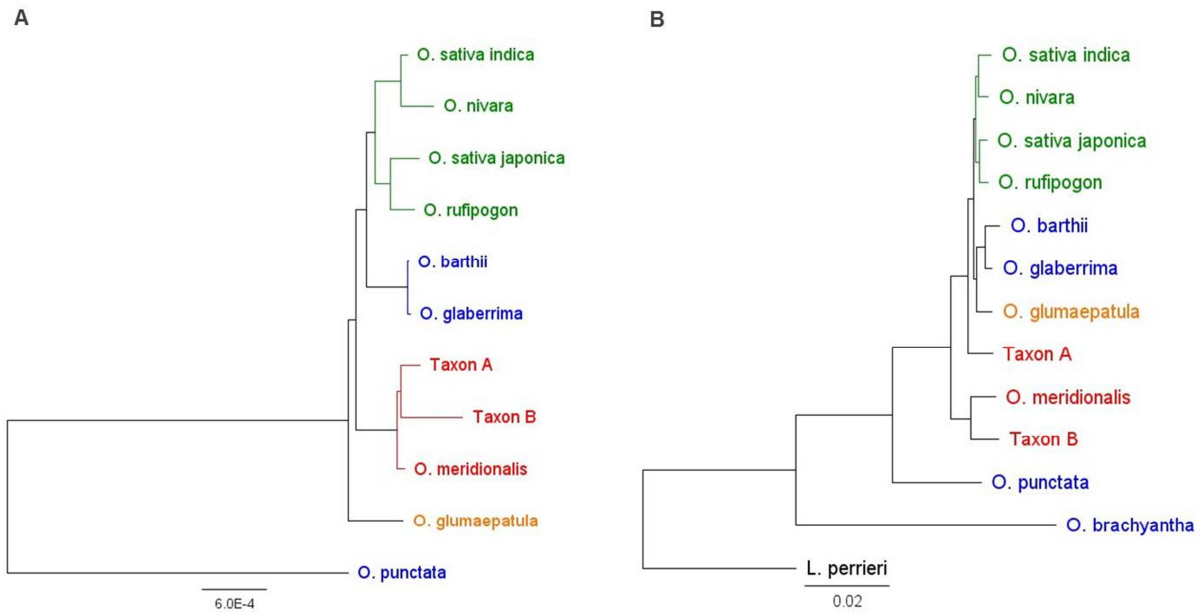


Figure 5.2 Phylogenetic relationships between A genome rice species. (A) Tree topology based upon analysis of supermatrix of 4,643 nuclear genes. (B) Tree topology based on whole chloroplast genome sequences. Figure adapted and modified from Wambugu *et al.* (2015). Taxa marked in green represent Asian rice species, in blue: African, in orange: South American, and in red: Australian. *L. perrieri* and *O. punctata* were used as outgroups in nuclear and chloroplast studies, respectively.

5.3 Methods

5.3.1 Plant material

The wild rice plants used in the study came from perennial wild rice populations in North Queensland, Australia. The first individual, referred to here as Taxon A, was collected from Abattoir Swamp Environmental Park near Julatten and was described by Sotowa *et al.* (2013) as *Oryza rufipogon*-like taxon (r-type) collected from Jpn1 site. The second individual, referred to here as Taxon B, was collected from a small wetland beside the Peninsula Developmental Road and in Sotowa *et al.* (2013) was called *Oryza meridionalis*-like (m-type) taxon collected from Jpn2 site. Specimens of these wild rice populations were collected from their natural habitats and are now kept and maintained in glasshouse conditions at The University of Queensland in Brisbane, Australia.

5.3.2 DNA extraction and sequencing

DNA from leaf tissue of Taxon A and Taxon B individuals was extracted using a modification of the CTAB method (Furtado 2014) and subsequently subjected to whole genome shotgun sequencing. Next generation sequencing platforms used were Illumina

HiSeq2000 (Illumina, San Diego, CA, USA) and Pacific Biosciences RSII with P6-C4 chemistry (PacBio, Menlo Park, CA, USA). The data generated on Illumina instrument were 101 bp reads with an average library insert of 550 bp (paired end reads, PE), 3,000 bp (mate pair reads, 3 Kb MP) and 5,000 bp (mate pair reads, 5 Kb MP). Samples for paired end sequencing were generated using TruSeq DNA PCR free library preparation kit, whereas mate pair libraries were prepared using the Nextera Mate Pair protocol. Illumina sequencing was performed by Macrogen (Seoul, Korea) and PacBio sequencing by The University of Queensland Diamantina Institute (Brisbane, Australia). The SMRTbell template libraries were prepared following the standard protocol for long-insert libraries according to the manufacturer's instructions (PacBio) with an insert size of 20 kbp. 20 SMRTcells per taxon were sequenced resulting in approximately 40-fold genome coverage for each sample.

5.3.3 Data processing and genome assembly

Raw reads from both platforms were assessed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc), a tool for evaluating the quality of sequencing reads in FASTQ files. Illumina reads were additionally used to estimate the genome size of the taxa. A preqc module (Simpson 2014) from the SGA de novo genome assembler package (Simpson & Durbin 2012) was used for this estimation. This utility also enabled an estimation of heterozygosity and repeat content in the genome.

Taxon A and Taxon B genomes were assembled *de novo* using two strategies. The first one utilised Illumina and PacBio sequencing reads together (hybrid assembly) and the second one – PacBio data only (PacBio-only assembly). The software used to accomplish the hybrid assembly was DBG2OLC package (Ye *et al.* 2014). This assembly included Illumina PE raw reads and raw PacBio reads. The first step within the analysis involved Illumina read error correction and subsequent assembly with SparseAssembler beta version (Ye *et al.* 2012). The next step was "overlap and layout" with the output contigs from the first step and PacBio reads. The last step called the consensus contigs. In PacBio-only assembly, first, the raw reads from the PacBio platform were corrected using the PBcR pipeline (Berlin *et al.* 2015) with the self-correction feature enabled. The assembly was done using The Celera Assembler (CA) version 8.3rc2 (Myers *et al.* 2000) leaving the parameters as default. The primary contigs were filtered in order to keep only the unique contigs.

5.3.4 Evaluation of genome assemblies

First, each of the assemblies was assessed using Assembly Stats (assemblathon tool) from the Assemblathon project (Earl *et al.* 2011) accessed through the iPlant Collaborative platform [iPlant; Goff *et al.* (2011)]. Second, core genes presence was assessed in the assemblies. This was done using both CEGMA [Core Eukaryotic Genes Mapping Approach; Parra *et al.* (2007)] and BUSCO [Benchmarking Universal Single-Copy Orthologs; Simao *et al.* (2015)]. CEGMA uses a set of 248 CEGs (Core Eukaryotic Genes) which are very highly conserved in eukaryotes and are present in low copy number (Parra *et al.* 2009). BUSCO also uses a set of universal single-copy orthologs and provides a set to evaluate plant genomes in particular with a set of 956 plant orthologs. Additionally, both evaluation tools were run for the rice reference sequence of *O. sativa japonica* and the values for completeness of this high quality rice genome were used for normalisation.

The assemblies were also aligned to rice reference genomes using QUAST [Quality Assessment Tool for Genome Assemblies; Gurevich *et al.* (2013)] with the default parameters and the minimum alignment length of 1,000 bp. The genome of *Oryza sativa* ssp. *japonica* var. Nipponbare [IRGSP_MSU.v7; Kawahara *et al.* (2013)] and *O. meridionalis* (GenBank assembly accession: GCA_000338895.2) were used as the reference sequences in this study.

5.3.5 Rice pseudomolecules

Assembled PacBio contigs were assigned to chromosome pseudomolecules using Genome Puzzle Master [GPM; Zhang *et al.* (2016)]. Rice reference genome sequences were used to guide the process: *Oryza sativa* ssp. *japonica* var. Nipponbare genome was used for Taxon A and *O. meridionalis* was used for Taxon B. The twelve pseudomolecules and unanchored contigs were then annotated for genes and other features.

5.3.6 Genome annotation

Protein coding genes were annotated using the MAKER-P v.2.3 annotation pipeline (Campbell *et al.* 2014). Within the pipeline the repeat elements were masked using the RepeatMasker (www.repeatmasker.org, v. 3.3.0). Expression evidence included Oryza expressed tags (ESTs) and full-length cDNA. These CDS and their corresponding protein sequences (used as the protein homology evidence) consisted of annotated genes models of *O. sativa* ssp. *japonica* var. Nipponbare RefSeq, *O. glaberrima* (Wang *et al.* 2014b) and *Brachypodium distachyon* (The International Brachypodium Initiative 2010). ESTs

comprised of *O. sativa* ssp. *japonica* var. Nipponbare (for Taxon A annotation) and *O. meridionalis* (for Taxon B annotation) transcripts generated by The International *Oryza* Map Alignment Project (Stein *et al.*, personal communication) and clustered at 95% similarity. The *ab initio* gene predictors run within MAKER-P were SNAP (Korf 2004) using O.sativa.hmm parameter and AUGUSTUS 3.1 (Stanke & Waack 2003) with rice as the gene prediction species model. Resulting gene models were filtered removing non-complete models, i.e. without valid start and or stop codons and with internal terminator codons, followed by removing transposable elements based on the specific rice- and Australian taxa-related libraries used at the genome masking step. The gene models predicted by MAKER-P were functionally analysed using InterProScan version 5.16.55 (Jones *et al.* 2014). The InterProScan results were further parsed for additional functional information (KEGG pathway and GO terms) using interproscanParser script available at iPlant.

The repeat annotation was obtained by merging the output of RepeatMasker and Blaster, a component of the REPET package (Flutre *et al.* 2011), using nucleotide libraries (PReDa and RepeatExplorer) from RiTE-db (Copetti *et al.* 2015) and an in-house curated collection of transposable element (TE) proteins. Additionally, for each of the two species, a custom repeat library was developed with RepeatExplorer and curated as described previously (Copetti *et al.* 2015) using short read data (Illumina sequencing reads). Infernal (Nawrocki & Eddy 2013) was adopted to identify non-coding RNAs (ncRNAs) using the Rfam library Rfam.cm.1_1. Hits above the e-value threshold of 1e-5 were filtered, as well as results with scores lower than the family-specific gathering threshold. When loci on both strands were predicted, only the hit with the highest score was kept. Transfer RNAs were also predicted using tRNAscan-SE v. 1.23 (Schattner *et al.* 2005) with default parameters.

5.3.7 Phylogenetic analysis

Phylogenetic analysis was undertaken using data from twelve fully sequenced diploid *Oryza* genomes including two taxa investigated in this study (Taxon A and Taxon B), and ten other species downloaded from GenBank (Table 5.1). Eleven of these species were A genome type rice relatives. Moreover, we used *O. punctata*, which belongs to BB genome group, and *O. brachyantha*, which has FF genome type. *Leersia perrieri* was used as the outgroup species.

Table 5.1 *Oryza* and the outgroup species used in the phylogenetic study.

***Oryza* species**	**Genome group**	**GenBank Assembly Accession**
O. sativa japonica	AA	GCA_000005425.2
O. sativa indica	AA	GCA_000004655.2
O. rufipogon	AA	GCA_000817225.1
O. nivara	AA	GCA_000576065.1
O. barthii	AA	GCA_000182155.3
O. glaberrima	AA	GCA_000147395.2
O. glumaepatula	AA	GCA_000576495.1
O. meridionalis	AA	GCA_000338895.2
Taxon A	AA	LONB00000000
Taxon B	AA	LONC00000000
O. punctata	BB	GCA_000573905.1
O. brachyantha	FF	GCA_000231095.2
L. perrieri	–	GCA_000325765.3

From these diploid genomes a set of putatively single copy orthologs was selected by blasting (BLASTn) the initial collection of 6,015 genes used in another I-OMAP study (Stein *et al.*, personal communication) against Australian *Oryza* genomes (Taxon A and Taxon B). A final subset of 4,643 genes sequences present in all *Oryza* assemblies was extracted from the genomes and used for further investigation. Nucleotide sequences of the 4,643 genes selected from each genome were separately aligned using CLUSTAL W multiple sequence alignment program (Thompson *et al.* 1994). Then, single gene alignments were concatenated to create a supermatrix of 6,272,851 base pairs, which was used in the following phylogenetic inference.

Phylogenetic tree reconstruction was conducted using maximum parsimony (MP) and Bayesian inference (BI) methods. MP was performed using PAUP* 4.0 software (Swofford 2003). The following tree search settings were enabled in MP reconstruction: heuristic search with tree bisection-reconnection branch swapping and 200 random addition sequence replications. The group support was assessed using 2,000 bootstrap pseudoreplications. Alignment gaps were treated as missing data. All characters were treated as unordered and weighted equally.

For the model-based approach (BI), jModelTest2 software (Darriba *et al.* 2012) was used to determine the model of nucleotide substitution that best fits the data based on the Akaike Information Criterion. The Bayesian analyses used the General Time Reversible

Model with gamma-shaped among-site rate variation with an estimated proportion of invariable sites (GTR+I+G; p-inv = 0.3730, four gamma categories and gamma shape = 0.8890). The BI analysis was performed using MrBayes version 3.2 (Ronquist *et al.* 2012). The branch length prior was set to exponential with parameter 10.0. Two independent and simultaneous analyses starting from distinct random trees were performed. Three heated (heating coefficient = 0.2) and one cold Monte Carlo Markov Chains (MCMC) were run for 1×10^6 generations, with a tree sampled every 200 generations. The first 10% of trees were discarded as burn-in and a 50% majority rule consensus tree was constructed and rooted using the outgroup method.

5.3.8 Divergence time estimates

Divergence times were estimated using the Bayesian evolutionary method implemented in the software package BEAST 2 version 2.3.1 (Bouckaert *et al.* 2014). A secondary clock calibration was used based on the estimated divergence time for the *Oryza* crown group (*O. brachyantha*-*O. punctata*) of 15 ± 0.5 mya (Stein *et al.*, personal communication). The 4,643 genes used were divided according to the chromosome they were found on and aligned using CLUSTAL W. The number of genes and the alignment length is shown in Table 5.2. The best fit evolutionary model, determined by jModelTest2, was the General Time Reversible Model (GTR+I+G) for each of the alignments. Evolutionary rates were modelled under a strict molecular clock and speciation was modelled employing the Yule Model. Posterior probabilities were estimated using MCMC algorithm with chain length 5,000,000 and a tree sampled every 1,000th generation. The first 10% of sampled trees was discarded as burn-in. The output from BEAST 2 was analysed in Tracer version 1.63 (www.beast.bio.ed.ac.uk/Tracer). The best supported tree with the highest product of the posterior probability of all its nodes (maximum clade credibility tree) and the mean heights of each node was summarised using TreeAnnotator distributed in the BEAST 2 package. The final tree estimates were visualized in FigTree version 1.4.2 (www.tree.bio.ed.ac.uk/software/figtree).

Table 5.2 Summary of data used in divergence time estimation.

Chromosome	# genes	Alignment length
1	659	885,116
2	748	996,507
3	593	805,108
4	535	706,297
5	511	658,924
6	398	551,896
7	372	533,899
8	272	350,596
9	256	345,315
10	65	93,087
11	98	134,216
12	136	211,890
Total	4,643	6,272,851

5.3.9 Genetic introgression

To test for genetic introgression between Australian and other *Oryza* species we used the software package HYBRIDCHECK (Ward & van Oosterhout 2016). We performed two tests and in each run four aligned sequences were analysed. In the first one we tested for introgression between Taxon A and either *O. rufipogon* or *O. barthii* and in the second we tested for introgression between Taxon A and either *O. rufipogon* or *O. nivara*. Block jackknife was used to calculate the statistic with the block size of 20,000 as well as Z-score to measure the statistical significance.

5.4 Results

5.4.1 Genome sequencing

The statistics of sequencing reads obtained in this study are shown in Table 5.3. Total data produced by the Illumina platform was 47.1 Gb and 41.4 Gb for Taxon A and Taxon B, respectively. The data generated on the PacBio instrument were long reads with an average length of 7,693 bp and 8,140 bp for Taxon A and Taxon B, respectively, with 14.8 Gb and 15.0 Gb of overall data for those taxa. The minimum and maximum read lengths for Taxon A were 50 bp and 49,742 bp, respectively, and for Taxon B: 50 bp and 50,242 bp.

Table 5.3 Illumina and PacBio sequencing reads statistics for Taxon A and Taxon B.

	# reads	# bases	Avg. length (bp)	Avg. GC (%)	Avg. ambiguous base (%)	Avg. quality (Phred)	Quality distribution (Phred)
Illumina							
Taxon A							
PE	200,095,430	20,209,638,430	101	43	0.11	35	10-40
3 Kb	155,539,872	15,709,527,072	101	44	0.04	35	2-40
5 Kb	110,916,270	11,202,543,270	101	44	0.04	35	2-40
Total	466,551,572	47,121,708,772	101	43	0.06	35	2-40
Taxon B							
PE	172,599,982	17,432,598,182	101	42	0.11	35	10-40
3 Kb	119,259,720	12,045,231,720	101	44	0.04	35	2-40
5 Kb	118,252,654	11,943,518,054	101	43	0.04	35	2-40
Total	410,112,356	41,421,347,956	101	43	0.06	35	2-40
PacBio							
Taxon A							
	1,928,732	14,781,976,594	7,664	43	NA	10	0-15
Taxon B							
	1,867,655	15,002,306,864	8,140	43	NA	10	0-15

PE – paired end reads, 3 Kb – mate pair reads with 3 Kb insert size, 5 Kb – mate pair reads with 5Kb insert size, NA – no ambiguous bases detected.

Table 5.4 Illumina and PacBio sequencing genome coverage for Taxon A and Taxon B calculated using the estimated genome sizes (390 Mb and 370 Mb for Taxon A and Taxon B, respectively).

Taxa	**Genome coverage (x)**				
	Illumina				**PacBio**
	PE	3 Kb	5 Kb	Total	Total
Taxon A	51.8	40.3	28.7	120.8	37.9
Taxon B	47.1	32.6	32.3	112.0	40.6

PE – paired end reads, 3 Kb – mate pair reads with 3 Kb insert size, 5 Kb – mate pair reads with 5Kb insert size.

The genome sizes were estimated *in silico* to be about 390 Mb and 370 Mb for Taxon A and Taxon B, in turn. These estimates were similar to other A genome rice species which fall between 341 and 413 Mb in size (Zhang *et al.* 2014). Considering the estimated genome sizes, we also assessed the genome coverage of each of the data sets

(Table 5.4). Furthermore, we used these estimations in evaluating completeness of genome assemblies.

5.4.2 Genome assemblies and evaluation

A total of 384.8 Mb (PacBio-only assembly) and 382.7 Mb (hybrid assembly) of the Taxon A, and 354.9 Mb (PacBio-only) and 446.4 Mb (hybrid) of the Taxon B genome sequences were assembled (Table 5.5). PacBio assemblies slightly outperformed hybrid assemblies in terms of standard assembly metrics, i.e. lower number of scaffolds, longest contig size, higher N50 and mean scaffold size. Both Taxon A and Taxon B assemblies exhibited high total lengths, as percentage of known genome sizes, with an unexpected high length of the Taxon B hybrid assembly that accounted for around 120% of estimated genome size (370 Mb). The high percentage of estimated genome size for Taxon B hybrid assembly may be due to the heterozygous and repetitive nature of this taxon's genome or to a non-precise genome size estimation. Predicted heterozygous sites, measured as the rate of variant branches caused by allelic differences in a de Bruijn graph, (1 in 400) and repeat content (1 in 300) rates were slightly higher for Taxon B compared to Taxon A (1 in 800 and 1 in 400, respectively; data not shown). Those traits might impact the assembly quality and completeness especially using short read data (Illumina) resulting in a more fragmented assembly with a higher number of repeated contigs.

Table 5.5 Taxon A and Taxon B hybrid and PacBio assembly statistics. The metrics were calculated for scaffolds and contigs for hybrid assembly and for scaffolds only for PacBio assembly.

	Taxon A		**Taxon B**	
Assembly	**Hybrid**	**PacBio-only**	**Hybrid**	**PacBio-only**
Assembler	Sparse Assembler + DBG2OLC	Celera Assembler	Sparse Assembler + DBG2OLC	Celera Assembler
Scaffolds				
Number of scaffolds	3,359	2,585	4,718	3,252
Total size of scaffolds	382,655,312	384,759,810	446,369,637	354,906,376
Total scaffold length as percentage of known genome size	98.1	98.7	120.6	95.9
Longest scaffold	1,305,248	1,692,155	2,079,733	3,232,522
Shortest scaffold	2,297	9,523	2,425	12,563
Number of scaffolds > 500 nt	3,359	2,585	4,718	3,252
Percentage of scaffolds > 500 nt	100	100	100	100

	Taxon A		Taxon B	
Assembly	**Hybrid**	**PacBio-only**	**Hybrid**	**PacBio-only**
Number of scaffolds > 1K nt	3,359	2,585	4,718	3,252
Percentage of scaffolds > 1K nt	100	100	100	100
Number of scaffolds > 10K nt	3,228	2,583	4,604	3,252
Percentage of scaffolds > 10K nt	96.1	99.9	97.6	100
Number of scaffolds > 100K nt	1,221	1,274	1,343	834
Percentage of scaffolds > 100K nt	36.4	49.3	28.5	25.6
Number of scaffolds > 1M nt	4	8	15	21
Percentage of scaffolds > 1M nt	0.1	0.3	0.3	0.6
Mean scaffold size	113,919	148,843	94,610	109,135
Median scaffold size	61,996	97,803	54,787	61,207
N50 scaffold length	217,336	219,409	163,003	159,640
L50 scaffold count	511	505	718	461
NG50 scaffold length	211,599	215,884	200,474	147,615
LG50 scaffold count	528	517	507	510
N50 scaffold - NG50 scaffold length difference	5,737	3,525	37,471	12,025
scaffold %A	28.29	28.03	28.49	28.42
scaffold %C	21.7	21.96	21.5	21.58
scaffold %G	21.71	21.96	21.52	21.58
scaffold %T	28.29	28.05	28.48	28.42
scaffold %N	0	0	0	0
Contigs				
Percentage of assembly in scaffolded contigs	4.1	-	5.6	-
Percentage of assembly in unscaffolded contigs	95.9	-	94.4	-
Average number of contigs per scaffold	1	-	1	-
Average length of break (>25 Ns) between contigs in scaffold	3.29	-	3.95	-
Number of contigs	3,425	-	4,808	-
Number of contigs in scaffolds	130	-	177	-
Number of contigs not in scaffolds	3,292	-	4,631	-
Total size of contigs	382,644,322	-	446,351,110	-
Longest contig	1,158,569	-	1,449,836	-
Shortest contig	1,139	-	790	-

	Taxon A		Taxon B	
Assembly	**Hybrid**	**PacBio-only**	**Hybrid**	**PacBio-only**
Number of contigs > 500 nt	3,425	-	4,808	-
Percentage of contigs > 500 nt	100	-	100	-
Number of contigs > 1K nt	3,425	-	4,807	-
Percentage of contigs > 1K nt	100	-	100	-
Number of contigs > 10K nt	3,280	-	4,677	-
Percentage of contigs > 10K nt	95.8	-	97.3	-
Number of contigs > 100K nt	1,227	-	1,347	-
Percentage of contigs > 100K nt	35.8	-	28	-
Number of contigs > 1M nt	3	-	13	-
Percentage of contigs > 1M nt	0.1	-	0.3	-
Mean contig size	111,721	-	92,835	-
Median contig size	61,459	-	54,495	-
N50 contig length	211,599	-	159,759	-
L50 contig count	522	-	743	-
NG50 contig length	209,177	-	196,877	-
LG50 contig count	540	-	528	-
N50 contig - NG50 contig length difference	2,422	-	37,118	-
contig %A	28.29	-	28.49	-
contig %C	21.71	-	21.50	-
contig %G	21.71	-	21.52	-
contig %T	28.3	-	28.48	-
contig %N	0	-	0	-

The number of scaffolds obtained as a result of the assemblies was between 2,585 and 3,252 for Taxon A and Taxon B, respectively for PacBio assemblies and between 3,359 and 4,718 for Taxon A and Taxon B, respectively for hybrid assemblies. The number of scaffolds was slightly higher for hybrid assemblies for both taxa. Higher number of scaffolds in Taxon B assemblies might also be a result of the more heterozygous and repetitive nature of this taxon in comparison with Taxon A.

Both of the core genes presence evaluation methods, CEGMA and BUSCO, indicated PacBio assemblies to be more complete than hybrid assemblies for both taxa (Table 5.6 and Table 5.7). Normalized values ranged from 94.8% to 99.1% of completeness using CEGMA and from 89% to 98% using BUSCO. Neither wild rice

assembly was found to have a higher number of mapped genes than the Nipponbare reference.

Table 5.6 Completeness of Taxon A and Taxon B assemblies evaluated by means of presence of CEGMA core genes.

Taxa	Assembly	Mapped CEGs	Not mapped CEGs	% complete[a]	Normalized (%)
Taxon A	Hybrid	218	30	87.9	95.2
	PacBio-only	227	21	91.5	99.1
Taxon B	Hybrid	217	29	87.5	94.8
	PacBio-only	225	23	90.7	98.3
Oryza sativa spp. *japonica*	IRGSP_MSU.v7	229	19	92.3	100

[a] Percentage of CEGs present.

Table 5.7 Completeness of Taxon A and Taxon B assemblies evaluated by means of presence of BUSCO orthologous genes.

Taxa	Assembly	Single	Duplicated	Fragmented	Missing	% complete[a]	Normalized
Taxon A	Hybrid	832	246	46	78	87	89
	PacBio-only	918	212	22	16	96	98
Taxon B	Hybrid	830	255	56	70	86	89
	PacBio-only	836	170	55	65	87	89
Oryza sativa spp. *japonica*	IRGSP_MSU.v7	936	216	10	10	97	100

[a] Percentage of Complete Single-Copy BUSCOs present, Single – Complete Single-Copy BUSCOs, Duplicated – Complete Duplicated BUSCOs, Fragmented – Fragmented BUSCOs, Missing – Missing BUSCOs.

A high fraction of the Nipponbare reference genome was aligned to Taxon A assemblies, 70.7% and 71.5% to hybrid and PacBio, respectively (Table 5.8). However, a significantly lower percentage was aligned to Taxon B assemblies, 42.4% and 37.3% to hybrid and PacBio, respectively. These values were not high enough to use the Nipponbare genome as a reference for Taxon B in orienting and ordering contigs into chromosome pseudomolecules using GPM (Zhang *et al.* 2016). Based on the previous finding that Taxon B shares numerous molecular markers with *O. meridionalis* (Sotowa *et*

al. 2013) and that they descended from a common ancestor (this study), this genome was evaluated as well. QUAST results showed higher *O. meridionalis* genome fraction aligning to Taxon B assemblies than Nipponbare genome, 62.4% and 56.3% to hybrid and PacBio, respectively. These values were satisfactory allowing the use of *O. meridionalis* instead of Nipponbare sequences as a guide in GPM for Taxon B assembly.

Table 5.8 Unaligned and partially unaligned contig metrics of Australian wild rice taxa.

	Taxon A				**Taxon B**			
Reference genome	*O. sativa*		*O. meridionalis*		*O. sativa*		*O. meridionalis*	
Assembly	Hybrid	PacBio-only	Hybrid	PacBio-only	Hybrid	PacBio-only	Hybrid	PacBio-only
# unaligned contigs[a]	67	14	803	265	301	33	676	240
Fully unaligned length (Mb)	1.1	0.5	22.7	14.8	6.5	1.5	20.6	12.0
# partially unaligned contigs	670	279	2,354	2,120	2,577	1,688	2,075	1,311
Partially unaligned length (Mb)	25.8	15.3	263.9	266.6	202.1	162.0	116.6	80.2
Unaligned length[b]	26.9	15.9	286.6	281.3	208.6	163.5	137.3	92.3
Genome fraction (%)	70.7	71.5	23.6	25.4	42.4	37.3	62.4	56.3
# genes	38,587	33,346	4,981	5,402	26,532	22,588	18,006	16,357
# partial genes	8,106	10,063	8,950	9,421	9,399	8,794	8,822	7,805

[a] Number of contigs without alignments to the reference sequences, fully or partially.

[b] Total length of the unaligned contigs (sum of fully and partially unaligned regions in the assembly).

5.4.3 Rice psuedomolecules

After the preliminary evaluation of the assemblies (basic assembly statistics, core gene presence and alignment to the reference genomes), the PacBio-only scaffolds were chosen for further analysis and investigation. Ordering and orientation of the contigs with GPM resulted in 12 pseudochromosomes for both taxa and 386 unordered contigs for Taxon A and 1,080 for Taxon B (Table 5.9). 94.9% and 83.1% of the assemblies' length were anchored and oriented to chromosomes for Taxon A and Taxon B, respectively. Alignment of wild rice pseudomolecules to their reference genomes revealed better coverage and less ambiguity between Taxon A and *O. sativa japonica* than between Taxon B and *O. meridionalis* genome (Figure 5.3).

Table 5.9 Length and GC content of Taxon A and Taxon B psuedosomocules and remaining contigs.

Pseudomolecule	**Taxon A**		**Taxon B**	
	%GC	Length	%GC	Length
Chr01	44.0%	41,912,295	43%	33,618,492
Chr02	43.8%	37,496,392	43%	27,986,264
Chr03	44.2%	39,112,528	43%	27,026,720
Chr04	44.1%	31,213,496	44%	31,144,760
Chr05	43.9%	29,753,236	43%	23,976,054
Chr06	43.8%	31,811,270	43%	22,865,571
Chr07	43.9%	28,478,632	43%	26,941,941
Chr08	43.6%	27,136,197	43%	22,577,491
Chr09	44.1%	23,188,492	43%	19,679,370
Chr10	43.8%	22,039,502	43%	18,408,352
Chr11	43.2%	27,446,056	43%	22,781,387
Chr12	43.2%	25,545,820	43%	17,848,930
Unordered	45.7%	19,846,994	43%	59,973,669
Total	43.9%	384,980,910	43%	354,829,001

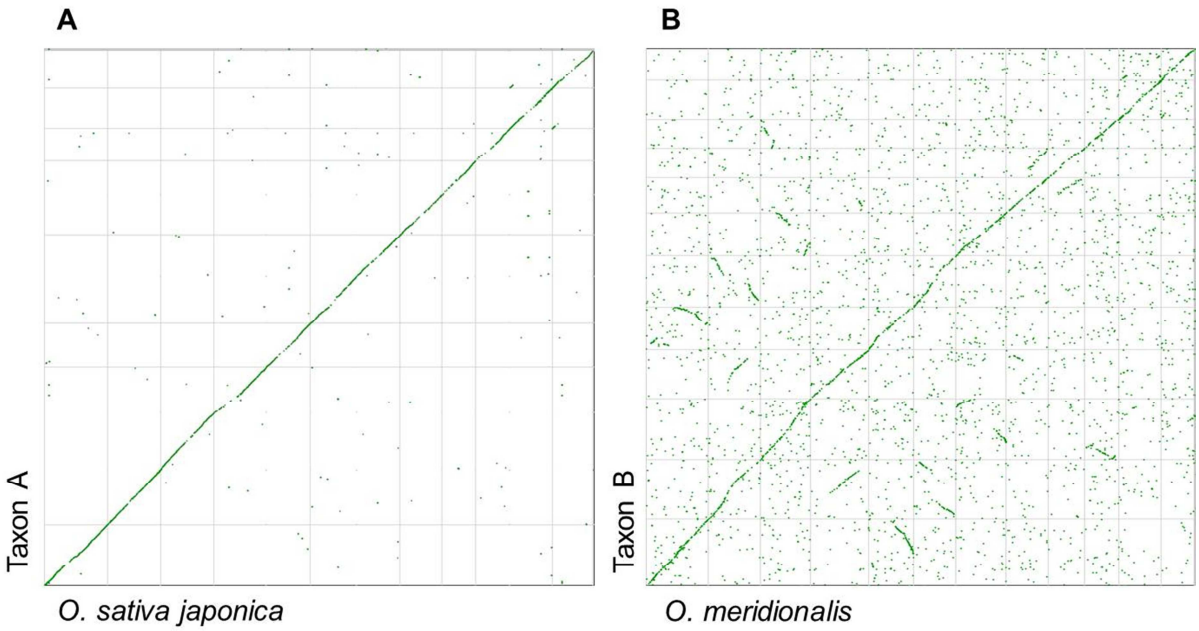


Figure 5.3 Dot plot of wild rice psuedomolecules with reference genomes chromosomes. (A) Taxon A and its reference, *O. sativa* ssp. *japonica* var. Nipponbare. (B) Taxon B and its reference, *O. meridionalis*. Each square corresponds to one of the 12 rice chromosomes.

5.4.4 Genome annotation

Repetitive elements, RNAs and protein coding genes were annotated in Taxon A and Taxon B draft genomes. The sequences subjected to the annotation were the 12 psuedomolecules and remaining unordered contigs for each of the taxa.

Total repeats found in Australian wild rices made up 36.5% and 46.4% of the Taxon A and Taxon B genomes, respectively (Table 5.10 and Table 5.11). The most abundant class of transposable elements found were retrotrasposons from the Gypsy superfamily. These represented 57.8% and 39.7% of all repeats described in Taxon A and Taxon B, respectively, followed by the Copia superfamily in Taxon A (8.3%) and Mutator in Taxon B (9.3%). The classes and fractions of other repetitive elements were similar in both taxa; however the numbers and lengths were significantly higher in the Taxon B genome.

Non-coding RNAs annotated in the wild rice genomes included tRNA, miRNA, snoRNA, sRNA, rRNA and other (Table 5.12). Overall, RNAs consisted of approximately 0.25% of both genomes which corresponded to length of 960,496 bp in Taxon A and 883,779 bp in Taxon B. 675 and 558 tRNAs models were predicted by tRNAscan in Taxon A and Taxon B respectively, whereas 629 and 581 tRNAs models were predicted by Infernal, respectively. Additional models predicted by Infernal but not by tRNAscan were added to the final annotations resulting in 677 and 615 tRNAs for Taxon A and Taxon B, respectively, of the combined length of 50,687 and 46,301 bp.

The number of gene models annotated in wild rice genomes is listed in Table 5.13. Slightly more genes were found in the Taxon A genome which was probably associated with the longer total assembly of this taxon (384.8 Mb as opposed to 354.9 Mb of Taxon B). In comparison with other wild rice species (Stein *et al.*, personal communication) these taxa showed a considerably lower number of genes. In the previous study the lowest number of annotated loci was found in *O. brachyantha* (24,208), which also carries the smallest genome described so far in the genus *Oryza* (261 Mb). Similar numbers of InterPro protein domains, KEGG pathways or GO terms were found in the two genomes (Table 5.13). Overall, just over 61% of annotated models had matches in the InterPro database, about 9% in KEGG and around 40% in GO.

Table 5.10 Repetitive elements annotated in Taxon A genome.

		Genome coverage (bp)	Repeat #	Repeat fraction (%)	Genome fraction (%)
Class I (retrotransposons)					
LTR	Copia	11,716,567	7,037	8.33	3.05
	Gypsy	81,228,038	27,133	57.77	21.11
	Retrovirus	609,338	182	0.43	0.16
	Other LTR	898,540	5,117	0.64	0.23
LINE	L1	4,284,820	5,478	3.05	1.11
	Other LINE	4,769	42	0.00	0.00
SINE	SINE	407,798	1,082	0.29	0.11
Other Class I		49,006	557	0.03	0.01
Class II (DNAt) subclass 1					
TIR	Tc1-Mariner	1,381,604	5,255	0.98	0.36
	hAT	2,183,755	3,465	1.55	0.57
	Mutator	8,417,944	9,364	5.99	2.19
	PIF-Harbinger	2,950,940	9,159	2.10	0.77
	CACTA	11,204,924	6,531	7.97	2.91
	Other DNAt	504,947	2,144	0.36	0.13
	MITE	1,715,292	5,719	1.22	0.45
Class II (DNAt) subclass 2					
Helitron		4,492,507	5,207	3.20	1.17
Other Class II		45,350	332	0.03	0.01
Total TEs		132,096,139	93,804	93.95	34.33
Ribosomal DNA		135,588	133	0.10	0.04
Structural Repeats		7,828,864	139,062	5.57	2.03
Unclassified		534,597	3,697	0.38	0.14
Total Repeats		140,595,188	236,696		36.54

Table 5.11 Repetitive elements annotated in Taxon B genome.

		Genome coverage (bp)	Repeat #	Repeat fraction (%)	Genome fraction (%)
Class I (retrotransposons)					
LTR	Copia	14,396,632	13,267	8.75	4.06
	Gypsy	65,230,139	37,092	39.65	18.38
	Retrovirus	367,124	263	0.22	0.10
	Other LTR	827,384	5,581	0.50	0.23
LINE	L1	4,842,831	9,820	2.94	1.36
	Other LINE	57,132	483	0.03	0.02
SINE	SINE	1,892,063	10,196	1.15	0.53
Other Class I		134,526	1,910	0.08	0.04
Class II (DNAt) subclass 1					
TIR	Tc1-Mariner	9,027,155	57,805	5.49	2.54
	hAT	4,338,182	17,762	2.64	1.22
	Mutator	15,222,784	50,417	9.25	4.29
	PIF-Harbinger	11,760,001	61,983	7.15	3.31
	CACTA	13,770,423	18,078	8.37	3.88
	Other DNAt	4,061,534	42,279	2.47	1.14
	MITE	5,690,040	45,488	3.46	1.60
Class II (DNAt) subclass 2					
Helitron		6,863,768	15,316	4.17	1.93
Other Class II		314,866	4,526	0.19	0.09
Total TEs		158,796,584	392,266	96.52	44.74
Ribosomal DNA		231,301	1,185	0.14	0.07
Structural Repeats		3,451,409	68,645	2.10	0.97
Unclassified		2,035,778	26,940	1.24	0.57
Total Repeats		164,515,072	489,036		46.35

Table 5.12 Non-coding RNA annotation in Taxon A and Taxon B genomes.

	Taxon A		Taxon B	
	number	length (bp)	number	length (bp)
tRNAs	677	50,687	615	46,301
miRNA	5,398	801,838	4,892	734,557
snoRNA	499	53,803	453	47,791
sRNA	91	13,717	87	13,341
rRNA	115	29,046	129	34,496
Other[a]	76	11,405	63	7,293
Total	6,179	909,809	5,624	837,478

tRNA – transfer RNA, miRNA – micro RNA, snoRNA – smll nucleolar RNA, sRNA – spliceosomal RNA, rRNA – ribosomal RNA.

[a] Other includes signal recognition particle RNA, nodulin 40, TPP riboswitch, iron stress repressed RNA, RNase MR.

Table 5.13 Protein coding genes annotation in Taxon A and Taxon B genomes.

	Taxon A	Taxon B
# genes	22,035	21,169
# transcripts	27,473	26,735
# exons	99,618	97,975
# InterPro domain	13,505	13,072
# KEGG pathway	1,926	1,904
# GO annotation	8,763	8,466

5.4.5 Phylogenetic analysis

Sequences of 4,643 genes were extracted from nuclear genome sequences generated by whole genome sequencing (Table 5.1; Stein *et al.*, personal communication). The alignment of these 4,643 gene sequences comprised of 6,272,851 bp. The sequence similarity between the *Oryza* species was very high (Table 5.14) ranging from 86.6% (between *O. meridionalis and O. brachyantha*) to up to 98.4% (between *O. rufipogon* and *O. sativa* ssp. *japonica*). Out of the total number of bases that were subjected to the MP analysis, 5,229,706 were constant, 741,498 were variable and parsimony uninformative and 301,647 were parsimony informative. Both phylogenetic inference methods used in this study, MP and BI, recovered the same optimal tree topology (Figure 5.2B) with the following values for the MP tree: length = 1,264,556 steps, consistency index CI = 0.90, retention index RI = 0.74, CI excluding uninformative characters = 0.72. The nodes on this

topology were all strongly supported with MP bootstrap values of 100% and the posterior probabilities of all nodes in the BI equal to 1.

Our results showed that *O. meridionalis* and Taxon B in Australia are sister to all other A genome species including the Australian Taxon A (Figure 5.2B). Taxon A is in turn a sister to the clade that includes the Asian and African domesticated species. African domesticated rice, *O. glaberrima*, and its wild progenitor, *O. barthii*, together with *O. glumaepatula* from South America are a clade distinct from the Asian species (Wambugu *et al.* 2013). *Indica* and *japonica* rice are represented by two well resolved clades. *Japonica* and *O. rufipogon* show a close relationship which is consistent with the long accepted view that *O. rufipogon* is the progenitor of *japonica* (Wei *et al.* 2012) while *indica* rice was found in a clade with the Asian annual *O. nivara*. Recent SNP analysis of genomic regions under selection suggests the independent domestication of *indica* rice from wild rice in an area from southern Indochina to the Brahmaputra valley (Civáň *et al.* 2015).

The nuclear phylogenies presented here showed a different relationship to those deduced from the chloroplast genomes suggesting different evolutionary histories for the maternally-inherited plastid genetic material and the nuclear genome (Figure 5.2). This difference in evolutionary path for these genomes is commonly observed in recently diverged plant taxa (Tsitrone *et al.* 2003).

Table 5.14 Sequence similarities between rice taxa in the supermatrix used for phylogenetic inference.

	***Oryza* species**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	*O. sativa indica*		97.7	97.6	97.5	96.7	97.2	97.2	96	94.8	94.7	93	87.1
2	*O. nivara*			98.1	98.2	97.4	97.9	98	96.6	95.3	95.2	93.5	87.6
3	*O. sativa japonica*				98.4	97.3	98	97.9	96.8	95.4	95.4	93.6	87.7
4	*O. rufipogon*					97.3	97.8	97.8	96.5	95.3	95.2	93.5	87.6
5	*O. barthii*						97.9	97.3	95.8	94.8	94.6	93	87.3
6	*O. glaberrima*							97.8	96.4	95.3	95.2	93.5	87.7
7	*O. glumaepatula*								96.4	95.4	95.2	93.5	87.7
8	Taxon A									94.4	95.2	92.7	87.1
9	*O. meridionalis*										95.1	92.5	86.6
10	Taxon B											92.4	86.7
11	*O. punctata*												87.6
12	*O. brachyantha*												

The panel represents percentage of bases that are identical in the supermatrix alignment between corresponding species. The heatmap shows individual values in a matrix as colours: red cells indicate high similarities; orange, yellow and green cells show gradually lower similarities.

The relationship of the Australian A genome populations was greatly clarified by this study. *O. meridionalis* and the morphologically similar perennial populations (Taxon B) are sister to all other A genome species. The Australian wild populations with morphology similar to *O. rufipogon* (Taxon A) were found to be sister to the clades including all other A genome species. The Australian (Taxon A) population has a large anther like that of *O. rufipogon* in Asia but is morphologically distinct from the other Australian A genome species and domesticated rice with small anthers. Taxon A and Taxon B can be most readily distinguished in the field by the open panicles of Taxon A and closed panicles of Taxon B (Figure 5.1). Taxon B generally has longer awns but the ranges of awn length for the two taxa overlap. The presence of these diverse A genome taxa makes northern Australia a key centre of diversity for rice and indicates the need for more collections from this poorly explored area and the need to ensure in situ conservation of these resources.

Analysis of the timing of the evolutionary events in this study (Figure 5.4) agrees well with that reported in analysis of the *Oryza* genomes (Stein *et al.*, personal communication) despite the use of a different method of analysis. The average rate of evolution was estimated to be 3.53E-3 ± 1.85E-6 (Table 5.15) and the root age (divergence between *O. brachyantha* and *O. punctata*) to be 14.98 ± 0.97 mya (Figure 5.4). Given this root age the A genome group diverged in the last 3 million years and the divergence of the *japonica* and *indica* clades dated at about 990,000 years ago (Figure 5.4). The chloroplast genomes appear to have diverged more recently (Wambugu *et al.* 2015) possibly due to some degree to the sharing of maternal genomes across this group.

Table 5.15 The average rate of evolution estimates for *Oryza* species by chromosome.

Chromosome 1		**Chromosome 2**		**Chromosome 3**		**Chromosome 4**	
mean	**stderr**	**mean**	**stderr**	**mean**	**stderr**	**mean**	**stderr**
3.29E-03	1.82E-06	3.51E-03	1.76E-06	3.02E-03	1.53E-06	3.66E-03	1.90E-06

Chromosome 5		**Chromosome 6**		**Chromosome 7**		**Chromosome 8**	
mean	**stderr**	**mean**	**stderr**	**mean**	**stderr**	**mean**	**stderr**
3.78E-03	1.95E-06	3.38E-03	1.67E-06	3.83E-03	1.90E-06	3.79E-03	1.93E-06

Chromosome 9		**Chromosome 10**		**Chromosome 11**		**Chromosome 12**	
mean	**stderr**	**mean**	**stderr**	**mean**	**stderr**	**mean**	**stderr**
3.78E-03	2.04E-06	3.29E-03	1.89E-06	3.38E-03	1.78E-06	3.69E-03	2.00E-06

Analysis of the phylogenetic relationship for each chromosome separately revealed some discordant results for chromosomes 5, 7, 10 and 11 (Figure 5.5, Table 5.16) that is likely due to introgression rather than incomplete lineage sorting. To test for potential recombination events between Australian Taxon A and other *Oryza* species we performed a four-taxon test, also known as the D-statistic (Durand *et al.* 2011; Green *et al.* 2010) separately for each chromosome. This test screens the aligned data for two biallelic mutation patterns: ABBA and BABA. The first species set we used was ((*O. rufipogon*, *O. barthii*), Taxon A, *O. punctata*) where *O. punctata* was used as the outgroup. In this analysis a negative D-statistic value would suggest introgression between Taxon A and the Asian species, whereas a positive value would mean an introgression between Taxon A and the African species (*O. barthii*). In this set all but two chromosomes showed negative D-statistic values indicating that introgression occurred between Australian Taxon A and Asian *O. rufipogon* (Figure 5.6 and Table 5.17). Eight of these statistics were significant (chromosomes 1, 3, 5, 6, 8, 9, 11 and 12). Among chromosomes with significant results was chromosome 5 for which Taxon A was found to be closer on the phylogenetic tree to the Asian clade (Figure 5.5) than it was on the consensus tree (Figure 5.2B). The D-statistic calculated for chromosome 7 was positive and statistically significant thereby providing evidence for introgression between Taxon A and *O. barthii*.

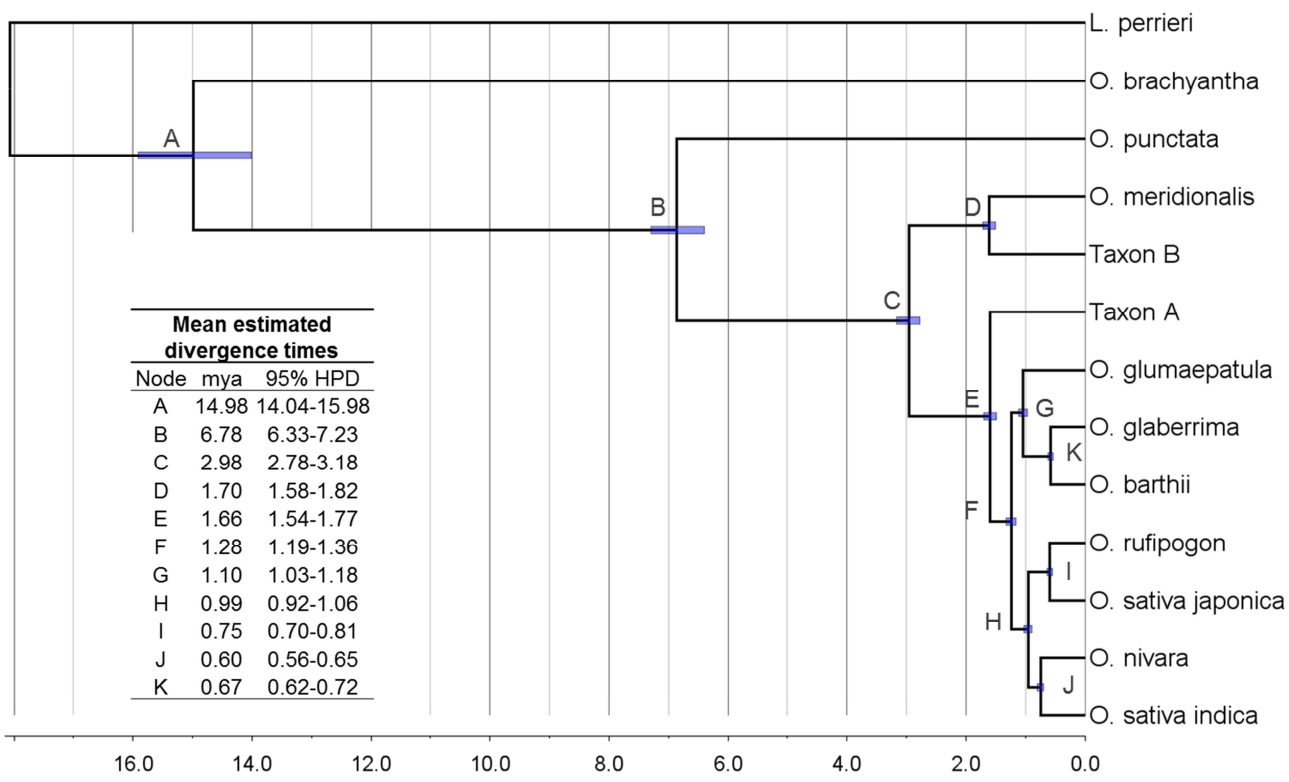


Figure 5.4 Molecular clock analyses for A genome rice evolution. The most frequent tree topology retrieved in analyses of alignments of separate chromosomes, inferred for 8 out of 12 chromosomes. Scale axis represents age in million years (mya). Node bars display 95% Highest Posterior Density (HPD) interval.

Table 5.16 Divergence times estimates for *Oryza* species by chromosome.

	Chromosome 1		Chromosome 2		Chromosome 3		Chromosome 4	
Node	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**
A	14.98	14.04-15.96	14.97	14.07-16.01	14.99	14.04-16.02	14.98	14.07-16.05
B	6.72	6.26-7.14	6.86	6.41-7.31	6.80	6.35-7.25	6.76	6.32-7.23
C	2.66	2.48-2.83	3.17	2.92-3.38	2.79	2.61-2.99	3.16	2.95-3.37
D	2.02	1.89-2.15	2.11	1.97-2.25	2.07	1.93-2.21	1.50	1.39-1.60
E	1.57	1.47-1.68	1.71	1.59-1.82	1.40	1.31-1.50	1.73	1.62-1.85
F	1.14	1.06-1.23	1.33	1.24-1.42	1.16	1.09-1.24	1.24	1.15-1.32
G	1.02	0.95-1.09	1.14	1.07-1.22	0.99	0.92-1.06	1.06	0.99-1.13
H	0.91	0.84-0.98	1.09	1.02-1.17	0.97	0.90-1.04	0.96	0.89-1.03
I	0.68	0.63-0.73	0.82	0.76-0.88	0.75	0.69-0.80	0.78	0.73-0.84
J	0.59	0.55-0.64	0.62	0.58-0.66	0.60	0.56-0.65	0.60	0.55-0.64
K	0.51	0.48-0.55	0.72	0.67-0.77	0.49	0.46-0.53	0.63	0.59-0.68

	Chromosome 6		Chromosome 8		Chromosome 9		Chromosome 12	
Node	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**
A	14.98	14.01-15.91	14.99	14.05-15.95	14.99	14.03-15.96	14.98	14.04-16.01
B	6.86	6.39-7.29	6.77	6.34-7.22	6.88	6.43-7.35	6.59	6.13-7.03
C	2.96	2.77-3.16	3.09	2.88-3.30	3.09	2.88-3.30	2.93	2.73-3.14
D	1.61	1.50-1.72	1.08	1.00-1.16	1.48	1.37-1.59	1.72	1.59-1.85
E	1.60	1.48-1.70	1.74	1.62-1.86	1.60	1.49-1.72	1.90	1.77-2.04
F	1.24	1.16-1.32	1.30	1.20-1.38	1.25	1.16-1.33	1.55	1.44-1.66
G	1.04	0.97-1.11	1.20	1.11-1.28	1.18	1.09-1.26	1.19	1.10-1.28
H	0.96	0.89-1.02	0.94	0.87-1.01	1.07	0.99-1.15	1.00	0.93-1.08
I	0.75	0.69-0.80	0.77	0.71-0.82	0.72	0.67-0.78	0.74	0.68-0.80
J	0.59	0.54-0.63	0.54	0.50-0.59	0.80	0.74-0.86	0.46	0.42-0.51
K	0.58	0.54-0.62	0.76	0.71-0.82	0.84	0.78-0.90	0.81	0.74-0.87

Chromosomes with different tree topology than the majority of chromosomes

	Chromosome 5		Chromosome 7		Chromosome 10		Chromosome 11	
Node	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**	**mya**	**95% HPD**
A	14.98	14.00-15.95	14.99	14.08-16.00	14.99	14.00-15.99	14.99	14.01-15.96
B	6.72	6.29-7.17	6.71	6.27-7.15	6.12	5.67-6.58	7.09	6.61-7.60
C	3.13	2.93-3.34	2.76	2.57-2.94	2.02	1.85-2.19	2.39	2.21-2.57
D	1.66	1.55-1.78	0.96	0.90-1.03	0.62	0.55-0.70	1.25	1.13-1.36
E	1.63	1.52-1.74	1.57	1.47-1.68	0.92	0.84-1.00	1.64	1.51-1.76
F	1.35	1.26-1.44	1.18	1.10-1.26	0.77	0.70-0.84	1.53	1.42-1.65
G	1.44	1.35-1.54	1.38	1.29-1.47	0.79	0.72-0.87	1.27	1.17-1.37
H	1.01	0.94-1.08	0.92	0.86-0.99	0.71	0.65-0.78	0.94	0.87-1.03
I	0.8	0.74-0.85	0.77	0.71-0.82	0.55	0.49-0.61	0.69	0.62-0.75
J	0.58	0.53-0.62	0.68	0.63-0.73	0.34	0.29-0.40	0.46	0.41-0.52
K	0.92	0.66-0.99	0.66	0.61-0.70	0.24	0.20-0.28	0.51	0.45-0.57

To investigate the relationship between Australian and Asian species we used the set ((*O. rufipogon*, *O. nivara*), Taxon A, *O. punctata*). For this set a negative D-statistic value would suggest introgression between Taxon A and the *japonica*/*O. rufipogon* clade, whereas a positive value would suggest introgression between Taxon A and the *indica*/*O. nivara* clade. Four chromosomes (1, 2, 4, and 12) returned significant positive values, whereas no negative value was meaningful. Introgression between the Asian and Australian populations was suggested with evidence for greater introgression between Taxon A and the *indica*/*O. nivara* clade than between Taxon A and the *japonica*/*O. rufipogon* clade.

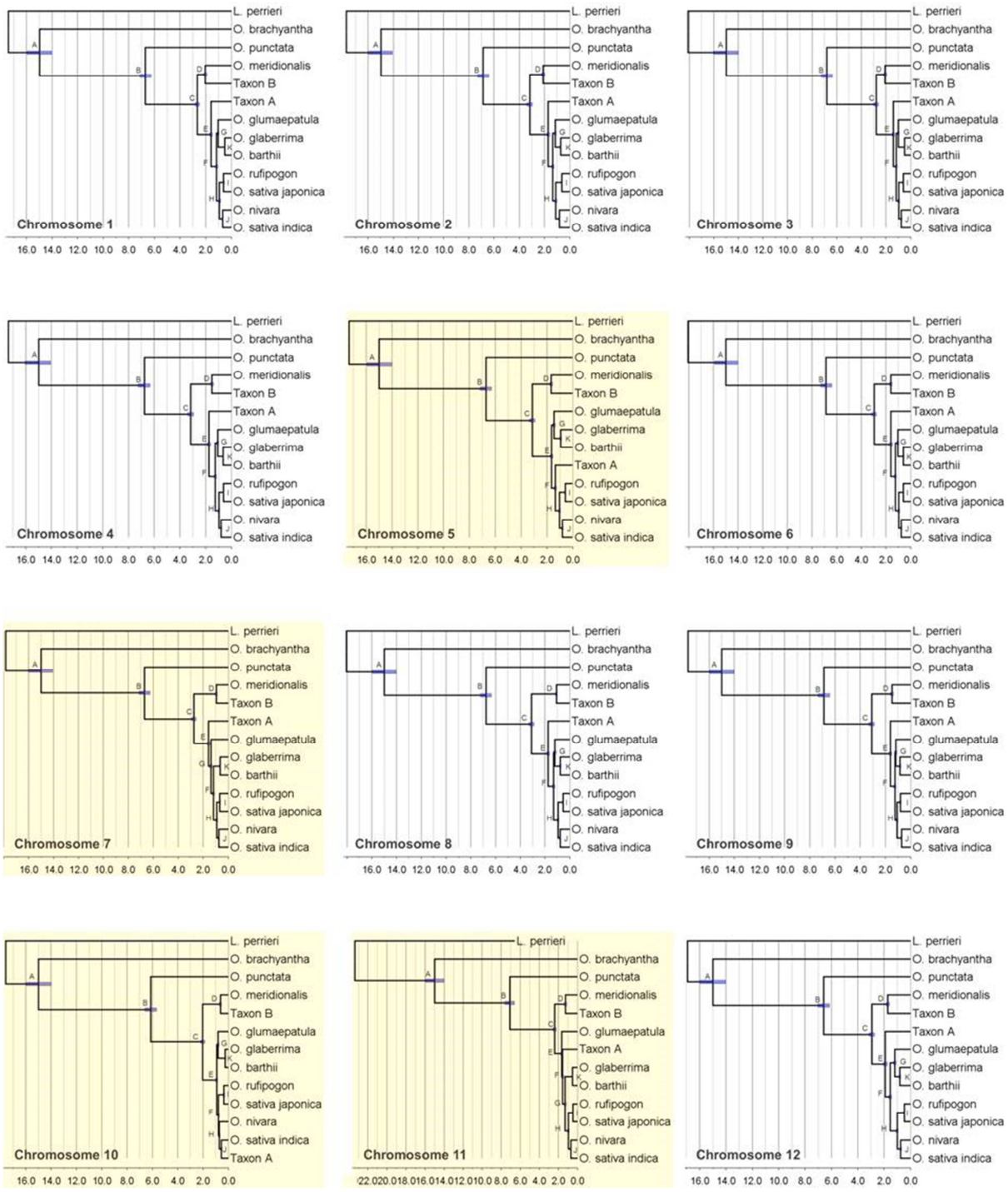


Figure 5.5 Bayesian phylogenies for each chromosome of investigated *Oryza* species and the ourgroup. Tree topologies for chromosomes 1, 2, 3, 4, 6, 8, 9 and 12 are identical. Tree topologies for chromosomes 5, 7, 10 and 11 are distinct and are highlighted in yellow. Scale axis represents age in million years (mya). Node bars display 95% Highest Posterior Density (HPD) interval.

Table 5.17 Global statistics calculated in four-taxon test for selected *Oryza* species per chromosome.

Four taxon test for a set (((*O. rufipogon, O. barthii*), Taxon A), *O. punctata*)						
	Chr 1	**Chr 2**	**Chr 3**	**Chr 4**	**Chr 5**	**Chr 6**
Base pairs	885,116	996,507	805,108	706,297	658,924	551,896
Polymorphic sites	71,201	96,385	61,486	74,967	70,350	52,933
ABBA	1,278	2,853	772	2,018	1,449	1,096
BABA	1,787	3,341	1,888	2,275	2,872	1,813
D-statistic	-0.16607	-0.07879	-0.41955	-0.05986	-0.32932	-0.2465
Z-score	-1.9633	-0.72932	-3.74382	-0.6485	-3.42562	-1.9795
	Chr 7	**Chr 8**	**Chr 9**	**Chr 10**	**Chr 11**	**Chr 12**
Base pairs	533,899	350,596	345,315	93087	134,216	211,890
Polymorphic sites	55,811	41,850	36,600	5588	12,562	19,252
ABBA	1,781	313	783	206	133	244
BABA	972	1,580	1,557	186	288	822
D-statistic	0.29386	-0.66931	-0.33077	0.05102	-0.36817	-0.5422
Z-score	2.55576	-10.03458	-2.85748	0.1037	-2.51294	-3.6627

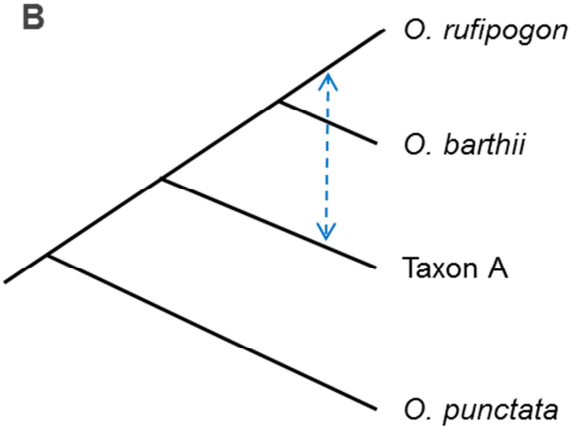
Four taxon test for a set (((*O. rufipogon, O. nivara*), Taxon A), *O. punctata*)						
	Chr 1	**Chr 2**	**Chr 3**	**Chr 4**	**Chr 5**	**Chr 6**
Base pairs	885,116	996,507	805108	706,297	658924	551896
Polymorphic sites	68,110	90,534	57,274	70,570	64,356	50,122
ABBA	1,638	2,363	852	1,821	1,319	1,227
BABA	957	1,417	1,090	848	2,012	1,028
D-statistic	0.26243	0.25026	-0.12255	0.36456	-0.20805	0.08825
Z-score	2.58619	2.37828	-0.99487	2.85723	-1.79152	0.56591
	Chr 7	**Chr 8**	**Chr 9**	**Chr 10**	**Chr 11**	**Chr 12**
Base pairs	533899	350596	345315	93087	134216	211,890
Polymorphic sites	51,798	37,595	33,296	5,682	11,661	18,963
ABBA	1,595	761	581	53	91	299
BABA	1,136	595	696	79	119	184
D-statistic	0.16807	0.12242	-0.09005	-0.19697	-0.13333	0.23810
Z-score	1.14980	0.66700	-0.74602	-0.50227	-1.51219	2.73644

Columns marked in green and red indicate significant Z-score for negative and positive D-statistic, respectively.

A

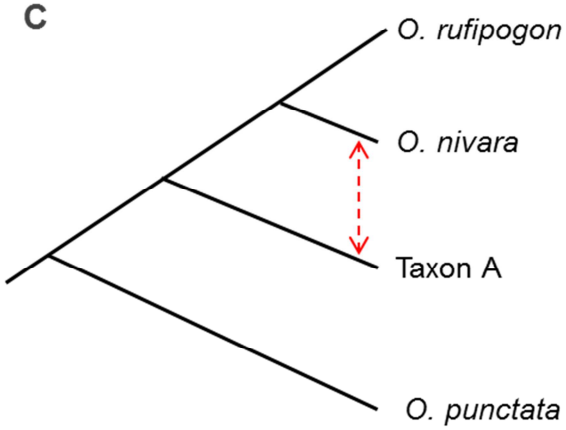
Chromosome	1	2	3	4	5	6	7	8	9	10	11	12
(((*O. rufipogon, O. barthii*), Taxon A), *O. punctata*)												
D-statistic	-0.167	-0.08	-0.42	-0.06	-0.33	-0.25	0.29	-0.67	-0.33	0.05	-0.37	-0.54
Z-score	-1.96*	-0.73	-3.74*	-0.65	-3.43*	-1.98*	2.56*	-10.03*	-2.86*	0.10	-2.51*	-3.66*
(((*O. rufipogon, O. nivara*), Taxon A), *O. punctata*)												
D-statistic	0.26	0.25	-0.12	0.37	-0.21	0.09	0.17	0.12	-0.09	-0.20	-0.13	0.24
Z-score	2.59*	2.38*	-1.00	2.86*	-1.79	0.57	1.15	0.67	-0.75	-0.50	-1.51	2.74*

B



D < 0 * chromosomes: 1, 3, 5, 6, 8, 9, 11, 12

C



D > 0 * chromosomes: 1, 2, 4, 12

Figure 5.6 Results of four-taxon test for Taxon A and selected *Oryza* species. (A) D-statistics and Z-scores calculated for two sets of selected species per chromosome: (((*O. rufipogon, O. barthii*), Taxon A), *O. punctata*) and (((*O. rufipogon, O. nivara*), Taxon A), *O. punctata*). Z-scores marked with asterisk (*) indicate statistically significant values. (B) Four-taxon tree used in the first test. Bidirectional arrow shows inferred introgression for chromosomes with significant D-statistics (listed below the tree). (C) Four-taxon tree used in the second test. Bidirectional arrows show inferred introgression for chromosomes with significant D-statistics (listed below the tree).

5.5 Discussion

The discovery of two novel Australian wild rice taxa expands the understanding of the genetic diversity within the genus *Oryza*. The two draft genomes generated in this study produce an excellent platform for exploring the potential of Australian wild rices. These genomic resources will complement the already extensive study of numerous wild and cultivated species within *Oryza* (Stein *et al.*, personal communication) providing unique data for comparative genomics, evolutionary studies of the entire genus and widening the species pool for rice improvement.

The sequencing and assembly performed in this study show the particular challenges of generating genome assemblies for wild heterozygous plants. It also shows that the replacement of short read data (Illumina) with long read data (PacBio) can

improve the overall completeness of a draft genome sequence. The results suggest that the best representation of the Taxon A and Taxon B genomes was obtained using the PacBio-only data. As a consequence, the PacBio-only assemblies for both taxa have been selected for further analysis. PacBio-only assembly improved almost all metrics, such as number of scaffolds, total size, the longest and the shortest scaffolds, mean and median lengths as well as completeness measured as presence or absence of orthologous genes.

The evolution of rice has been the subject of ongoing debate in particular regarding whether there was a single or multiple domestication in Asia (Bouchenak-Khelladi *et al.* 2010; Kellogg 2009; Vaughan *et al.* 2008). The distinctness of the *indica* and *japonica* genomes suggests separate origins for most of each genome (Wei *et al.* 2012). However the presence of many shared domestication related alleles has led to suggestions that some level of introgression between the two genomes has also been a feature of their domestication history (Huang *et al.* 2012a; Huang *et al.* 2012b; Molina *et al.* 2011). Geographic separation may have allowed early populations to diverge resulting in distinct *O. rufipogon*-like populations in Asia and Australia. Chloroplast transfer or capture is common between recently diverged plant taxa which may explain the distinct Asian and Australian chloroplast genomes in the wild populations descended from the taxa that were domesticated in Asia.

Wild rice populations are a key genetic resource for rice improvement (Anacleto *et al.* 2015). The Australian populations may provide an especially useful resource for evaluation of rice domestication due to their isolation from significant impact of gene flow from domesticated rice. An improved understanding of the geographic variation in A genome wild rice species provided by genome sequencing should guide the search for useful alleles in wild populations (Krishnan *et al.* 2014). The taxonomy of the A genome wild rice species in Australia and Asia needs to be re-evaluated in the light of the molecular data now available to determine if distinct wild taxa need to be recognized. Movement of flora between Sahul (New Guinea and Australia) and Sunda (Malay Peninsula, Sumatra, Borneo, Java) in both directions may be an important part of the evolutionary history of the A genome *Oryza* species (Crayn *et al.* 2015; Prasad *et al.* 2011; Tang *et al.* 2010). The current diversity of A genome *Oryza* in northern Australia suggests the possibility of an Australian and or south east Asian origin for the A genome clade, but further historical biogeographical analyses based on more extensive datasets are required to evaluate this hypothesis.

Chapter 6

General Discussion

6.1 Fulfilment of objectives

The motivation for this study has been a recent discovery and identification of two distinct types of perennial wild rice in northern Australia (Sotowa *et al.* 2013). It was suggested that Australian species that had been described as *O. rufipogon* might be indeed two different taxa. In the previous study, they were characterised based on morphological characteristics and a small number of loci from chloroplast and nuclear genomes. In the current study, these populations have been described in more detail by reporting their chloroplast and nuclear genome sequences delivered by state of the art genomic technologies and analysis methods.

In this thesis research, Chapters 3 and 4, focused on chloroplast sequence investigation, whereas Chapter 5 was devoted to whole genome sequence analysis. Chapter 3 examined the reliability of two distinct next generation sequencing platforms coupled with diverse data analysis methods when applied to creating full chloroplast sequences of rice species with application to other plant species. In Chapter 4, the chloroplast sequences of Australian rices generated as a part of the project were annotated, compared with other available *Oryza* chloroplast genomes and employed in inference of evolutionary relationships between these and other rice species. Chapter 5 covered sequencing, assembly and annotation of the whole genomes of the two Australian wild rice taxa applying, as in Chapter 3, two different high throughput sequencing technologies and different data analysis approaches. This chapter also utilized these genome sequences to analyse the genome-wide phylogeny of numerous *Oryza* species, including speciation times, with an emphasis on the Australian taxa.

The current chapter reviews and discusses key findings as well as indicates future directions for more thorough characterisation of Australian wild rice populations.

6.2 Chloroplast genomes of Australian wild rice species

The objective of generating the chloroplast genomes of Australian wild rices in this study was to utilize them as barcodes to investigate molecular taxonomy. Nevertheless, these genomes can also be used for plant identification, ecological studies and food purity assessment. Advances in DNA sequencing have made it possible to sequence large genomes, transcriptomes and exomes or even small genomes such as the chloroplast. However, every sequencing platform performs better for certain applications and not all can be used with a similar success for the same exercise. In the first part of this thesis (Chapter 3) two next generation sequencing platforms were assessed for use in the specific application of whole chloroplast sequencing. The sequencing platforms used were Illumina and Ion Torrent. Herein, only two Illumina instruments were tested (HiSeq2000 and GAIIx) but this company offers a broad range of instruments, from large high-throughput sequencers able to sequence thousands of genomes per year (HiSeq X Ten) to bench-top machines capable of sequencing smaller genomes in faster modes (MiSeq). With this extensive portfolio of instruments Illumina covers virtually all of the NGS applications, from human whole genome sequencing, through RNA- and ChIP-seq to metagenomics. Illumina has been dominating the next-generation sequencing market for many years now. This success is mainly owed to high read accuracy, low error rate, high yield of data per run and low cost per Gb of data. The Ion Torrent instrument tested was a Personal Genome Machine (PGM). The main applications of Ion Torrent systems have been small genomes and targeted sequencing. The output from the PGM has expanded rapidly and it is the cheapest sequencer available on the market (approximately US$50k). Additionally, it allows very fast runs (~2 hours) and requires inexpensive consumables. However, it generates significantly lower output than Illumina with considerably lower read accuracy and higher error rates.

One of the key requirements for a practical analysis method in assessing a genome assembly is the availability of a reference sequence. For rice, a complete and high-quality reference sequence is available and has been used throughout the study in several occasions. In this case, thanks to the *O. sativa* ssp. *japonica* var. Nipponbare chloroplast genome, it was possible to accurately assess the sequencing platforms and analysis methods. As a consequence, one approach for creating the chloroplast consensus sequences involved the use of the reference (read mapping), while in the second the sequences were created without this support (*de novo* assembly). In the read mapping

method, chloroplast consensus sequences were obtained from data from both sequencing platforms. However, each of them exhibited certain errors. Illumina reads were highly accurate in terms of reconstructing the reference sequences but missed long variations (insertions and deletions) in the Australian genotype. On the other hand, these indels were correctly called using the Ion Torrent data; however the overall consensus accuracy was significantly lower predominantly due to short indel errors in the final sequence. Ion Torrent was reliable for *de novo* assembly allowing generation of a full length chloroplast genome with no gaps but not fully accurate with a considerable number of short indels. Illumina data did not prove suitable for *de novo* assembly where the full length consensus sequence was obtained for neither of genotypes because of gaps present between the contigs. Not surprisingly, the read length (long Ion Torrent vs. shot Illumina reads) was the main factor influencing contiguity of *de novo* genome assembly.

Generating a *de novo* sequence for an organism of interest is particularly important if there is no reference available from the same species. Although a close relative from the same genera might be used in a mapping-only approach, some longer polymorphisms can be overlooked. This was exemplified here by the long indels not detected by Illumina read mapping. *De novo* assembly can significantly improve the result and avoid misassemblies in a mapping-only consensus. As demonstrated, in order to generate an accurate chloroplast genome a combination of sequencing technologies and data analysis should be used. The advances in sequencing technologies, including the cost as well as ease of analysis, make barcoding by sequencing of the whole chloroplast a reliable, fast and inexpensive method that, in consequence, has application in plant identification in many research and industrial applications.

Knowledge of the phylogenetic relatedness between and within Australian populations, especially now in the light of the new discoveries, is important for effective use of wild genetic resources. The phylogenetic relationships inferred here based on chloroplast genomes were the first attempt to clarify the relationships between the two novel Australian taxa and other, both wild and cultivated, A genome rice species from Asia and Australia. In this study we confirmed the distinctness of Asian and Australian wild rices as shown by Waters *et al.* (2012). All of the Australian species, including the two new taxa, formed a monophyletic group with a clear distinction between them; specifically Taxon A and Taxon B were shown to be different from one another. Furthermore, for the first time, we showed that, despite very close morphological resemblance between Asian *O. rufipogon* and Australian Taxon A, they are distinct species. Moreover, Taxon B and

O. meridionalis were found to be more closely related to each other than to Taxon A. This chloroplast analysis was the first indication that these taxa could be new Australian wild rice species. The results from chloroplast genome investigation and the discovery of distinctness between the two novel Australian wild rice populations and Asian species became the drive for exploration of the evolutionary relationships between these species based on their nuclear genomes.

6.3 Nuclear genomes of Australian wild rice taxa

Whole nuclear genomes of Australian wild rice taxa were created by combination of data from two high throughput sequencing platforms, namely Illumina (described in detail previously) and Pacific Biosciences (PacBio). Although the output from the PacBio machine is comparatively low, the very long read lengths (with a mean of ~15k) make this platform an attractive option for a number of applications. It has become a good fit for *de novo* genome assembly, closing gaps in already existing sequences by sequencing difficult regions, large structural variation or full length transcriptome discovery. An interesting and unusual feature of PacBio is its use in epigenetic studies by detection of modified nucleotides. Other advantages of this technology include relatively quick sequencing runs and ability to sequence high GC content regions where other platforms struggle. Unfortunately, the read accuracy is low and, therefore, the overall error rate is relatively high.

Exploring the genome characteristics before the assembly, including genome size estimation, is a good way of discovering how challenging the assembly might be. This approach was adapted in this study leading to discovery of a higher heterozygosity and repetitive content rates for Taxon B compared to Taxon A. Accordingly, Taxon B assembly was more challenging and resulted in a slightly lower quality of the final sequence in terms of basic statistics and completeness. Out of the two different assembly strategies, hybrid and PacBio-only, the latter was proven to be more appropriate to produce more complete genome assemblies for both taxa. Addition of short read data (Illumina) and performing a hybrid assembly did not improve the final sequences. On the contrary, it caused an inflated assembly of one of the taxa (Taxon B).

The whole A genome *Oryza* group phylogeny has been recently resolved using chloroplast (Wambugu *et al.* 2015) and nuclear genomes (Stein *et al.*, personal communication); however the latter did not include the two novel Australian taxa. These

taxa were incorporated in the current study showing discrepancy between chloroplast and whole genome phylogenetic tree topologies. The most unexpected difference was the position of Taxon A where it formed a part of the Australian species clade based on the chloroplast genomes but was found closer to the clades with domesticated rices based on the nuclear sequences. Incomplete lineage sorting, interspecific hybridization and coalescent processes have been shown to cause incongruence between chloroplast and nuclear phylogenies (Comes & Abbott 2001; Hudson 1990; Mason-Gamer *et al.* 1995). Herein, we demonstrated a possible introgression into Australian Taxon A from the *indica*/*O. nivara* clade in Asia, which might be responsible for the divergence between chloroplast and genome-wide tree topologies. Another reason might be different evolutionary paths for these two genomes; a case that has been reported for recently diverged taxa (Tsitrone *et al.* 2003).

O. meridionalis and Taxon B were the A genome species that first diverged from other species. This event was estimated to have occurred around 3 million years ago. The next speciation event in this group took place about 1.6 million years ago when Taxon A diverged. The *Oryza* tribe is probably of Gondwanan origin (Bouchenak-Khelladi *et al.* 2010), but according to our estimated speciation times, the *Oryza* A genome speciation took place after the breakup of the supercontinent supporting the hypothesis of long range dispersal of *Oryza* species across the world by humans and animals.

This thesis confirms that Australian perennial wild rice populations comprise at least two types; the m-type, represented by Taxon B, which is morphologically similar to *O. meridionalis*, and the r-type, Taxon A, similar to *O. rufipogon*. Detailed genomic analyses have showed that all Australian A genome rice species carry *meridionalis*-type plastid genomes as revealed by phylogenetic inference based on chloroplast genomes. However, only one of the novel taxa, Taxon B, also carries a *meridionalis*-type nuclear genome, whereas the other taxon, Taxon A, has a nuclear genome that is more closely related to domesticated species as revealed by nuclear phylogenies.

6.4 Future directions

The close relatives of cultivated plants, crop wild relatives, carry a potentially valuable gene pool for use in breeding of domesticated species and their improvement. The genetic diversity present in today's crop plants is scarce and their immediate wild relatives might contribute a reservoir of genetic diversity for these crops. The discovery of new wild rice

populations in Australia, which are close wild relatives of cultivated rice, expands the effective gene pool for rice available for rice improvement programmes.

The draft genomes created in this study provide the first detailed genomic resources for the Australian Taxa A and B and are the first stage in unveiling the genetic potential of these Australian wild rice populations. Future steps should focus on mining and thorough examination of agronomically important traits. Tolerance to abiotic and resistance to biotic stresses as well as yield enhancing genes are of particular value in breeding programmes. However, establishing a high quality genome assembly is essential for consistent and exhaustive whole genome exploration. Given the current status of the Australian rice genomes, which are in thousands of scaffolds ordered into pseudochromosomes based on reference sequences, there is a need for improvement of the overall quality and completeness of these sequences. A few common strategies exist for improving genome assemblies that could be applied to the draft genomes of the Australian taxa. One method, typically used in genomics for closing gaps in genomes and or placing scaffolds onto chromosomes, is genome or genetic mapping. Genome maps represent the physical distance between genome positions measured in base pairs [physical maps; Deonier (2005)] while genetic maps determine the relative order between various markers or genes based on their recombination frequencies (Tiwari *et al.* (2016)]. Recent technological advances have allowed the generation of optical maps (Schwartz *et al.* 1993), which are a type of high-resolution restriction map where the size and location of individual restriction fragments is determined. The most suitable approach to mapping the Australian wild rice taxa genomes would be a single-nucleotide polymorphism (SNP) map that is one the most straightforward and most reliable ways to map genome loci (Davis & Hammarlund 2006). SNPs occur in high-density and are easy to identify by whole genome sequencing. In this study, scaffolds were placed onto pseudochromosomes based on the available reference genomes. Having a map would improve the assemblies by orienting and ordering scaffolds into chromosome pseudomolecules based on the true relationships between them. Consequently, potential large-scale structural variations between theses taxa and other *Oryza* species could be discovered and investigated.

Discovery of potentially new and valuable alleles relies on the quality of functional annotation of the genome. The results from gene space annotation (low numbers of predicted gene models) have indicated the need for improvement of protein coding gene annotation. This could be facilitated by sequencing the mRNA transcripts from wild rice samples by means of whole transcriptome shotgun sequencing, also called RNA-seq. The

annotation pipeline adopted here requires expression evidence to be able to predict gene models. Collecting more evidence in the form of full-length transcripts would positively influence the numbers of gene models detected and the overall success of the wild rice genome annotation.

Molecular and morphological evidence gathered in this and other studies (Sotowa *et al.* 2013; Tikapunya *et al.* 2016; Wambugu *et al.* 2015) clearly indicates that the two recently recognized Australian wild rice taxa might be new species that are not only distinct from the species they were previously defined as, *O. rufipogon*, but also from one another. At this point, these taxa require a formal taxonomic description, scientific naming and publishing as new species. However, the description of a newly discovered species is a task for a specialist and further collaborations will be necessary to achieve this according to established taxonomic guidelines.

The findings of this thesis show that tropical Australia may hold more genetic diversity within wild rice populations than had been previously recognized. Herein only two individuals coming from two different sites have been characterised. Therefore, there is a need for further collections and descriptions of more populations in order to determine their distribution across northern Australia as well as variations between and within the populations. Great opportunity exists for discovery of new sites where wild rice is present, new taxa and, possibly, more new species with the ultimate goal of protection and conservation, both *ex situ* and *in situ*, of those populations.

Chapter 7

References

Abberton M, Batley J, Bentley A*, et al.* (2016) Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnology Journal* **14**, 1095-1098.

Aggarwal RK, Brar DS, Nandi S, Huang N, Khush GS (1999) Phylogenetic relationships among *Oryza* species revealed by AFLP markers. *Theoretical and Applied Genetics* **98**, 1320-1328.

Albers CA, Lunter G, MacArthur DG*, et al.* (2011) Dindel: accurate indel calls from short-read data. *Genome Research* **21**, 961-973.

Ammiraju JSS, Fan C, Yu Y*, et al.* (2010a) Spatio-temporal patterns of genome evolution in allotetraploid species of the genus *Oryza*. *The Plant Journal* **63**, 430-442.

Ammiraju JSS, Luo M, Goicoechea JL*, et al.* (2006) The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Research* **16**, 140-147.

Ammiraju JSS, Song X, Luo M*, et al.* (2010b) The *Oryza* BAC resource: a genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. *Breeding Science* **60**, 536-543.

Anacleto R, Cuevas RP, Jimenez R*, et al.* (2015) Prospects of breeding high-quality rice using post-genomic tools. *Theoretical and Applied Genetics* **128**, 1449-1466.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815.

Bajgain P, Rouse MN, Tsilo TJ*, et al.* (2016) Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PLoS ONE* **11**, e0155760.

Bao Y, Ge S (2004) Origin and phylogeny of *Oryza* species with the CD genome based on multiple-gene sequence data. *Plant Systematics and Evolution* **249**, 55-66.

Bell DC, Thomas WK, Murtagh KM*, et al.* (2012) DNA base identification by electron microscopy. *Microscopy and Microanalysis* **18**, 1049-1053.

Berlin K, Koren S, Chin CS*, et al.* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* **33**, 623-630.

Bernatzky R, Tanksley SD (1986) Genetics of actin-related sequences in tomato. *Theoretical and Applied Genetics* **72**, 314-321.

Bininda-Emonds ORP (2004) *Phylogenetic supertrees: combining information to reveal the tree of life.* Kluwer Academic Publishers, Dordrecht, Boston.

Bouchenak-Khelladi Y, Verboom GA, Savolainen V, Hodkinson TR (2010) Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. *Botanical Journal of the Linnean Society* **162**, 543-557.

Bouckaert R, Heled J, Kühnert D*, et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10**, e1003537.

Boutet G, Alves Carvalho S, Falque M*, et al.* (2016) SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. *BMC Genomics* **17**, 1-14.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLOS Computational Biology* **9**, e1003031.

Brar DS, Khush GS (1997) Alien introgression in rice. *Plant Molecular Biology* **35**, 35-47.

Brar DS, Singh K (2011) Oryza. In: *Wild crop relatives: genomic and breeding resources, cereals.* (ed. Kole C), pp. 321–365. Springer, Berlin.

Brenchley R, Spannagl M, Pfeifer M*, et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705-710.

Brocchieri L (2001) Phylogenetic inferences from molecular sequences: review and critique. *Theoretical Population Biology* **59**, 27-40.

Brozynska M, Furtado A, Henry RJ (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal* **14**, 1070-1085.

Brozynska M, Omar ES, Furtado A*, et al.* (2014) Chloroplast genome of novel rice germplasm identified in Northern Australia. *Tropical Plant Biology* **7**, 111-120.

Campbell MS, Law M, Holt C*, et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology* **164**, 513-524.

Cao ZH, Ding JL, Hu ZY*, et al.* (2006) Ancient paddy soils from the Neolithic age in China's Yangtze River Delta. *Naturwissenschaften* **93**, 232-236.

Carroll BJ, Klimyuk VI, Thomas CM*, et al.* (1995) Germinal transpositions of the maize element Dissociation from T-DNA loci in tomato. *Genetics* **139**, 407-420.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. *Evolution* **32**, 550–570.

Chang TT (1976) The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. *Euphytica* **25**, 425-441.

Chen J, Huang Q, Gao D*, et al.* (2013) Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying Oryza genome evolution. *Nature Communications* **4**, 1595.

Civáň P, Craig H, Cox CJ, Brown TA (2015) Three geographically separate domestications of Asian rice. *Nature Plants* **1**, 15164.

Clarke J, Wu HC, Jayasinghe L*, et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**, 265-270.

Clifford HT, Simon BK (1981) The biogeography of Australian grasses. In: *Ecological biogeography of Australia* (ed. Keast A). W. Junk, The Hague.

Comes HP, Abbott RJ (2001) Molecular phylogeography, reticulation, and lineage sorting in Mediterranean Senecio sect. Senecio (Asteraceae). *Evolution* **55**, 1943-1962.

Copetti D, Zhang J, El Baidouri M*, et al.* (2015) RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**, 1-10.

Crayn DM, Costion C, Harrington MG, Richardson J (2015) The Sahul–Sunda floristic exchange: dated molecular phylogenies document Cenozoic intercontinental dispersal dynamics. *Journal of Biogeography* **42**, 11-24.

Cronn R, Liston A, Parks M*, et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* **36**, e122.

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

Davis MW, Hammarlund M (2006) Single-Nucleotide Polymorphism mapping. In: *C. elegans: Methods and Applications* (ed. Strange K), pp. 75-92. Humana Press, Totowa, NJ.

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. *Trends in Ecology & Evolution* **22**, 34-41.

Deonier RC (2005) Physical mapping of DNA. In: *Computational Genome Analysis: An Introduction*, pp. 99-119. Springer New York, New York, NY.

Diaz-Sanchez S, Hanning I, Pendleton S, D'Souza D (2013) Next-generation sequencing: the future of molecular genetics in poultry production and food safety. *Poultry Science* **92**, 562-572.

Doi K, Yasui H, Yoshimura A (2008) Genetic variation in rice. *Current Opinion in Plant Biology* **11**, 144-148.

Doust AN, Kellogg EA, Devos KM, Bennetzen JL (2009) Foxtail millet: a sequence-driven grass model system. *Plant Physiology* **149**, 137-141.

Duan S, Lu B, Li Z*, et al.* (2007) Phylogenetic analysis of AA-genome *Oryza* species (Poaceae) based on chloroplast, mitochondrial, and nuclear DNA sequences. *Biochemical Genetics* **45**, 113-129.

Duistermaat H (1987) A revision of *Oryza* (Gramineae) in Malaysia and Australia. *Blumea* **32**, 157-193.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239-2252.

Earl D, Bradnam K, St John J*, et al.* (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research* **21**, 2224-2241.

Eid J, Fehr A, Gray J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Ferrarini M, Cestaro A, Sargent DJ*, et al.* (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* **14**, 670-670.

Fierst JL (2015) Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics* **6**, 220.

Fleischmann R, Adams M, White O*, et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in *De Novo* annotation approaches. *PLoS ONE* **6**, e16526.

Food and Agriculture Organization (2016) *http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf.*

Fujii S, Kazama T, Yamada M, Toriyama K (2010) Discovery of global genomic re-organization based on comparison of two newly sequenced rice mitochondrial genomes with cytoplasmic male sterility-related genes. *BMC Genomics* **11**, 209-209.

Fuller DQ (2006) Agricultural origins and frontiers in South Asia: a working synthesis. *Journal of World Prehistory* **20**, 1-86.

Fuller DQ, Sato YI, Castillo C*, et al.* (2010) Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeological and Anthropological Sciences* **2**, 115-131.

Furtado A (2014) DNA extraction from vegetative tissue for next-generation sequencing. In: *Cereal Genomics. Methods in Molecular Biology* (eds. Henry RJ, Furtado A), p. 1. Springer Science+Business Media, New York.

Galimberti A, De Mattia F, Losa A*, et al.* (2013) DNA barcoding as a new tool for food traceability. *Food Research International* **50**, 55-63.

Gao LZ, Innan H (2008) Nonindependent domestication of the two rice subspecies, *Oryza sativa* ssp. *indica* and ssp. *japonica*, demonstrated by multilocus microsatellites. *Genetics* **179**, 965-976.

Garg R, Verma M, Agrawal S*, et al.* (2014) Deep transcriptome sequencing of wild halophyte rice, *Porteresia coarctata*, provides novel insights into the salinity and submergence tolerance factors. *DNA Research* **21**, 69-84.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169**, 1631-1638.

Ge S, Li A, Lu BR, Zhang SZ, Hong DY (2002) A phylogeny of the rice tribe Oryzeae (Poaceae) based on matK sequence data. *American Journal of Botany* **89**, 1967-1972.

Ge S, Sang T, Lu BR, Hong DY (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences USA* **96**, 14400–14405.

Ge S, Sang T, Lu BR, Hong DY (2001) Phylogeny of the genus *Oryza* as relealed by molecular approaches. In: *Rice genetics IV* (eds. Khush GS, Brar DS, Hardy B), pp. 89-105. Science Publishers, Inc., Enfield, NH, USA.

Giuliani R, Koteyeva N, Voznesenskaya E*, et al.* (2013) Coordination of leaf photosynthesis, transpiration, and structural traits in rice and wild relatives (genus *Oryza*). *Plant Physiology* **162**, 1632-1651.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources* **11**, 759-769.

Global Rice Science Partnership (2010) *http://africarice.org/publications/GRiSP.pdf.*

Goff SA, Ricke D, Lan TH*, et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92-100.

Goff SA, Vaughn M, McKay S*, et al.* (2011) The iPlant Collaborative: cyberinfrastructure for plant biology. *Frontiers in Plant Science* **2**, 34.

Gramazio P, Blanca J, Ziarsolo P*, et al.* (2016) Transcriptome analysis and molecular marker discovery in *Solanum incanum* and *S. aethiopicum*, two close relatives of the common eggplant (*Solanum melongena*) with interest for breeding. *BMC Genomics* **17**, 1-17.

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. *The Plant Journal* **66**, 34-44.

Green RE, Krause J, Briggs AW*, et al.* (2010) A draft sequence of the Neandertal genome. *Science* **328**, 710-722.

Guang A, Zapata F, Howison M, Lawrence CE, Dunn CW (2016) An integrated perspective on phylogenetic workflows. *Trends in Ecology & Evolution* **31**, 116-126.

Guo YL, Ge S (2005) Molecular phylogeny of *Oryzeae* (*Poaceae*) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *American Journal of Botany* **92**, 1548–1558.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075.

Hartwig B, James GV, Konrad K, Schneeberger K, Turck F (2012) Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiology* **160**, 591-600.

He R, Salvato F, Park JJ*, et al.* (2014) A systems-wide comparison of red rice (*Oryza longistaminata*) tissues identifies rhizome specific genes and proteins that are targets for cultivated rice improvement. *BMC Plant Biology* **14**, 1-21.

He Z, Zhai W, Wen H*, et al.* (2011) Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genetics* **7**.

Henry RJ (2001) *Plant genotyping: the DNA fingerprinting of plants* CABI Pub, New York.

Henry RJ (2012) Next-generation sequencing for understanding and accelerating crop domestication. *Briefings in Functional Genomics* **11**, 51-56.

Henry RJ (2013) *Molecular Markers in Plants* John Wiley & Sons, Hoboken, N.J.

Henry RJ, Rice N, Waters DLE*, et al.* (2010) Australian *Oryza*: Utility and Conservation. *Rice* **3**, 235-241.

Higman C, Lu TLD (1998) The origins and dispersal of rice cultivation. *Antiquity* **72**, 867 - 877.

Hiratsuka J, Meng BY, Li YQ*, et al.* (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome - intermolecular recombination between distinct transfer-rna genes accounts for a major plastid dna inversion during the evolution of the cereals. *Molecular and General Genetics* **217**, 185-194.

Ho SYW (2008) The molecular clock and estimating species divergence. *Nature Education* **1**, 168.

Ho SYW, Duchêne S (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology* **23**, 5947-5965.

Hou L, Chen X, Wang M*, et al.* (2015) Mapping a large number of QTL for durable resistance to stripe rust in winter wheat druchamp using SSR and SNP markers. *PLoS ONE* **10**, e0126794.

Huang PU, Molina J, Flowers JM*, et al.* (2012a) Phylogeography of Asian wild rice, *Oryza rufipogon*: a genome-wide view. *Molecular Ecology* **21**, 4593-4604.

Huang X, Kurata N, Wei X*, et al.* (2012b) A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497-501.

Huang X, Wei X, Sang T*, et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* **42**, 961-967.

Hudson RR (1990) Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds. Havey PH, Partridge L), pp. 1-44. Oxford University Press, New York.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**, 793-800.

Ishii T, Brar DS, Multani DS, Khush GS (1994) Molecular tagging of genes for brown planthooper resistance and earliness in introgressed from *Oryza australiensis* into cultivated rice, *Oryza sativa*. *Genome* **37**, 217-221.

Ishii T, Nakano T, Maeda H, Kamijina O (1996) Phylogenetic relationships in A-genome species of rice as revealed by RAPD analysis. *Genes and Genetic Systems* **71**, 195-201.

Iwamatsu Y, Aoki C, Takahashi M*, et al.* (2008) UVB sensitivity and cyclobutane pyrimidine dimer (CPD) photolyase genotypes in cultivated and wild rice species. *Photochemical & Photobiological Sciences* **7**, 311-320.

Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M, Hayashi T (2016) Genomics-assisted breeding in fruit trees. *Breeding Science* **66**, 100-115.

Jacquemin J, Bhatia D, Singh K, Wing RA (2013) The International *Oryza* Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Current Opinion in Plant Biology* **16**, 1-10.

Jansen RK, Raubeson LA, Boore JL*, et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. In: *Methods in Enzymology* (eds. Elizabeth AZ, Eric HR), pp. 348-384. Academic Press.

Jena KK, Jeung JU, Lee JH, Choi HC, Brar DS (2006) High-resolution mapping of a new brown planthopper (BPH) resistance gene, *Bph18(t)*, and marker-assisted selection for BPH resistance in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* **112**, 288-297.

Jeung JU, Kim BR, Cho YC*, et al.* (2007) A novel gene, *Pi40(t)*, linked to the DNA markers derived from NBS-LRR motifs confers broad spectrum of blast resistance in rice. *Theoretical and Applied Genetics* **115**, 1163-1177.

Jones P, Binns D, Chang H*, et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240.

Joshi SP, Gupta VS, Aggarwal RK, Ranjekar PK, Brar DS (2000) Genetic diversity and phylogenetic relationship as revealed by inter simple sequence repeat (ISSR) polymorphism in the genus *Oryza*. *Theoretical and Applied Genetics* **100**, 1311-1320.

Kasem S, Waters DLE, Rice N, Shapter FM, Henry RJ (2010) Whole grain morphology of Australian rice species. *Plant Genetic Resources* **8**.

Kawahara Y, de la Bastide M, Hamilton J*, et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4.

Kaya HB, Cetin O, Kaya HS*, et al.* (2016) Association mapping in Turkish olive cultivars revealed significant markers related to some important agronomic traits. *Biochemical Genetics* **54**, 1-28.

Kececioglu JD, Myers EW (1995) Combinatorial algorithms for DNA sequence assembly. *Algorithmica* **13**, 7-51.

Kellogg EA (2009) The evolutionary history of Ehrhartoideae, Oryzeae, and *Oryza*. *Rice* **2**, 1-14.

Kharabian-Masouleh A, Waters DLE, Reinke RF, Ward R, Henry RJ (2012) SNP in starch biosynthesis genes associated with nutritional and functional properties of rice. *Scientific Reports* **2**, 557.

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. *Plant Molecular Biology* **35**, 25-34.

Kim H, Hurwitz B, Yu Y*, et al.* (2008) Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. *Genome Biology* **9**, R45.

Kokoris MS, McRuer RN (2008) High throughput nucleic acid sequencing by expansion. US Patent 7, 939, 259.

Körber N, Bus A, Li J*, et al.* (2016) Agronomic and seed quality traits dissected by genome-wide association mapping in *Brassica napus*. *Frontiers in Plant Science* **7**, 386.

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59-59.

Krishnan SG, Daniel LEW, Henry RJ (2014) Australian wild rice reveals pre-domestication origin of polymorphism deserts in rice genome. *PLoS ONE* **9**, e98843.

Kurtz S, Phillippy A, Delcher AL*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biology* **5**, 12-R12.

Kwon SJ, Lee JK, Hong SW*, et al.* (2006) Genetic diversity and phylogenetic relationship in AA *Oryza* species as revealed by *Rim2*/*Hipa* CACTA transposon display. *Genes & Genetic Systems* **81**, 93-101.

Lang BF, Burger G (2007) Purification of mitochondrial and plastid DNA. *Nature Protocols* **2**, 652-660.

Lee KH, Park J, Williams DS*, et al.* (2013) Defective chloroplast development inhibits maintenance of normal levels of abscisic acid in a mutant of the Arabidopsis *RH3* DEAD‐box protein during early post‐germination growth. *Plant Journal* **73**, 720-732.

Leite DC, Pinheiro JB, Campos JB, Di Mauro AO, Unêda-Trevisoli SH (2016) QTL mapping of soybean oil content for marker-assisted selection in plant breeding program. *Genetics and Molecular Research* **15**.

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. *Science* **311**, 1936-1939.

Li J, Stein D, McMullan C*, et al.* (2001) Ion-beam sculpting at nanometre length scales. *Nature* **412**, 166-169.

Li ZM, Zheng XM, Ge S (2011) Genetic diversity and domestication history of African rice (*Oryza glaberrima*) as inferred from multiple gene sequences. *Theoretical and Applied Genetics* **123**, 21-31.

Linares OF (2002) African rice (*Oryza glaberrima*): history and future potential. *Proceedings of the National Academy of Sciences USA* **99**, 16360-16365.

Liu L, Lee GA, Jiang L, Zhang J (2007) Evidence for the early beginning (c. 9000 cal. BP) of rice domestication in China: a response. *The Holocene* **17**, 1059–1068.

Liu L, Li Y, Li S*, et al.* (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology* **2012**, 251364.

Liu S, Fan C, Li J*, et al.* (2016) A genome-wide association study reveals novel elite allelic variations in seed oil content of *Brassica napus*. *Theoretical and Applied Genetics* **129**, 1203-1215.

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41**, W575–W581.

Loman NJ, Misra RV, Dallman TJ*, et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**, 434-434.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences USA* **103**, 9578-9583.

Lu F, Ammiraju JSS, Sanyal A*, et al.* (2009) Comparative sequence analysis of *MONOCULM1*-orthologous regions in 14 *Oryza* genomes. *Proceedings of the National Academy of Sciences USA* **106**, 2071-2076.

Luan B, Peng H, Polonsky S*, et al.* (2010) Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Physical Review Letters* **104**, 238103.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology* **46**, 523-536.

Marcotuli I, Houston K, Schwerdt JG*, et al.* (2016) Genetic diversity and genome wide association study of beta-glucan content in tetraploid wheat grains. *PLoS ONE* **11**, e0152590.

Martin C, Juliano A, Newbury HJ*, et al.* (1997) The use of RAPD markers to facilitate the identification of *Oryza* species within a germplasm collection. *Genetic Resources and Crop Evolution* **44**, 175-183.

Martinez C, Arumuganathan K, Kikuchi JH, Earle ED (1994) Nuclear DNA content of ten rice species as determined by flow cytometry. *Japanese Journal of Genetics* **69**, 513-523.

Mason-Gamer RJ, Holsinger KE, Jansen RK (1995) Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae). *Molecular Biology and Evolution* **12**, 371-381.

Masood MS, Nishikawa T, Fukuoka SI*, et al.* (2004) The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. *Gene* **340**, 133-139.

McIntosh SK (1995) *Excavations at Jenne-Jeno, Hambarketolo, and Kaniana (Inland Niger Delta, Mali), the 1981 Season.* University of California Press, Berkerley, US.

McNally B, Singer A, Yu Z*, et al.* (2010) Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Letters* **10**, 2237-2244.

McNally KL, Childs KL, Bohnert R*, et al.* (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences USA* **106**, 12273-12278.

McPherson H, Rossetto M, van der Merwe M*, et al.* (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology* **13**, 8-18.

Meacham F, Boffelli D, Dhahbi J*, et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451-451.

Merchant CA, Healy K, Wanunu M*, et al.* (2010) DNA translocation through graphene nanopores. *Nano Letters* **10**, 2915-2921.

Michener CD, Sokal RR (1956) A quantitative approach to a problem in classification. *Evolution* **11**, 130-162.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327.

Molina J, Sikora M, Garud N*, et al.* (2011) Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences USA* **108**, 8351-8356.

Moore MJ, Bell CD, Soltis DE, Soltis PS (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences USA* **104**, 19363-19368.

Moore MJ, Soltis PS, Soltis DE*, et al.* (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences USA* **107**, 4623-4628.

Morishima H, Oka HI (1960) The pattern of interspecific variation in the genus *Oryza*: its quantitative representation by statistical methods. *Evolution* **14**, 153-165.

Murphy WJ, Eizirik E, O'Brien SJ*, et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348-2351.

Myers EW (2005) The fragment assembly string graph. *Bioinformatics* **21**, ii79-ii85.

Myers EW, Remington KA, Anson EL*, et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* **287**, 2196-2204.

Nagarajan N, Pop M (2009) Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *Journal of Computational Biology* **16**, 897-908.

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935.

Nguyen B, Brar D, Bui B*, et al.* (2003) Identification and mapping of the QTL for aluminum tolerance introgressed from the new source, *Oryza rufipogon*., into indica rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* **106**, 583-593.

Nock CJ, Waters DL, Edwards MA*, et al.* (2011) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* **9**, 328-333.

Notsu Y, Masood S, Nishikawa T*, et al.* (2002) The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Molecular Genetics and Genomics* **268**, 434-445.

Oka H (1974) Experimental studies on the origin of cultivated rice. *Genetics* **78**, 475-486.

Ouyang S, Zhu W, Hamilton J*, et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research* **35**, D883-887.

Pareja-Tobes P, Manrique M, Pareja-Tobes E, Pareja E, Tobes R (2012) BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS ONE* **7**, e49239.

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* **7**, 84.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* **23**, 1061-1067.

Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**, 289-297.

Paterson AH, Freeling M, Sasaki T (2005) Grains of knowledge: genomics of model cereals. *Genome Research* **15**, 1643-1650.

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences USA* **98**, 9748-9753.

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

Porteres R (1970) *Primary Cradles of Agriculture in the African Continent.* Cambridge University Press, Cambridge.

Prasad V, Strömberg CAE, Leaché AD*, et al.* (2011) Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. *Nature Communications* **2**, 480.

Qin C, Yu C, Shen Y*, et al.* (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proceedings of the National Academy of Sciences USA* **111**, 5135-5140.

Quail MA, Smith M, Coupland P*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341-341.

Ram T, Majumder ND, Mishra B, Ansari MM, Padmavathi G (2007) Introgression of broad-spectrum blast resistance gene(s) into cultivated rice (*Oryza sativa* ssp. *indica*) from wild rice *O. rufipogon*. *Current Science* **92**, 225-230.

Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* **43**, 304-311.

Raubeson LA, Peery R, Chumley TW*, et al.* (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* **8**, 174.

Ren F, Lu BR, Li S, Huang J, Zhu Y (2003) A comparative study of genetic relationships among the AA-genome *Oryza* species using RAPD and SSR markers. *Theoretical and Applied Genetics* **108**, 113-120.

Ricegrowers' Association of Australia http://www.rga.org.au/.

Richards P (1996) *Redefining Nature: Ecology, Culture and Domestication.* Berg Publishers, Oxford.

Rockwood WG, Chandler RF (2001) *Rice research and production in the 21st century : symposium honoring Robert F. Chandler, Jr.* Makati, Metro Manila, Philippines: International Rice Research Institute.

Ronquist F, Teslenko M, van der Mark P*, et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**, 539-542.

Ross MG, Russ C, Costello M*, et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51.

Rothberg JM, Edwards M, Hoon J*, et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**, 5463-5467.

Scafaro AP, Yamori W, Carmo-Silva AE*, et al.* (2012) Rubisco activity is associated with photosynthetic thermotolerance in a wild rice (*Oryza meridionalis*). *Physiologia Plantarum* **146**, 99-109.

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**, W686-W689.

Schnable PS, Ware D, Fulton RS*, et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115.

Schwartz D, Li X, Hernandez L*, et al.* (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110-114.

Second G (1985) Geographic origins, genetic diversity and the molecular clock hypothesis in the Oryzeae. In: *Genetic differentiation and dispersal in plants* (ed. Jacquard P), pp. 41-56. Springer-Verlag, Heidelberg.

Sharma SD (2003) Species of the genus Oryza and their interrelationships. In: *Monograph on the genus Oryza.* (eds. Nanda JS, Sharma SD), pp. 73–111. Science Publishers, Enfield, New Hampshire.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145.

Shimamoto K, Kyozuka J (2002) Rice as a model for comparative genomics of plants. *Annual Review of Plant Biology* **53**, 399-419.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

Simon BK, Jacobs SWJ (1990) Gondwanan grasses in the Australian flora. *Austrobaileya* **3**, 239–260.

Simpson JT (2014) Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**, 1228-1235.

Simpson JT, Durbin R (2012) Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* **22**, 549-556.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15**, 121-132.

Sober E (1988) *Reconstructing the past: parsimony, evolution, and inference* MIT Press, Cambridge.

Song Z, Zhu W, Rong J*, et al.* (2006) Evidences of introgression from cultivated rice to *Oryza rufipogon* (Poaceae) populations based on SSR fingerprinting: implications for wild rice differentiation and conservation. *Evolutionary Ecology* **20**, 501-522.

Sotowa M, Ootsuka K, Kobayashi Y*, et al.* (2013) Molecular relationships between Australian annual wild rice, *Oryza meridionalis*, and two related perennial forms. *Rice* **6**, 26.

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215-225.

Stevens PF (2014) *Poaceae at the Angiosperm Phylogeny Website http://www.mobot.org/mobot/research/apweb/orders/poalesweb.htm#Poaceae.*

Straub SCK, Parks M, Weitemier K*, et al.* (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* **99**, 349-364.

Su Y, Liu Y, Li Z*, et al.* (2015) QTL analysis of head splitting resistance in cabbage (*Brassica oleracea* L. var. *capitata*) using SSR and InDel makers based on whole-genome re-sequencing. *PLoS ONE* **10**, e0138073.

Suh JP, Roh JH, Cho YC*, et al.* (2009) The *Pi40* gene for durable resistance to rice blast and molecular analysis of *Pi40*-advanced backcross breeding lines. *Phytopathology* **99**, 243-250.

Sun CQ, Wang XK, Li ZC, Yoshimura A, Iwata N (2001) Comparison of the genetic diversity of common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*O. sativa* L.) using RFLP markers. *Theoretical and Applied Genetics* **102**, 157-162.

Sundaramoorthi J, Babu C, Ram GS (2009) Molecular diversity in the primary and secondary gene pools of genus *Oryza*. *Plant Systematics and Evolution* **279**, 115-123.

Sweeney M, McCouch S (2007) The complex history of the domestication of rice. *Annals of Botany* **100**, 951-957.

Sweeney MT, Thomson MJ, Pfeil BE, McCouch S (2006) Caught red-handed: *Rc* encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* **18**, 283-294.

Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next‐generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21**, 2045-2050.

Tan L, Li X, Liu F*, et al.* (2008) Control of a key transition from prostrate to erect growth in rice domestication. *Nature Genetics* **40**, 1360-1364.

Tang J, Yu J, Yang H*, et al.* (2004) A comparison of rice chloroplast genomes. *Plant Physiology* **135**, 412-420.

Tang L, Zou XH, Achoundong G*, et al.* (2010) Phylogeny and biogeography of the rice tribe (Oryzeae): Evidence from combined analysis of 20 chloroplast fragments. *Molecular Phylogenetics and Evolution* **54**, 266-277.

Tang T, Lu J, Huang J*, et al.* (2006) Genomic variation in rice: genesis of highly polymorphic linkage blocks during domestication. *PLoS Genetics* **2**, e199.

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

Tian X, Zheng J, Hu S, Yu J (2006) The rice mitochondrial genomes and their variations. *Plant Physiology* **140**, 401-410.

Tikapunya T, Fox G, Furtado A, Henry RJ (2016) Grain physical characteristic of the Australian wild rices. *Plant Genetic Resources: Characterisation and Utilisation*, 1-12.

Tiwari VK, Faris JD, Friebe B, Gill BS (2016) Genome Mapping. In: *Encyclopedia of Food Grains*, pp. 365-375.

Tobkes N, Wallace BA, Bayley H (1985) Secondary structure and assembly mechanism of an oligomeric channel protein. *Biochemistry* **24**, 1915-1920.

Tsitrone A, Kirkpatrick M, Levin DA (2003) A model for chloroplast capture. *Evolution* **57**, 1776-1782.

Uozu S, Ikehashi H, Ohmido N*, et al.* (1997) Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. *Plant Molecular Biology* **35**, 791–799.

Van de Peer Y, Lemey Pe, Salemi Me (2009) Phylogenetic inference based on distance methods: theory (ed. Vandamme AM). Cambridge.

Vaughan DA (1989) The genus *Oryza* L.: Current Status of Taxonomy. *Research Paper Series 138. International Rice Research Institute, Manila, The Philippines*.

Vaughan DA, Kadowaki KI, Kaga A, Tomooka N (2005) On the Phylogeny and Biogeography of the Genus *Oryza*. *Breeding Science* **55**, 113-122.

Vaughan DA, Lu BR, Tomooka N (2008) The evolving story of rice evolution. *Plant Science* **174**, 394-408.

Venter JC, Adams MD, Myers EW*, et al.* (2001) The sequence of the human genome. *Science* **291**, 1304-1351.

Vicentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA (2008) The age of the grasses and clusters of origins of C4 photosynthesis. *Global Change Biology* **14**, 2963-2977.

Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ (2015) Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Scientific Reports* **5**, 13957.

Wambugu PW, Furtado A, Waters DLE, Nyamongo DO, Henry RJ (2013) Conservation and utilization of African *Oryza* genetic resources. *Rice* **6**, 1-13.

Wang L, Yu S, Tong C*, et al.* (2014a) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biology* **15**, 1-13.

Wang M, Yu Y, Haberer G*, et al.* (2014b) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nature Genetics* **46**, 982-988.

Wang Y, Fan G, Liu Y*, et al.* (2013) The sacred lotus genome provides insights into the evolution of flowering plants. *The Plant Journal* **76**, 557-567.

Ward BJ, van Oosterhout C (2016) hybridcheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources* **16**, 534-539.

Waters DL, Nock CJ, Ishikawa R, Rice N, Henry RJ (2012) Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. *Ecology and Evolution* **2**, 211-217.

Wei X, Qiao WH, Chen YT*, et al.* (2012) Domestication and geographic origin of *Oryza sativa* in China: insights from multilocus analysis of nucleotide variation of *O. sativa* and *O. rufipogon. Molecular Ecology* **21**, 5073-5087.

Whelan S, Liò P, Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17**, 262-272.

Wing R, Ammiraju JSS, Luo M*, et al.* (2005) The *Oryza* Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species. *Plant Molecular Biology* **59**, 53-62.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA* **84**, 9054-9058.

Wu W, Stupi BP, Litosh VA*, et al.* (2007) Termination of DNA synthesis by N(6)-alkylated, not 3′ -O-alkylated, photocleavable 2′ -deoxyadenosine triphosphates. *Nucleic Acids Research* **35**, 6339-6349.

Xie X, Jin F, Song M-H*, et al.* (2008) Fine mapping of a yield-enhancing QTL cluster associated with transgressive variation in an *Oryza sativa* × *O. rufipogon* cross. *Theoretical and Applied Genetics* **116**, 613-622.

Xu X, Liu X, Ge S*, et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* **30**, 105-111.

Xu Y, McCouch S, Zhang Q (2005) How Can We Use Genomics to Improve Cereals with Rice as a Reference Genome? *Plant Molecular Biology* **59**, 7-26.

Yamamoto T, Yonemaru J, Yano M (2009) Towards the understanding of complex traits in rice: substantially or superficially? *DNA Research* **16**, 141-154.

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303-314.

Ye CX, Hill C, Ruan J, Ma ZSS (2014) DBG2OLC: efficient assembly of large genomes using the compressed overlap graph. *arXiv preprint arXiv:1410.2801v3*.

Ye CX, Ma ZSS, Cannon CH, Pop M, Yu DW (2012) Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics* **13**, S1.

Yoshimura A, Nagayama H, Kurakazu T*, et al.* (2010) Introgression lines of rice (*Oryza sativa* L.) carrying a donor genome from the wild species, *O. glumaepatula* Steud. and *O. meridionalis* Ng. *Breeding Science* **60**, 597-603.

Yu J, Hu S, Wang J*, et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92.

Yu LX, Zheng P, Zhang T, Rodringuez J, Main D (2016) Genotyping-by-sequencing based genome-wide association studies on Verticillium wilt resistance in autotetraploid alfalfa (*Medicago sativa* L.). *Molecular Plant Pathology*.

Zao X, Wu T, Xie Y, Wu R (1989) Genome-specific repetitive sequences in the genus *Oryza*. *Theoretical and Applied Genetics* **78**, 201-209.

Zhang J, Kudrna D, Mu T*, et al.* (2016) Genome Puzzle Master (GPM) - An integrated pipeline for building and editing pseudomolecules from fragmented sequences. *Bioinformatics* **32**, 3058-3064.

Zhang Q, Wing (eds.) RA (2013) *Genetics and Genomics of Rice* Springer Science+Business Media, New York.

Zhang QJ, Zhu T, Xia EH*, et al.* (2014) Rapid diversification of five *Oryza* AA genomes associated with rice adaptation. *Proceedings of the National Academy of Sciences USA* **111**, E4954-E4962.

Zhao Z (1998) The Middle Yangtze region in China is one place where rice was domesticated: Phytolith evidence fron the Diaotonghuan Cave, Northern Jiangxi. *Antiquity* **72**, 885-897.

Zheng G, Patolsky F, Cui Y, Wang WU, Lieber CM (2005) Multiplexed electrical detection of cancer markers with nanowire sensor arrays. *Nature Biotechnology* **23**, 1294-1301.

Zhou S, Bechner M, Place M*, et al.* (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**, 278.

Zhu Q, Ge S (2005) Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytologist* **167**, 249-265.

Zou XH, Zhang FM, Zhang JG*, et al.* (2008) Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biology* **9**, R49.