THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Genetic diversity of Australian wild rice

Ali Imad Mohammad Moner

M.Sc. plant protection

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2018

Queensland Alliance for Agriculture and Food Innovation

Abstract

Rice (*Oryza sativa*) is the most important crop in the world. Two thirds of the world population consume rice as main part of their daily diet. Crop wild relatives are essential to provide new genetic resources in order to improve crops to meet food demand and cope with environmental changes. Domestication of rice led to loss of many important genes through application of strong selection for the traits favoured by humans. Australian wild rice has unique features and is found growing in areas isolated from domesticated rice. This avoids the risk of contamination by gene flow from domesticated rice into the wild rice populations as in Asia where wild rice is mixed with cultivated rice in the same areas. These populations retain the genetics of rice prior to domestication.

We took the advantage of next generation sequencing to study the Australian and Asian wild relatives of rice. We assembled high quality chloroplast sequences and used them to investigate the phylogeny of these populations, providing more details on the biogeography of the major groups of wild AA genome rices globally. Interestingly, the Australian chloroplast type was distinct from all others and was found to extend north to the Philippines. The groups of Asian wild relatives had substantially overlapping distributions across the area studied. This suggested a complex evolutionary history of the rice progenitors leading to the domestication of rice. Genome sequencing has suggested that the wild rice populations in northern Australia may include novel taxa, Analysis of the chloroplast and nuclear data demonstrated very clear evidence of distinctness from other AA genome *Oryza* species with significant divergence between Australian populations. Phylogenetic analysis suggested the Australian populations represent the earliest-branching AA genome lineages and may be critical resources for global rice food security. Populations of apparent hybrids between the taxa were also identified suggesting ongoing dynamic evolution of wild rice in Australia. These introgressions model events similar to those likely to have been involved in the domestication of rice.

Starch quality and quantity are crucial for rice consumers and the rice industry. Starch properties have been linked directly to impact on human health. Many genes have been involved in determining rice starch properties. The genetic relationship of the starch related genes: *ISA2, ISA3, PUL, SBE1, SBE3, SBE4, SSI, SSII-1, SSII-2, SSII-3, SSIII, SSIV and GBSSI* in the Australian wild rice populations of Cape York were studied. Many SNPs/FNPs were recorded in the UTRs and exonic regions of these genes that could possibly impact on their expression and function. CDS prediction of the *GBSSI* gene showed an extra 120bp in some populations. This was due to a change in the predicted splicing site that would lead to intron retention and add 40 amino acid to the predicted protein. It seems that this addition would not affect the active site, however this may explain the differences in starch properties of this taxa reported previously. Australian wild rice populations have potential as a novel source of starch related genes which may help to improve the health of rice consumers.

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, financial support 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 higher degree by research 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 and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications during candidature

Moner A.M.M., Agnelo Furtado, Ian Chivers, Glen Fox, Darren Crayn, Henry RJ. 2018. Diversity and Evolution of Rice Progenitors in Australia Ecology and Evolution accepted

Mondal TK, Henry RJ. 2018. The Wild *Oryza* Genomes. In: Springer. Book Chapter 16 *Oryza meridionalis* N.Q.Ng Ali Mohammad Moner and Robert J. Henry p177-182.

Conference presentation

(Underline denotes oral presentation, " * " denotes poster presentation)

A. Moner, T Tikapunya, H Badro, M Brozynska, A Furtado, H Smyth,QQ Liu1, R G Gilbert and R J Henry 2017. Australian Wild Rice: Diverse and Tasty. TropAg conference Brisbane Australia 2017

A. Moner*, Agnelo Furtado and R.J. Henry. Phylogenetic analysis of the Asian and Australian AA genome wild rice 2017 Plant Genome Evolution conference Barcelona Spain

A. Moner*, Agnelo Furtado and R.J. Henry. Rice Genetic Resources of Cape York. TropAg conference Brisbane Australia 2015

Publications included in this thesis

Moner A.M.M., Agnelo Furtado, Ian Chivers, Glen Fox, Darren Crayn, Henry RJ. 2018. Diversity and Evolution of Rice Progenitors in Australia Ecology and Evolution accepted

Incorporated as Chapter 4.

Contributor	Statement of contribution
Moner A.M.M. (Candidate)	Conception and design (60%) Analysis and interpretation (70%) Drafting and production (70%)
Agnelo Furtado	Conception and design (10%) Analysis and interpretation (10 %) Drafting and production (5%)
Ian Chivers	Conception and design (0%) Analysis and interpretation (0%) Drafting and production (5%)
Glen Fox	Conception and design (0%) Analysis and interpretation (0%) Drafting and production (5%)
Darren Crayn	Conception and design (0%) Analysis and interpretation (0%) Drafting and production (5%)
Henry R.J.	Conception and design (30%) Analysis and interpretation (20%) Drafting and production (10%)

Incorporated as Appendix 1

Contributor	Statement of contribution
Moner A.M.M. (Candidate)	Conception and design (60%) Analysis and interpretation (80%) Drafting and production (80%)
Henry R.J.	Conception and design (40%) Analysis and interpretation (20%) Drafting and production (20%)

Contributions by others to the thesis

Principal advisor, Prof. Robert J. Henry contributed in the conception and design of this project. He edited and critically revised all sections.

Associated advisor, Agnelo Furtado contributed in the conception and design of this project. He edited and critically revised all sections. He also designed and drafted the appendix 3

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

None

Research Involving Human or Animal Subjects

No animal or human participants were involved in this research.

Acknowledgements

First and foremost I would like to appreciate my supervisors Prof. Robert J. Henry and Dr. Agnelo Furtado for their generous guidance and advice. I am thankful to all my colleagues in Robert's research group especially Annie Morley, Marta Brozynska and Nam Van Hoang. Also I sincerely thank the HCED Iraq program for providing a PhD scholarship for me. I would like to say thanks to thesis committee panel Andrew Geering and Ian Godwin who accompanied me during my candidature

I thank John Thurlow and Alan Lambert (Mareeba Shire Council), Kerry Walsh (Department of Environment and Heritage Protection), Annabelle Olsson and Tim Nevard (Mareeba Tropical Savanna and Wetland Reserve Wildlife Conservancy of Tropical Queensland), Ernest Madua (Traditional Owner and Ranger Napranum), Les Harrigan and Conrad Yeatman (Land Trust Traditional Owners of Rinyirru), Gil Hainey (Northern Peninsula Area Regional Council), Louise Stone (Mapoon Aboriginal Shire Council), Uncle Shorty (Traditional Owner, Far North Cape Area), Peta Standley (Cape York Natural Resource Management), Paul Ryan (Paul Ryan Global), Warren Strevens (Northern Peninsula Area Regional Council / Injinoo Ranger Base), Janie White (Shire of Cook QPWS Office), Tony Cockburn and Matt Wallace (Rinyirru National Park), Ray Byers, Lisa Hamilton, Tiparat Tikapunya, Xavier Tierney, Andrew Geering for assistance with sample and data collection.

Special thanks to my family for their patient and endless support. Finally, it is honor to dedicate this work to my parents and my wife, Alaa and my lovely children, Abduladheem and Lmar

Financial support

HCED Iraq program provided a full PhD scholarship.

Keywords

Asian wild rice, *Oryza*, chloroplast sequence, rice evolution, phylogeny, Australian wild rice, GBSSI, starch genes

Australian and New Zealand Standard Research Classifications (ANZSRC)

Provide data that links your thesis to the disciplines and discipline clusters in the Federal Government's Excellence in Research for Australia (ERA) initiative.

Please allocate the thesis a **maximum of 3** Australian and New Zealand Standard Research Classifications (ANZSRC) codes at the **6 digit level** and include the descriptor and a percent weighting for each code. Total percent must add to 100.

ANZSRC code: 060408, Genomics, 40%

ANZSRC code: 060309, Phylogeny and Comparative Analysis,30%

ANZSRC code: 070305, Crop and Pasture Improvement (Selection and Breeding),30%

Fields of Research (FoR) Classification

Allows for categorisation of the thesis according to the field of research.

Please allocate the thesis a **maximum of 3** Fields of Research (FoR) Codes at the **4 digit level** and include the descriptor and a percent weighting for each code. Total percent must add to 100.

FoR code: 0604, Genetics, 40%

FoR code: 0603, Evolutionary Biology, 30%

FoR code: 0703, Crop and Pasture Production, 30%

Table of Contents

1 Preface and Study objective 1

1.1 Rice importance and challenges 1

1.2 Rice genomics 1

1.3 Australian wild rice 2

1.4 Advanced technologies 2

1.5 Aim and Objectives of the project 2

1.6 Research plan 3

2 Literature review 5

2.1 Genetic diversity and the environmental impact 5

2.2 Diversity in genus *Oryza* spp 5

2.3 From wild to domesticated evolutionary background 6

2.3.1 Traits influenced by domestication 9

2.3.1.2 Shattering genes 9

2.3.1.3 Seed colour genes 10

2.4 Valuable Characteristics of Wild rice 12

2.4.1 Disease and pest resistance 12

2.4.2 Abiotic stress resistance 13

2.4.3 Productivity 13

2.4.4 Health and nutrition importance 14

2.5 Wild *Oryza* species in Australia 14

2.5.1 *Oryza* australiensis 14

2.5.2 *Oryza* officinalis 15

***2.5.3 Oryza rufipogon* 15**

***2.5.4 Oryza meridionalis* 15**

***2.5.5 Oryza nivara S.D.* and *O. minuta* 16**

2.5.6 *Oryza spp.* Taxon A and Taxon B in North Queensland *Oryza spp.* Taxon A and Taxon B in North Queensland 16

2.6 Genetic diversity analysis 17

2.6.1 Molecular genotyping tools: 17

2.7 NGS application in rice genetic diversity analysis 18

2.7.1 Specific gene sequences 18

2.7.2 Chloroplast DNA sequencing 18

2.7.3 Whole genome sequencing 19

2.8 Starch related genes 22

3 Chloroplast phylogeography of AA genome rice species 25

3.1 Abstract 25

3.2 Introduction 25

3.3 Materials and methods 27

3.3.1 Chloroplast genome assembly 27

3.3.2 Phylogenetic analysis 28

3.3.3 Genome annotation 28

3.4 Results 29

3.4.1 Raw data 29

3.4.2 Chloroplast assembly 32

3.4.3 Chloroplast alignment 32

3.4.4 Phylogenetic analysis 38

3.4.5 SNPs and FNPs variation 38

3.5 4. Discussion 40

4 Diversity and Evolution of Rice Progenitors in Australia 52

4.1 Abstract 52

4.2 Introduction 52

4.3 Material and methods 53

4.3.1 Field collections 53

4.3.2 Morphological measurement 53

4.3.3 DNA extraction and sequencing 53

4.3.4 Chloroplast genome assembly ... 54

4.3.5 Chloroplast phylogenetic analysis ... 55

4.3.6 Chloroplast genome annotation ... 55

4.3.7 Phylogenetic analysis of nuclear genes ... 55

4.4 Results and Discussion ... 56

5 Starch gene diversity in Australian wild rice ... 61

5.1 Abstract ... 61

5.2 Introduction ... 62

5.3 Materials and methods ... 64

5.3.1 Australian wild rice collection ... 64

5.3.2 Starch related Gene’s sequence ... 64

5.3.3 Phylogenetic analysis ... 65

5.3.4 CDS prediction ... 65

5.3.5 Protein model ... 65

5.3.6 Protein alignment and 3D structure ... 65

5.4 Results ... 67

5.5 Discussion ... 69

6 General discussion ... 78

6.1 Fulfilment of objectives ... 78

6.2 Chloroplast genomes of Asian wild rice ... 78

6.3 Phylogeny of Australian wild rice populations ... 80

6.4 Starch related genes in wild rice populations ... 81

6.5 Future directions ... 82

7 References ... 84

1 Appendix 1. *Oryza meridionalis* ... 95

1.1 Economic/Academic importance ... 95

1.2 Brief botanical descriptions including distribution ... 95

1.3 Cytological details of genome including karyotype data ... 96

1.4 Physiological studies 96

1.5 Enumeration of sequences 96

1.6 Assembly 96

1.7 Repetitive sequences 96

1.8 Gene annotation 97

1.9 Organelle genome 97

1.10 Impact on plant breeding including pre-breeding work 97

1.11 Comparative genomics 98

1.12 Future prospects 98

2 Appendix 2. 100

3 Appendix 3. Chloroplast Assembly Pipeline 139

3.1 Abbreviations 139

3.2 Sequence data statistics 140

3.3 Abbreviations and denotations used to identify assembled Cp sequences 140

3.4 Chloroplast Assembly Pipeline (CAP details) 141

3.5 Mapping Optimisation (MOpt) process and the CM-Rule: rationale 141

3.5.1 Mapping Improvement process: rationale 143

3.6 *De novo*-assembly process and *de novo*-improvement process: rationale 144

3.7 Manual curation to obtain a Cp-CAP 145

3.7.1 Important steps in the CAP 146

3.8 M-Component robustness- OsNipp35bp-PEreads and CpW as a reference 146

4 Appendix 4. 167

5 Appendix 5 192

List of tables

Table 1. *Oryza* species, genome group, chromosome number and the geographical origin (Joseph et al., 2008; Koh et al., 2015) and http://www.gramene.org/) 8

Table 2. Genes related to the domestication process. A functional nucleotide polymorphism in a specific region leads to changes in traits. 11

Table 3. Abiotic resistance genes in *O. sativa* 13

Table 4. SNPs and their impact on related agronomical traits (Huang et al., 2010)........ 20

Table 5. Starch analysis genes SNPs and InDels in Swarna cultivar modified (Rathinasabapathi et al., 2015).. 23

Table 6 Geographic origin of wild rice *O. rufipogon* accessions. The location (latitude and longitude) of collection, ecotype, sequence coverage (whole genome basis) and total number of sequence reads are provided for each accession............................... 30

Table 7 Chloroplast sequence analysis by mapping, the chloroplast coverage and the number of gaps following mapping is given for each accession. 33

Table 8 Variants among AA chloroplast genomes. Deletion, Insertions SNPs when compared with O. sativa subsp. *japonica* Nipponbare GU592207.1 Del: deletion, Del.T.R. : deletion tandem repeat, Ins.: insertion, Ins.T.R. : insertion tandem repeat, SNP Tr.: SNP tran transition, SNP Trv. :SNP transversion and Subs. : substitution. 35

Table 9 Polymorphisms between the clades defined by the chloroplast phylogeny SNPs, InDel and Deletions between the clades as defined in Figure 4.................................... 42

Table 10 Functional variation in chloroplast genome sequences. FNPs location, amino acid substitution, codon changed and polymorphism type. 42

Table 11 Summary of variants identified for all Asian wild rice samples analysed........ 49

Table 12 Details of thirteen starch related genes in rice reference gene name, ID and size are shown.. 65

Table 13. SNPs/FNP summary for starch related genes compared to the *O. sativa japonica* assembly Build 4.0... 70

Table 14 Description of *Oryza meridionalis* (Groves et al., 2009).. 95

Table 15 Hybrid and PacBio assembly statistics calculated for scaffolds and contigs for hybrid assembly and for scaffolds only for PacBio assembly (Brozynska et al., 2016). 97

TABLE 16 *ORYZA* SPECIES THE GENOME GROUP, CHROMOSOME NUMBER AND THE GEOGRAPHICAL ORIGIN IS PROVIDED FOR EACH SPECIES (JOSEPH & THOMAS, 2008) AND (KOH & THOMSON, 2015). 100

TABLE 17 CHLOROPLASTS SEQUENCES OF *ORYZA* SPP. (HTTP://WWW.NCBI.NLM.NIH.GOV/GENOME). REFSEQ, SIZE, GENES NUMBER AND RELEASED DATE WERE DEMONSTRATED. LAST UPDATE 15.1.2018 102

TABLE 18 COMPARISON OF CHLOROPLAST SEQUENCE GENERATED BY MAPPING AND DE NOVO PROCEDURES. TWO DIFFERENT REFERENCE GENOMES WERE USED. THE DEGREE OF MANUAL CORRECTION REQUIRED FOR ASSEMBLY AND THE FINAL CHLOROPLAST SIZE IS GIVEN. 104

TABLE 19 PHYLOGENETIC SOFTWARE TOOLS APPLIED TO CHLOROPLAST GENOME ANALYSIS, ANALYSIS MODEL AND BOOTSTRAP NUMBER USED IN THIS STUDY. 107

TABLE 20 SNP FREQUENCIES IN EACH CLADE AS DESCRIBED BELOW 108

TABLE 21 SNPS / INDELS MARKERS DISTINGUISHING THE CLADES DEFINED IN THE CHLOROPLAST PHYLOGENY. 109

TABLE 22 SUMMARY STATISTICS OF THE NEXT GENERATION SEQUENCE DATA USED FOR THE ASSEMBLY OF CHLOROPLAST GENOME SEQUENCE. 147

TABLE 23 READ-MAPPING PARAMETERS AND THEIR SETTINGS DETAILS USED IN THE MAPPING-OPTIMISATION PROCESS AT THE READ MAPPING STEP (MOPT:R) AND USING THE 35BP PARED-END ILLUMINA READS OF *ORYZA* SATIVA CV. NIPPONBARE. 148

TABLE 24 DETAILS OF MISMATCHES BETWEEN DIFFERENT CPN/MAPPING CONSENSUS SEQUENCES DERIVED WHEN O. SATIVA NIPPONBARE 35BP ILLUMINA READS WERE MAPPED TO THE O.SATIVA CHLOROPLAST SEQUENCE (GU592207) UNDER VARIOUS MAPPING SETTINGS. 149

TABLE 25 MANUAL CURATION OF MISMATCHES BETWEEN CHLOROPLAST (CP) SEQUENCES. 151

TABLE 26 MANUAL CURATION OF CHLOROPLAST (CP) SEQUENCES DERIVED FROM THE MAPPING ASSEMBLY COMPONENT (M-COMPONENT) AND TWO SEQUENCES FROM THE DENOVO-ASSEMBLY COMPONENT (D-COMPONENT); THE DENOVO ASSEMBLY PROCESS (D-PROCESS) AND FROM THE DENOVO IMPROVEMENT PROCESS (DIMP2-PROCESS). ALL CP SEQUENCES WERE GENERATED USING 35 BP PAIRED END ILLUMINA READS OF *O. SATIVA* NIPPONBARE (ACCESSION GU592207) AND THE CP (KF428978) OF THE AUSTRALIAN WILD RICE TAXA-A (CPW) WAS USED AS A REFERENCE CP SEQUENCE FOR THE M-PROCESS. ALL ANALYSIS STEPS WERE UNDERTAKE USING CLC GENOMICS WORKBENCH. 152

TABLE 27 DETAILS OF COLLECTIONS OF WILD RICE FROM NORTH QUEENSLAND MADE IN 2015, 2016 AND 2017. INCLUDING SITE DESCRIPTION, GPS COORDINATES, PANICLE SHAPE, AWN AND ANTHER LENGTH FOR WILD POPULATIONS FROM EACH COLLECTION SITE. 167

Table 28 Details of sequence coverage of Australian wild rice samples. Including whole genome coverage with total number of reads, and minimum, maximum and mean coverage of the chloroplast genome. 172

Table 29 Variants in chloroplast genomes insertions, deletions and SNPs compared with the O. sativa subsp. *japonica* Nipponbare GU592207.1 reference genome. Abbreviations are as follows: Del: deletion, Del.T.R.: deletion tandem repeat, Ins.: insertion, Ins.T.R.: insertion tandem repeat, SNP Tr.: SNP transition, SNP Trv.:SNP transversion and Subs.: substitution. 173

Table 30 Chloroplast functional nucleotide polymorphisms (FNPs) in Australian wild rice populations. including position, gene name, gene product, amino acid substitution and codon change. 174

Table 31 Comparison of the SNPs, FNPs and the unique FNPs in Australian wild rice populations. 176

Table 32 Phylogenetic analysis tools applied to chloroplast genome analysis. 177

Table 33 Unique chloroplast SNPs found in the Australian taxa. 178

Table 34 Chromosomes phylogenetic analysis topology agreement. 179

Table 35 Non synonyms nucleotide polymorphism in 13 starch related gene. Gene, protein and amino acid substitutions are shown. Colours: Green taxon A, orange taxon B yellow in both 192

List of figures

FIGURE 1. THE PICTURE SHOWS THE DIFFERENCE AMONG 12 *ORYZA* SPECIES AT THE SAME DEVELOPMENT STAGE (SANCHEZ ET AL., 2013) 7

FIGURE 2. THE RELATIONSHIP BETWEEN *ORYZA* SPECIES AA GENOMES BASED ON CHLOROPLAST DNA ANALYSIS (WAMBUGU ET AL., 2015) 7

FIGURE 3 GRAIN APPEARANCE OF THE AUSTRALIAN WILD RICES (TIKAPUNYA ET AL., 2017) 16

FIGURE 4. DISTRIBUTION OF 79 ASIAN WILD RICE ACCESSIONS. THE ACCESSIONS WERE DIVIDED INTO THOSE FROM 5 DIFFERENT GEOGRAPHIC ZONES FOR COMPARISON. MAP SOURCED FROM GOOGLE MAPS. 28

FIGURE 5 PHYLOGENETIC RELATIONSHIP OF *ORYZA* CHLOROPLAST AA GENOME. ANALYSIS USING MRBAYES GTR MODEL WITH 2000 BOOTSTRAPS AND O. AUSTRALIENSIS AS AN OUT GROUP. NUMBERS ON BRANCHES REFER TO PROBABILITY PERCENTAGE. 39

FIGURE 6 CHLOROPLAST GENE MAP. POLYMORPHIC GENES ARE MARKED WITH *. THE INNER CIRCLE REPRESENTS THE FOUR CHLOROPLAST REGIONS LSC, IRB, SSC AND IRA. THE GC CONTENT IS SHOWN IN THE GREY AREA 40

FIGURE 7 PHYLOGEOGRAPHIC DISTRIBUTION OF DIVERSITY IN *ORYZA* SPP. AA CHLOROPLAST GENOMES. THE *O. SATIVA* SPP. *INDIC*A AND *O. NIVARA* CLADE GROUP ARE REPRESENTED BY BLUE AND GREEN DOTS RESPECTIVELY. THE YELLOW DOTS REPRESENT THE CLADE RELATED TO *O. SATIVA SUBSP.* JAPONICA. THE AUSTRALIAN CLADE IS MARKED WITH RED DOTS. THE BLACK DOT REPRESENTS W1977 WHICH WAS AN OUT GROUP RELATIVE TO THE TWO SUB CLADES INCLUDING *O. NIVARA* AND *O. SATIVA* SUBSP. *INDICA*. *ASIAN AND AUSTRALIAN ACCESSION POSITIONS WERE BASED ON COLLECTION SITE GPS LOCATIONS. MAP SOURCED FROM GOOGLE MAPS. 41

FIGURE 8 AUSTRALIAN WILD RICE COLLECTION SITES. RED DOTS INDICATE COLLECTION SITES. 54

FIGURE 9 DIVERSITY OF CHLOROPLAST GENOMES A, PHYLOGENETIC TREE BASED ON MP ANALYSIS OF WHOLE CHLOROPLAST GENOME SEQUENCES COLOURS RELATE TO THE MAIN CLADES. RED AND BROWN CLADES ARE FROM AUSTRALIA. BOOTSTRAP VALUES (MP 1000 REPLICATES) ARE SHOWN ON THE BRANCHES; B, GENETIC DISTANCES BETWEEN POPULATIONS IN AUSTRALIA AND ELSEWHERE 58

FIGURE 10 INDIVIDUAL CHROMOSOME ANALYSIS SHOWING DIVERSITY OF NUCLEAR GENOMES A, PHYLOGENETIC TREE BASED ON MP ANALYSIS OF THE CONCATENATED ALIGNMENT OF ALL NUCLEAR GENES. COLOURS RELATE TO THE MAIN CLADES. RED AND BROWN CLADES ARE FROM AUSTRALIA. BOOTSTRAP VALUES (MAXIMUM PARSIMONY, 1000 REPLICATES) ARE SHOWN ON THE BRANCHES; B, INDIVIDUAL CHROMOSOME LENGTH AND NUMBER OF GENES PER CHROMOSOME. 60

Figure 11 A. Gene structure of 13 starch related genes. Green bars are complete gene sequences, yellow bars are exons. B. Australian wild rice collection sites North of Queensland 67

Figure 12 Phylogenetic tree based on maximum likelihood and Bayesian analysis (both agreed in topology) of 13 starch gene sequences. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids 68

Figure 13. Two type of reads as evidence of hybridisation in Australian wild rice population from North Queensland. 69

Figure 14. Prediction of the CDS and determined exons boundary in GBSSI taxa B compared to the reference O. sativa japonica. Differences are highlighted by the red rectangle. In Taxa B exon 11 and 12 were combined and included the intron between them. 71

Figure 15 A. SNP in the intron 11 splicing enhancer of the GBSSI gene B. intron 11 retention and 120 bp insertion in the CDS C. insertion of extra 40 amino acid as a consequence of the intron retention. 73

Figure 16. JSmol display of the GBSSI 3D structure alignment (superposition) Taxa B with Reference O.sativa *japonica* (3vue.1.A) using FATCAT. Taxa B and reference are in grey and red respectively. White arrow indicates the difference in the structure between these genes. left top is Taxa B left bottom reference 74

Figure 17. 3D structure filled of the GBSSI gene protein Taxa B blue, green and red colour referring to the pocket KTGGL, 40 amino acid insertion and one amino acid change Ser to Arg 76

Figure 18. Three dimension structure of the GBSSI gene of Taxa B. The closest distance between the Thr in the active site and the 40 amino acid insertion Phe was 15 A° 76

Figure 19 A, GBSSI gene of Taxa B; B, O. sativa japonica. Disulfide bond shown by white arrows 77

Figure 20 Distrubution of O. meridionalis http://www.ala.org.au 99

Figure 21 *Oryza* meridionalis in northern Australia. 99

Figure 22 FNPs percentages found in different genes in Asian wild rice chloroplast genomes 137

Figure 23 Geographical distribution of Asian wild rice zone 1 including India and zone 2 India and Burma. High coverage samples was selected from each circle. 137

Figure 24 Geographical distribution of Asian wild rice zone 3 including China and zone 4 including Thailand, Vietnam and Cambodia. High coverage samples were selected from each circle 138

Figure 25 Geographical distribution of Asian wild rice zone 5 including Oceania Australia, Papua New Guinea, Indonesia, Malaysia and Singapore. High coverage samples were selected from each circle. 138

Figure 26 Details of the Chloroplast Assembly Pipeline (CAP) R; read mapping tool, P; extract paired end mapped reads and remapping, S; structural variant plus realignment tools. **i,** The CAP pipeline consists of two distinct components the M-Component and the D-Component. **ii,** All mappings steps included Cost (C) settings of C1 and C2 comprising of 2, 3, 3 and 1, 2, 2 for Mismatch Cost, Insertion Cost and Deletion Cost respectively. Each of the Cost Settings had a combination of five Fraction (F) settings of 1.0, 1.0 and 1.0, 0.95 and 1.0, 0.8 and 0.8, 0.8 and 0.8, 0.5 for Length Fraction and Similarity Fraction respectively. **iii**, All *de novo* assembly steps were undertaken using the “Fast” (F) mode and at various “Word” (W) and “Bubble” (B) settings. The DImp-process involved subjecting the CpX-D to the 3-Map-Tool of R+P+S-Tool. X, code name of the genotype whose Cp is being generated. All analysis are designed to be undertaken in the CLC Genomics Workbench (CLCBio, Qiagen, Denmark). 159

Figure 27 Steps of the Mapping Optimisation process and the “Cost” and “Fraction” mapping settings used to obtain an accurate mapping-derived chloroplast assembled sequence. N/, abbreviation of the reference Cp genome sequence used a suffix to denote the reference Cp sequence used; C, Cost mapping setting used; F, Fraction mapping settings used; MOpt, Mapping Optimisation process; R, read mapping tool; P, extracting of mapped paired-end reads and remapping tool: S, structural variant analysis plus local realignment tool. Mapping assembled Chloroplast sequences were progressively passed through various MOpt steps. Increasing C- and F-values represents decreasing stringency in mapping of sequence reads to the Cp reference used. All analyses were conducted using the CLC Genomics Workbench analysis software. 160

Figure 28 Mapping Optimisation (MOpt) process-derived chloroplast genome (Cp) sequences using CpW as a reference and mismatches when compared to CpN. CpW, Cp sequence of the Australian Wild rice Taxon-A (Genbank accession KF428978); CpN, Cp sequence of *Oryza sativa* Nipponbare (CpN, Genbank accession GU592207); LF, SF, length and similarity fraction; MC, IC, DC, mismatch, insertion and deletion cost. **i,**

ii, iii, iv, data related to the MOpt process-derived Cp sequences derived using a fixed setting for MC, IC and DC of 2, 3, 3 and of 1, 2, 2 respectively and within these six combinations of LF and SF Fraction settings. Y-axis indicates mismatches in the MOpt process-derived Cp genome sequences when compared to the CpN (**i, ii**) and when compared to CpW (**iii, iv**). Read mapping was carried out using 35 bp paired-end Illumina reads of *O. sativa* Nipponbare (Genbank accession GU592207) and using the publically available CpW. MOpt process involves the read mapping tool (R), extracting the mapped paired-end reads and remapping tool (P) and the structural variant plus local realignment tool (S), implemented is sequence as the R+P+S or R+S+P with the aim of reducing the mismatches in the Cp sequences obtained from the preceding step. Number of mismatches when compared to CpN, shown above each bar, is a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions. Consistent number of mismatches in consensus Cp sequences derived from the R+P+S step and the R+S+P step, at each of the C and F setting used, are highlighted in blue. All mapping analysis was carried out using CLC Genomics Workbench v7.5.1. 161

Figure 29 Mapping Optimisation (MOpt) process-derived chloroplast genome (Cp) sequences using CpW as a reference and mismatches when compared to CpN. CpW, Cp sequence of the Australian Wild rice Taxon-A (Genbank accession KF428978); CpN, Cp sequence of *Oryza sativa* Nipponbare (CpN, Genbank accession GU592207); LF, SF, length and similarity fraction; MC, IC, DC, mismatch, insertion and deletion cost. i, ii, iii, iv, data related to the MOpt process-derived Cp sequences derived using a fixed setting for MC, IC and DC of 2, 3, 3 and of 1, 2, 2 respectively and within these six combinations of LF and SF Fraction settings. Y-axis indicates mismatches in the MOpt process-derived Cp genome sequences when compared to the CpN (i, ii) and when compared to CpW (iii, iv). Read mapping was carried out using 35 bp paired-end Illumina reads of *O. sativa* Nipponbare (Genbank accession GU592207) and using the publically available CpW. MOpt process involves the read mapping tool (R), extracting the mapped paired-end reads and remapping tool (P) and the structural variant plus local realignment tool (S), implemented is sequence as the R+P+S or R+S+P with the aim of reducing the mismatches in the Cp sequences obtained from the preceding step. Number of mismatches when compared to CpN, shown above each bar, is a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions. Consistent number of mismatches in consensus Cp sequences derived from the R+P+S step and the R+S+P step, at each of the C and F setting used, are highlighted in blue. All mapping analysis was carried out using CLC Genomics Workbench v7.5.1........163

Figure 30 Mapping Improvement (MImp) and De novo Improvement (DImp) process reduces mismatches in the Cp sequence from MOpt-process and CpN-D sequence respectively. Cp; chloroplast sequence; MOpt, Mapping optimisation process; CpN-D, *de novo* assembly-derived Cp sequence; MC, IC and DC, mismatch, insertion and deletion cost (C) setting; LF and SF, length and similarity fraction (F) setting ; R, read mapping tool; P, extracting mapped paired-end reads and remapping tool; S, structural variant analysis and local realignment tool. Read mapping and *de novo* assembly was carried out using 35bp Illumina Paired end reads of *Oryza sativa* cv Nipponbare. The Cp sequence of the Australian Wild rice Taxon-A (CpW, KF428978) was used as a reference for read mapping assembly at C1F3 mapping settings representing a C setting of 2, 3, 3 for MC, IC and DC respectively, and a F setting of 1.0 and 0.8 for LF and SF respectively. The X-axis indicates the various Cp sequences and the mapping settings used. Mismatches in mapping-derived Chloroplast (Cp) sequences when compared to the publically available Cp sequence of *O. sativa* Nipponbare (CpN, GU592207) (**i, iii**) and to the Australian Wild rice Taxa-A (CpW, KF428978) (**ii**). Number of mismatches are a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions and are shown at top of each bar while those in blue highlight represent mismatches and bases covered. **i**; The Cp sequence from the MOpt process at the C2F5 had the highest mismatches of 42 over 78 bases, while at the C1F3 setting had 5 over 24 bases to 3 mismatches over 5 bases when using the 3-Map tools and further reduced at the MImp process to 2 over 5 bases. **ii**; All of the consensus Cp sequences discussed above had 124 mismatches when compared to CpW. iii, The *de novo* assembly derived CpN-D sequence was also improved when passed through the DImp process with 17 mismatches over 96 bases reduced to 4 over 86 bases. 165

Figure 31 Comparisons of Chloroplast (Cp) sequences at a T nucleotide homopolymer sequence between CpN-CAP sequence of *Oryza* sativa cv Nipponbare to CpN and CpW respectively. CpN-CAP, chloroplast sequence assembled using the Chloroplast assembly pipeline (CAP) using paired end Illumina reads (35 bp) of *Oryza sativa*; CpW, chloroplast sequences of Australian Wild rice Taxa-A (KF428978); CpN, Chloroplast sequence of *O. sativa* (GU592207). CpW has 16 T and one of the T is replaced with a G nucleotide while CpN has 17 T nucleotide (i). The assembled CpN-CAP has 16 T nucleotides (B) and its alignment to CpW is shown in (ii) and to CpN is shown in (iii). The mapping of the paired end Illumina reads (35 bp) to CpW is shown in

(iv) and to CpN-CAP is shown in (v). All alignments were undertaken Geneious V 9 and all mapping of reads to Cp sequences using CLC genomics Work Bench V 9.0...........166

Figure 32 Wild rice habitat in northern Queensland Jpn2 site S:15.43943° E:144.21111° 180

Figure 33 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 1 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...180

Figure 34 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 2 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...181

Figure 35 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 3 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...182

Figure 36 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 4 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...183

Figure 37 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 5 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...184

Figure 38 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 6 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...185

Figure 39 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 7 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...186

Figure 40 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 8 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...187

Figure 41 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 9 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches...188

Figure 42 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 10 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches. 189

Figure 43 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 11 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches. 190

Figure 44 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 12 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches. 191

Figure 45 Phylogenetic tree based on Bayesian analysis of GBSSI gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids 196

Figure 46 Phylogenetic tree based on Bayesian analysis of ISA2 gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids. 197

Figure 47 Phylogenetic tree based on Bayesian analysis of ISA3 gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids. 198

Figure 48 Phylogenetic tree based on Bayesian analysis of *PUL* gene. Bootstrap values (1000 replicates) are shown on the branches........ 199

Figure 49 Phylogenetic tree based on Bayesian analysis of *SBE1* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids. 200

Figure 50 Phylogenetic tree based on Bayesian analysis of *SBE3* gene. Bootstrap values (1000 replicates) are shown on the branches........ 201

Figure 51 Phylogenetic tree based on Bayesian analysis of *SBE4* gene. Bootstrap values (1000 replicates) are shown on the branches........ 202

Figure 52 Phylogenetic tree based on Bayesian analysis of *SSI* gene. Bootstrap values (1000 replicates) are shown on the branches........ 203

Figure 53 Phylogenetic tree based on Bayesian analysis of *SSII-1* gene. Bootstrap values (1000 replicates) are shown on the branches...204

Figure 54 Phylogenetic tree based on Bayesian analysis of *SSII-2* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids. 205

Figure 55 Phylogenetic tree based on Bayesian analysis of *SSII-3* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids. 206

Figure 56 Phylogenetic tree based on Bayesian analysis of *SSIII* gene. Bootstrap values (1000 replicates) are shown on the branches...207

Figure 57 Phylogenetic tree based on Bayesian analysis of *SSIV* gene. Bootstrap values (1000 replicates) are shown on the branches...208

Chapter 1

1 Preface and Study objective

1.1 Rice importance and challenges

There is no doubt that rice (*Oryza sativa* L.) is one of the most essential crops in the world. It is planted in one and half billion hectares in over 100 countries and accounts for approximately 30 % of global cereal production. By 2025, rice production will need to meet the demand of 4.6 billion people who rely mainly on rice. Moreover, it is a key source of carbohydrates (calories) as well as a source of many other essential nutrients (minerals and amino acids) in the human diet (Gnanamanickam, 2009). To meet this need with current production efficiencies, the area which is currently cultivated for rice would need to be doubled over the next few decades. However, it is unlikely that such an expansion in the area of land cultivated would be possible, as land resources are very limited, especially in relation to soil suitability and water availability.

Moreover, environmental stresses (biotic and abiotic stresses), including those associated with climate change and global warming, are reducing the available area that is suitable for rice growing. According to the FAO Rice Market Monitor Report of October 2015, world production was then around 740 million tons, which is less than that predicted previously by 6.5 million tons. This productivity is 0.4% (2.6 million tons) less than that of 2014 (which was also less than predicted), indicating that there has been negative growth in rice production for those two years (FAO, 2015). As a consequence of all these issues, improving rice cultivars is essential, not optional, to ensure increased productivity to fill the gap between production and demand for rice.

1.2 Rice genomics

Rice is the first food crop for which a genome sequence was completed. It is an ideal model plant for investigating the genetics of grasses, due to its small genomic size (approximately 430 Mb) in comparison with other major crops like wheat. A high-quality reference genome is available now. This resource has accelerated rice research to improve it in all aspects: yield, environmental stress tolerance, pest and disease resistance, quality and nutrition.

1.3 Australian wild rice

The *Oryza* genus has 26 species, all of them wild except two, and it is believed that they have many genes that will be very useful in rice improvement. Among those wild species, the Australian wild rice species AA genome group has vital importance. *O. rufipogon* and *O. meridionalis* populations from northern Australia represent intact genomic rice resources due to: isolation from other rice species both domesticated and wild; being far from cultivated rice fields in Australia; and being geographically isolated by sea from Asian populations. This has helped preserve the Australian populations from the genetic impact by gene flow from domesticated rice, which has been found in the Asian wild population. The uniqueness of the Australian wild rice, morphologically and genetically, suggests it is very valuable to plant breeders.

1.4 Advanced technologies

Classical breeding has improved both the quality and quantity of rice production. However, this process takes a long time and effort and is also expensive, so there is a serious need to develop and employ new methods that are effective, consume less time and are less costly. Next generation sequencing (NGS) has great potential for use in developing crops generally and rice in particular. This new approach promises the discovery of new genetic resources. NGS provides an opportunity to comprehensively view the whole genome and allow us to dig deeper into these resources to contribute to solving food security problems.

1.5 Aim and Objectives of the project

The aim of this study was to conduct a wide survey of all wild rice plants in north Queensland starting from Townsville up to the tip of Cape York. Sample collection was designed to cover all easily reached areas. The wild rice populations in this area are important because they can be considered as genetically intact, because they are isolated geographically from large scale domesticated rice production in southern Australia and are separated from Asian populations by sea. They are unlike the other wild rice accessions in the world (Asia, Africa and South America) that are close to domesticated rice fields and have no barrier to prevent mixing with domesticated rice physically as whole seeds, or via pollen transfer.

The whole genome was sequenced to study the genetic relationships in these populations and other domesticated rices at two levels: the chloroplast genome to track the maternal inheritance, and the nuclear genome. This will clarify the genetic distinctness of two potential taxa described recently in these populations (Taxa A and B). Because of the potential role they have as a major part of the primary gene pool of rice, it will be very important for the global rice research community to verify the status of these populations and answer other questions: how many divergent taxa are there? To what extent do they differ from other populations? and, Are these differences sufficient to consider any of the populations as new species?

Understanding the relationship between these populations and cultivated rice may allow researchers to develop enriched breeding programs with potential reservoirs of new genes that have not been used before in the development of rice cultivars, and thereby to provide appropriate new resources that meet the challenges posed by global climate change and satisfy food security insurance. Early rice selection and breeding focused on just a few traits, and this may have led to parts of the genome that have traits now considered useful, being omitted during the domestication process. Studying starch related genes in the north Queensland uniquely wild populations will give a better understanding of how we can use these genes to enhance the quality and nutrition of rice, especially after the linking of starch properties with recent disease threats such as colon cancer and diabetes.

1.6 Research plan

The research plan was to:

1. Collect samples between Townsville and the tip of Cape York. Vegetative material and seeds were to be collected if available. Additionally a site description was to be written, the GPS noted and pictures taken.
2. Extract high molecular DNA and measure the quality and quantity, as only good quality would be used for sequencing.
3. Sequence samples using the 150bp paired end technique and Illumina Hiseq 4000 machine.
4. Obtain an assembled chloroplast genome sequence from the NCBI data base. This would be used as a reference to assemble the chloroplast of the Asian and Australian wild populations
5. Assemble the chloroplast of the Asian and Australian wild rice with dual pipeline in order to reduce the assembly errors.
6. Study the genetic relationship of Asian wild rice with other *Oryza* AA genome based on chloroplast level.
7. Study the genetic relationships of Australian wild rice Cape York populations with other *Oryza* AA genome species at the chloroplast level.
8. Study the genetic relationships between Australian wild rice Cape York populations and other domesticated rice populations at the nuclear genome level.

9. Study starch related genes in Australian wild rice population. Thirteen genes were nominated for studying their relationships, namely: *ISA2, ISA3, PUL*, *SBE1*, *SBE3, SBE4*, *SSI*, *SSII-1*, *SSII-2, SSII-3, SSIII*, *SSIV* and *GBSSI.*

Chapter 2

2 Literature review

2.1 Genetic diversity and the environmental impact

Diversity occurs among plants due to a combination of factors (mutation, migration, recombination, selection and drift). Basically, it arises from the interaction between the reproductive system of a species and the environment. Changes in the environment have influenced different genetic selection processes during the evolutionary history of plant species. In addition, the reproductive system of plants plays an important role in the development of the species. For instance, in terms of its being sexual or asexual, unisexual or bisexual and whether it is monoecious or dioecious (De Vicente et al., 2004).

The relationship between genetic diversity and the environment is reciprocal. In other words, the impact of the environment leads to diversity within the population and the diversity within the population leads to a population's ability to cope with the harshest environments. A population which has less variation in its genome will be faced with extinction faster than a population with more variants in its genome. An important point which needs be considered, is whether the differences between populations are related to the genome itself or are a response to the impact of the environment (phenotype). There is no way of knowing the basis of variations in populations without examining the genetic material. Recent applications of new molecular techniques have proved that the phenotype is not necessarily a complete reflection of the genotype and that there are some silent genes that do not express because they are either controlled by other genes, or they need a specific environmental effect to express. Therefore, evaluation studies need to be at the molecular level to escape environmental interference. The sample number may have a great influence on the allele frequency detected and it needs to be large enough to represent all genotypes in the population (De Vicente et al., 2004; Huang et al., 2016).

2.2 Diversity in genus *Oryza* spp

The *Oryza* genus, which belongs to the *Poaceae*-grass family, has 26 species. Of Asian origin, are *Oryza sativa,* sub species *japonica* and *indica;* and of African origin, is *Oryza glaberrima,* both of which were domesticated thousands of years ago. In addition, there are 24 wild type species (Table 1). This diversity resulted from natural selection over millions of years, commencing with the ancient breeds. Wild species are quite distinct from each other morphologically and genetically (Figure 1) (Sanchez et al., 2013). Determining the relationships among these wild species and domesticated rice is interesting; thus it has been studied extensively. In order to maximise the benefits of these diverse resources and improve the current varieties (Li and Zhang, 2012; Wambugu et al., 2015) described the distribution of the *Oryza* species AA genome based on chloroplast DNA analysis, within five main groups (Figure 2).

2.3 From wild to domesticated evolutionary background

Oryza sativa was domesticated 9000 years ago. There are two theories as to the origin of its domestication. The first theory is about a single origin for the domesticated rice, which suggests that *O. sativa japonica* and *O. sativa indica* came from the domestication of the wild rice, *O. rufipogon.* The second theory concerns multiple independent domestications, which means domestication processes occurred separately (He et al., 2011; Londo et al., 2006; Sang, 2009; Sang and Ge, 2007).

According to Vaughan et al. (2008), the evidence that supports a single event in rice domestication history relates to shattering and seed colour genes (*sh4*, *rc*) and strong bottlenecks in local geographic areas. Secondly, the single event is supported by the reappearance of the characteristics of wild species in the segregations that come from crossing O. *sativa* ssp. *japonica* and *O. sativa* ssp. *indica*. The fact that a group of cultivars tends to present unique alleles from unrelated wild populations, supports the single event theory. Furthermore, there is diversity of the cytoplasm when comparing wild and cultivated rice. However, the sequencing of the genes and genotyping methods indicate that *indica* and *japonica* are related to different ancestors. Finally, the separation between *japonica* and *indica* is estimated to be 0.4-0.2 Mya and this date is distant from the rice domestication event.

The SNP pattern of 630 genes on three selected chromosomes (8, 10 and 12) from wild and domesticated accessions showed 20 apparent discriminating sweeps, which supports the single origin theory. As well, domestication dates back 8200 to 13500 years before the present (B.P.) based on the molecular clock, while the estimated time of separation between domesticated and wild is around 3900 years (B.P.) when based upon the archaeological evidence (Molina et al., 2011). Both *O. sativa* ssp. *japonica* and ssp. *indica* show genes concentrated in limited regions, causing their density to be high compared to that of the wild *O. rufipogon*. This distribution is subsequent to strong selection during the domestication process (Flowers et al., 2012). From a sequence of about 1500 cultivated and wild rice covering the Asian continent, 55 selective sweeps related to domestication were found. The conclusion is that O. *sativa japonica* was first domesticated in southern China in the Pearl River area, whereas O. *sativa indica* was developed as a result of crossing between O. *sativa japonica* and local wild rice, which then spread to South East and South Asia (Huang et al., 2012).

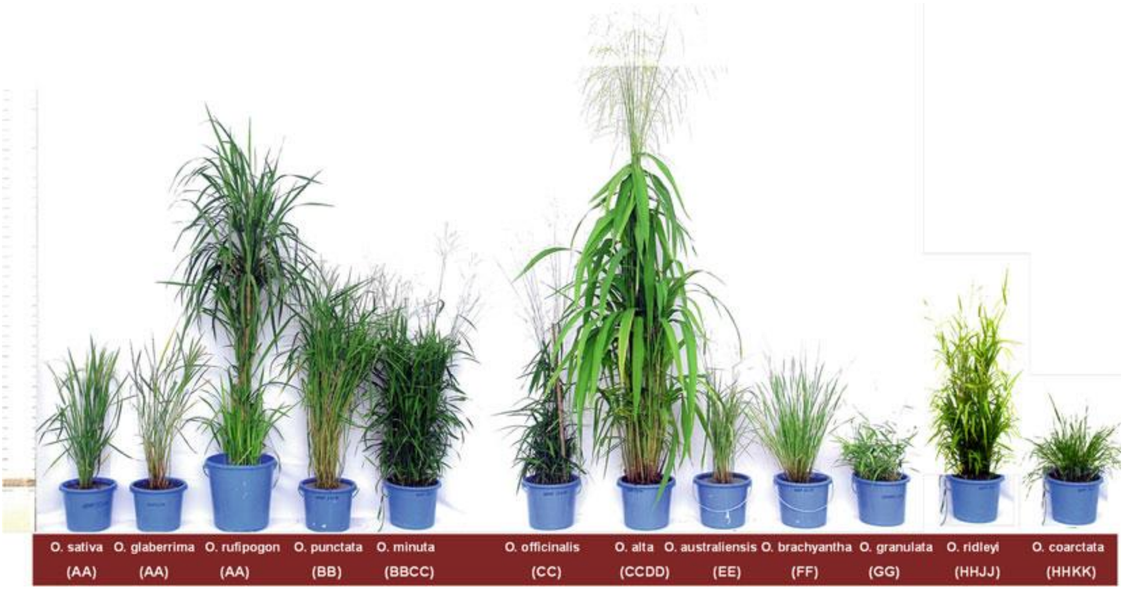


Figure 1. The picture shows the difference among 12 *Oryza* species at the same development stage (Sanchez et al., 2013)

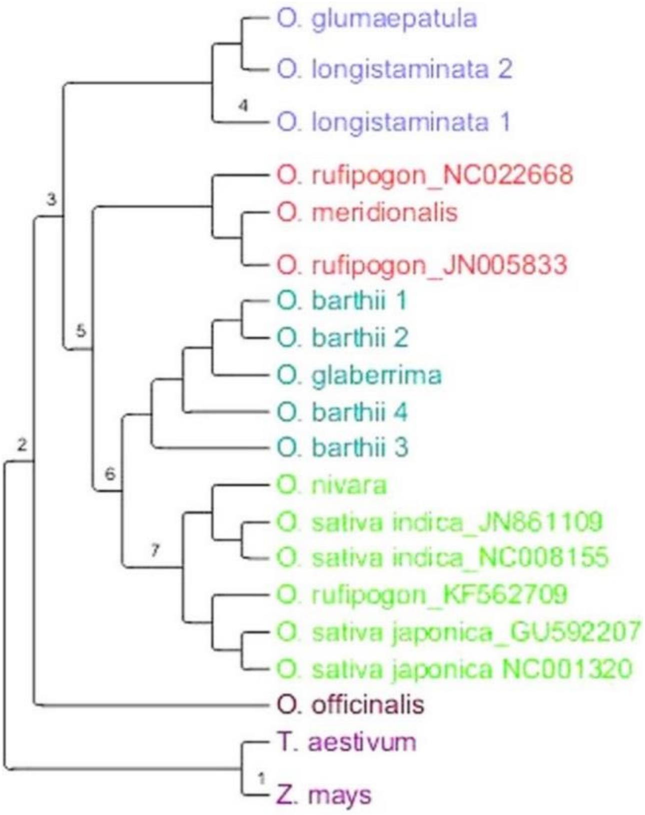


Figure 2. The relationship between *Oryza* species AA genomes based on chloroplast DNA analysis (Wambugu et al., 2015)

Table 1. *Oryza* species, genome group, chromosome number and the geographical origin (Joseph et al., 2008; Koh et al., 2015) and http://www.gramene.org/)

	Oryza species	Genome group	Chromo. number	Origin	Wild / Domesticated
1	*O. officinalis* Wall ex. Watt	CC	24	Tropical Asia	Wild
2	*O. perennis*	AA	24		Wild
3	*O. punctata* Kotschy ex Steud.	BB, BBCC	24, 48	Philippines and Papua New Guinea	Wild
4	*O. rhizomatis* Vaughan	CC	24	Sri Lanka	Wild
5	*O. ridleyi* Hook	HHJJ	48	South Asia	Wild
6	*O. rufipogon* Griff.	AA	24	Tropical Asia	Wild
7	*O. sativa ssp japonica* and *ssp indica*	AA	24		Domesticated
8	*O. schlechteri* Pilger	HHKK	48	Papua New Guinea	Wild
9	*O. alta* Swallen	CCDD	48	South America	Wild
10	*O. australiensis* Domin.	EE	24	Tropical Australia	Wild
11	*O. barthii* Chev. et Roehr	AA	24	Africa	Wild
12	*O. brachyantha* Chev. et Roehr	FF	24	Africa	Wild
13	*O. coarctata* Roxb.	KKLL	48	India	Wild
14	*O. eichingeri* Peter	CC	24	South Asia and East Africa	Wild
15	*O. glaberrima*	AA	24	Africa	Domesticated
16	*O. glumaepatula* Steud. (*Oryza glumaepatula*)	AA	24	South and central America	Wild
17	*O. grandiglumis* Prod.	CCDD	48	South America	Wild
18	*O. granulata* Nees et Arn. ex. Watt	GG	24	Southeast Asia	Wild
19	*O. latifolia* Desv.	CCDD	48	South America	Wild
20	*O. longiglumis* Jansen	HHJJ	48	Indonesia	Wild
21	*Oryza malampuzhaensis*	BBCC	48	South India	Wild
22	*O. meridionalis* Ng	AA	24	Tropical Australia	Wild
23	*O. meyeriana* Baill	GG	24	Southeast Asia	Wild
24	*O. minuta* J.S. Presl. ex C.B. Presl.	BBCC	48	Philippines and Papua New Guinea	Wild
25	*O. nivara* Sharma et Shastry (*Oryza sativa f. spontanea*)	AA	24	Tropical Asia	Wild
26	*O. longistaminata* Chev. et Roehr (*Oryza glumaepatula*)	AA	24	Africa	Wild

In contrast, Civáň et al. (2015) reanalysed the previous data (1500 rice accessions) and they identified three independent regions for the domesticated rice event. They suggested the *japonica* population originated in Southern China and the Yangtze valley; that *indica* could be traced back to the Indochina population and Brahmaputra Valley, and the *aus* back to central India and Bangladesh. Finally, *aromatic* rice was found consequent to hybridisation of the *japonica* and *aus* strains. This confusion should be clarified, as in some cases, nucleotide polymorphism might fail to explain the history of rice selection and domestication. There are four possibilities to be considered in order to clarify the confusion: 1. the gene is not part of the selection target; 2. Variation could have assigned polymorphism to different regions; 3. the statistical design of the experiment is not sufficient to detect variations; 4. use history knowledge to track back the evolution of this population (Doebley, Gaut & Smith, 2006).

2.3.1 Traits influenced by domestication

QTL comparison between the domesticated rice and wild ancestor *O. rufipogon* shows three regions in chromosome 3 are associated with five domesticated traits: seed shattering, tillering, flowering time, grain weight, and seed percentage per set. Tropical *japonica* shows low nucleotide variation compared to the wild varieties, with only 37 SNPs, 36 of them in silent sites. On the other hand, *indica* shows high variation–288 SNPs, 276 of them located in silent sites. In other words, the diversity of silent sites in wild species is six times higher than in the domesticated species (Xie et al., 2011).

2.3.1.1 Panicle shape (open /closed)

The *OsLG1* gene controls ligule development in rice and gives the panicle shape. The expression of the *OsLG1* gene was found to be much higher in the open panicle than in the closed one. In addition, it has been found that there are 12 SNPs and six base pair insertions/ deletions between the wild type *O. rufipogon* and O. *sativa*. One of those SNPs (G) was highly consistent in all wild types, whereas (A) was found in all domesticated cultivars (Zhu et al., 2013).

2.3.1.2 Shattering genes

Shattering related genes have received much attention due to their relation to the beginning of domestication. There are several types of mutations on chromosome 4 that control the shattering trait: A. one base pair substitution; B. mutations in the first exon 15 bp or 5 amino acid; C. 3bp or one amino acid insertion/ deletion; D. 1 bp or amino acid substitution and three mutations in the 5` of the starting codon; E. 1 bp substitution at site 55; F. 3 bp insertions /deletions between sites 343 and 344;

and G. 8bp insertions/ deletions between sites 558 and 559 (Li et al., 2006).

The seed shattering gene *sh4* showed probability of taking a role in the cell death event sequence or in releasing hydrolic enzymes. This enzyme is responsible for softening the cell bonds in the abscission layer, which leads to release of the seed from the spike. However, variation in one nucleotide in the cis-regulation of the *qSH1* gene causes diminishment in its expression in the cell and produces a non-shattering trait (Doebley et al., 2006; Sang, 2009). The *SHA1* gene has control of the seed shattering in *O. sativa japonica* and *indica*; a single nucleotide change from G to T leads to change in one amino acid–from lysine to asparagine, which switches the phenotype from shattering to non-shattering (Lin et al., 2007; Zhang et al., 2009).

2.3.1.3 Seed colour genes

White rice seeds (non-pigmented) have been found to exist through loss-of-function mutations which are encrypted to a protein that regulates the pathway of proanthocyanidin synthesis (Gross & Olsen, 2010). In white rice, the *Rc* gene has divided into two independent mutations: either 14 base pair fragment deletion, which has been found in 98% of white rice (this deletion basically was found in *japonica* cultivars then transferred to *indica* cultivars); or a single nucleotide substitution that causes a stop codon (Sang, 2009).

Later, it was discovered that the *Rc* gene is controlled by three different mutations which regulate anthocyanin production in rice grain. These mutations are responsible for removing the red pigment in the seed originally found in the wild ancestor, *O. rufipogon.* The deletion of 14-bp in exon 7 (causing frameshift translation) is the only mutation that has been found consistently in all white seed species and was not in all wild accessions. The other two mutations are almost variations of this mutation. One of the mutations seems to be fixed in O. *sativa japonica* cultivars only, while the other mutation is likely to cause a light red colour (Meyer and Purugganan, 2013).

2.3.1.4 Awnless seeds

Awns are controlled by a major gene (*awn1*) *LABA1* on chromosome 4. This gene is involved in cytokinin enzyme activation, which plays a role in cell division and growth. A frame-shift deletion in *LABA1* that has been found in cultivated rice, causes a significant reduction in the concentration of cytokinin in awn primordia. This leads to disruption of primordia elongation in the awn (Hua et al., 2015).

2.3.1.5 Other traits

The *BADH2* gene, with several mutations, has controlled the aromatic trait in most aromatic rice accessions (Gross and Olsen, 2010). Table 2 shows gene variations related to domestication events. The differences vary from SNP in the intron or in the open reading frame region, to the deletion range of nucleotides 14-1000 bp (Doebley et al., 2006; Gross and Olsen, 2010; Izawa et al., 2009).

Table 2. Genes related to the domestication process. A functional nucleotide polymorphism in a specific region leads to changes in traits.

	Genes	FNP	Trait	Functions that are affected by changes
1	Wx	SNP at the first intron on 5′ splice site	Texture/taste of rice	The synthesis of granule-bond in starch
2	sh4	SNP leads to changes in amino acid in the ORF	Seed shattering	MYB transcriptional activator protein
3	qSH1	SNP leads to changes in the expression pattern in the promoter region	Seed shattering	BELL (homeobox) transcript factor
4	Rc	Deletion of 14 bp leads to premature stop codon	The color of seed pericarp	bHLH transcript factor
5	Rd	Two separate SNPs cause premature stop codons	The color of seed pericarp	DFR (Dihydroflavanol-4-reductase)
6	qSW5	1 kb deletion	The width of seed	unknown
7	Gn1a	Deletion of 16 bps in the ORF	The number of grains per panicle	Cytokinin oxidase
8	Ghd7	Several FNPs	Flowering time	CCT motif protein
9	sd1	Deletion of 383 bps	Plant height	GA20 oxidase
10	PROG1	SNP leads to changed amino acid in the ORF	Plant stature	Zn-finger transcript factor
11	GIF1	FNPs in promoter region	Grain filling	Cell wall invertase
12	Sdr4		Seed dormancy	
13	GS3		Grain size	
14	GW2		Grain width and grain weight	
15	BADH2	Deletion	Fragrance	
16	Ghd7	Deletion	Grain number, plant height and heading date	
17	Phr1	Insertion / deletion	Grain discoloration	
18	Gn1a and gn1	Deletion Stop codon	Grain number	Dehydrogenase / Cytokinin oxidase
19	ehd1	Changes in one amino acid	Flowering time	Type B regulates the response
20	hd1	Dislocated in coding sequence	Flowering time	Transcriptional regulator
21	hd6	Stop codon	Flowering time	Protein Kinase

2.4 Valuable Characteristics of Wild rice

Wild species in general, and wild rice in particular, are in danger of extinction. Many factors have impacted on these valuable natural resources. For example, climate change in terms of changes in temperature and rainfall, has had great impact on the survival of wild plants. Moreover, competition with other weedy plants and the grazing of animals destroy their chances for survival. Diversity is the key to species survival. Simply put, if there are no differences or if there is less heterozygosity between populations, they will become extinct or unfit for purpose at the first unsuitable circumstance they face (Henry et al., 2010; Reed and Frankham, 2003; Zhu et al., 2000).

The domestication process added great value to cultivated rice by focusing on people's favourite traits (like large fruit size, more kernels, coincidence of flowering and maturity, removal of shattering in grain crops, reduction in seed dormancy or elimination of it in some crops *etc*). However, much valuable genetic material has been lost during the processes of grain refinement, such as closed hybridisation and back crossing (De Vicente et al., 2004; Krishnan et al., 2014). Between 50 and 60% of allele numbers have been lost when comparing the cultivated variety to the wild. In other words, 40-50% of the genepool has been lost (Sun et al., 2001). Moreover, artificial selection during domestication processes negatively affects cultivars by allowing the accumulation of several deleterious mutations. These deleterious mutations lead to reduction in cultivar reproductive fitness for facing climate change (Lu et al., 2006).

2.4.1 Disease and pest resistance

There are many examples of useful traits that have been successfully introduced to cultivated rice from its wild relatives. The first is disease resistance, from Blast resistance genes *Pi-9* (t) and Pi-40, which were introduced from the wild rice *O. minuta* and *O. australiensis* respectively (Kole, 2011). The *Pirf2-1*(t) gene located on chromosome 2 *O. Rufipogon* has an important role in providing non-specific resistance to rice Blast disease and contributes to a dominant mode of resistance to it (Utani et al., 2008). Furthermore, successfully introduced blight resistance genes *Xa21, Xa23, Xa27 Xa29(t)* and *Xa30* were from wild relatives *O. longistaminata, O. rufipogon, O. minuta, O. officinalis* and *O. nivara* respectively. In addition, viral resistance to Tungro disease comes from the *RTSV* gene that is derived from the ancestor, wild *O. rufipogon.* Secondly, pest resistance, in particular yellow stem borer resistance, comes from the wild rice *O. longistaminata,* and the brown plant hopper resistance genes *Bph10 and Bph18(t)* from *O. australiensis; Bph14, Bph15* from *O. officinalis; bph11, bph12(t)* from *O. eichingeri;* and *Bph20(t), Bph21(t)* from *O. minuta*. (Zhang and Xie, 2014). Also, *O. nivara* has a dominant gene resistant to grassy stunt disease (Khan et al., 2015).

2.4.2 Abiotic stress resistance

Soil salinity has serial impact on seed germination, reduces plant growth, damages the chloroplast structure and decreases photosynthesis. *O. coaretata* has a salt resistance trait. This species has specific unicellular hairs (trichomes) which are responsible for maintaining the salt concentration at the lowest level in leaf tissue. Cold resistance at seedling stage, aluminium toxicity tolerance and tolerance to acid sulphate traits have been found in *O. rufipogon,* and iron toxicity tolerance in both *O. rufipogon* and *O. glaberrima* (Bal and Dutt, 1986), as well as other abiotic resistance genes in *aus* accessions (Schatz et al., 2014). Moreover, (Duan and Cai, 2012; Hadiarto and Tran, 2011) reported several genes related to abiotic stress resistance (Table 3).

Table 3. Abiotic resistance genes in *O. sativa*

	Gene	stress	Species
1	*SUB1A*	flooding	*O. sativa*
2	*SK1* and *SK2*	flooding	*O. sativa*
3	*HKT1;5*	Saline soil	*O. sativa*
4	*NRAT1*	High Al3+	*O. sativa*
5	*PSTOL1* at the *Pup1 locus*	Low Phosphorus	*O. sativa*
6	*(OsPIP1, OsPIP2),*	Reduced transpiration, water use efficiency	*O. sativa*
7	*(OsCDPK)*	Rooting system efficiency	*O. sativa*
8	*OsLEA3-2*	Salt / drought	*O. sativa*

2.4.3 Productivity

Many QTLs responsible for increasing the yield have been reported found in *O. rufipogon* accessions from China and Malesia, and successfully transferred to the domesticated rice *O. sativa* (Fu et al., 2010; JinHua et al., 1996; Li et al., 2002). High expression of the *Os11Gsk* gene was found associated with high yield in the introgression line *O. rufipogon* (Thalapati et al., 2012). Furthermore, yld1-1 on chromosome 1 marker RM5, and yld2-1 on chromosome 2 marker RG256, were linked to yield improvement in *O. rufipogon* (Zhang and Wing, 2013). Moreover, agronomical traits (days to heading, number of spikelets in panicle, and shape and weight of the grains) of *O. sativa* have been improved by introducing new alleles from the wild relative *O. grandiglumis*. (Yoon et al., 2006)

2.4.4 Health and nutrition importance

Recently, there has been a rapid increase in type -2 diabetes cases throughout the world. This has had increasing association with rice consumption, which constitutes the main meal of more than half the world's population and is regularly eaten by about another 11%. Many studies have been focused on starch characteristics as a major carbohydrate component of the grains because of the emphasis on increased glucose percentage in the blood (Glycemia), also known as postprandial hyperglycemia PPHG. According to that, rice starch has been categorised as both high and low on the Glycemic Index. The low indexed rice is preferable, because it keeps PPHG under control after consumption and there is less risk of developing type -2 diabetes if it is eaten (Garaycochea et al., 2015). Starch synthesis is a process that is formed by about 18 combined genes. They are all, together, responsible for starch amount, and for the amylose/amylopectin ratio, and for other starch properties. This leads to the configuration of the Glycemic Index (GI) (Hu et al., 2012; Kharabian-Masouleh et al., 2012). Most recently, it has been found that Australian wild rice has the highest amylose content, which can improve the glycemic index of the current cultivars, and provide healthier products (Tikapunya et al., 2017b).

2.5 Wild *Oryza* species in Australia

Four *Oryza* wild species have been natively found in the northern part of Australia, namely: O. *meridionalis, O. australiensis, O. officinalis* and *O. rufipogon*. These species were indigenous in remote areas and so uncontaminated by human bred cultivars, which kept it as an intact genepool for potential new abiotic, biotic resistance genes and nutrient grain quality (Henry et al., 2010). Reports have shown zinc, phosphorus and magnesium percentages are higher in the wild rice grains compared to commercial cultivars, which is possibly because their nutrients can be taken up more efficiently. Furthermore, the sodium concentration in the wild leaf was lower than in the cultivated, which means the wild plants must be using special mechanisms to avoid accumulating sodium in their cells (Wurm, 2012).

2.5.1 *Oryza* australiensis

Oryza australiensis is found in the North of Queensland, the northern part of the Northern Territory and in Western Australia, according to Australia's Virtual Herbarium (http://avh.chah.org.au). This species has grown in areas considered relatively dry for the *Oryza* species. They usually overcome the dry season as rhizomes or seeds. The *O. australiensis* genome size has doubled (965 Mb) as a result of the accumulated retrotransposon copies through its lineage over millions years (Henry et al., 2010); (Piegu et al., 2006).

2.5.2 *Oryza* officinalis

The information about this species is very poor and some reports refer to collections of it from two sites in the north of Queensland (Moa Island) and the Northern Territory. Further investigation is required according to (Henry et al., 2010). A recent study (Wambugu et al., 2015) showed that this species stands out from all the other *Oryza* species AA genome groups, based on chloroplast sequencing analysis (Figure 2).

2.5.3 Oryza rufipogon

This species has been found to be widespread in many locations in the North of Queensland, the Northern part of the Northern Territory and Western Australia, as reported by Australia's Virtual Herbarium (http://avh.chah.org.au). However, these reports were conducted years ago, and were based on classical classification keys. Many of these records mixed up the *O. rufipogon* and *O. meridionalis*, especially before 1981 (the date of separating this species out and giving it a new name). Both were classified as *O. rufipogon*. This has been proven by molecular analysis using *SINE* marker. Fourteen of 24 accessions were classified as *O. rufipogon*, but in fact they are *O. meridionalis* (Xu et al., 2005b).

Recently, (Sotowa et al., 2013), found that the *O. rufipogon* samples in the north of Queensland have a unique morphological characterisation and are distinct from the Asian *O. rufipogon.* This has led to a huge argument about whether this finding applies to all Australian wild rice from other places, or just to these samples from North Queensland, due to its isolated location. Most recent molecular analysis based on the chloroplast genome for these samples, showed that Australian and Asian *O. rufipogon* is divided into two different clades. This point has opened the door to describing it as a new species (Wambugu et al., 2015). All the above considerations lead to this question: Can we treat *O. rufipogon* in all Australian states as the one species or not?

2.5.4 Oryza meridionalis

O. meridionalis is widespread and endemic to Australia and New Guinea. It is found in the north of Queensland, the Northern Territory and Western Australia. It is an annual species, surviving the harshest seasons as seeds. Before separation as a new species in 1981, its samples were classified as *O. rufipogon* (Ng et al., 1981). The interaction between both *O. rufipogon* and *O. meridionalis* which has been found in Australia, needs more investigation to explain the extent to which these species are genetically distinct from each other (Henry et al., 2010).

2.5.5 Oryza nivara S.D. and *O. minuta*

Some reports have suggested that *O. nivara* and *O. minuta* may be found in Australia; however, these reports have probably confused *O. officinalis* with *O. minuta*. This confusion probably applies to *O. nivara* as well, due to its high similarity to O. *meridionalis* (Groves et al., 2009).

2.5.6 *Oryza spp.* Taxon A and Taxon B in North Queensland *Oryza spp.* Taxon A and Taxon B in North Queensland

A recent study that discovered that two perennial populations in Australia are distinct genetically from the *O. rufipogon* found in Asia has been undertaken on a collection of *Oryza* AA genome species, gathered from throughout the Asian continent and Oceania. The first species has a similar appearance to *O. meridionalis* (hereafter referred to as Taxa B). The first one has a similar appearance to *O. meridionalis* (hereafter referred to as Taxa B), and the second one is more closely aligned to *O. rufipogon* (hereafter referred to as Taxa A (Sotowa et al., 2013). Furthermore, these studies suggest that the origin of Taxa B was evolutionary mixed mutations, segregation and natural selection from the ancient form of the *O. meridionalis,* which led to its becoming a new perennial species. The differences between the two are clearly seen in the shape of spikelets and lemma. On the other hand, Taxa A is possibly derived from Asian *O. rufipogon* and was later introduced to Australia.

Later studies (Brozynska et al., 2014b; Brozynska et al., 2017; Moner et al., 2018; Wambugu et al., 2015) using NGS data on both the chloroplast and nuclear levels showed the unique characterisation of the Australian wild rice (Figure 2 andFigure 3). The importance of these taxa lies in their having been found in remote areas geographically and far from human intervention and cross pollination with domesticated rice, which kept them as pure as ancient wild rice.

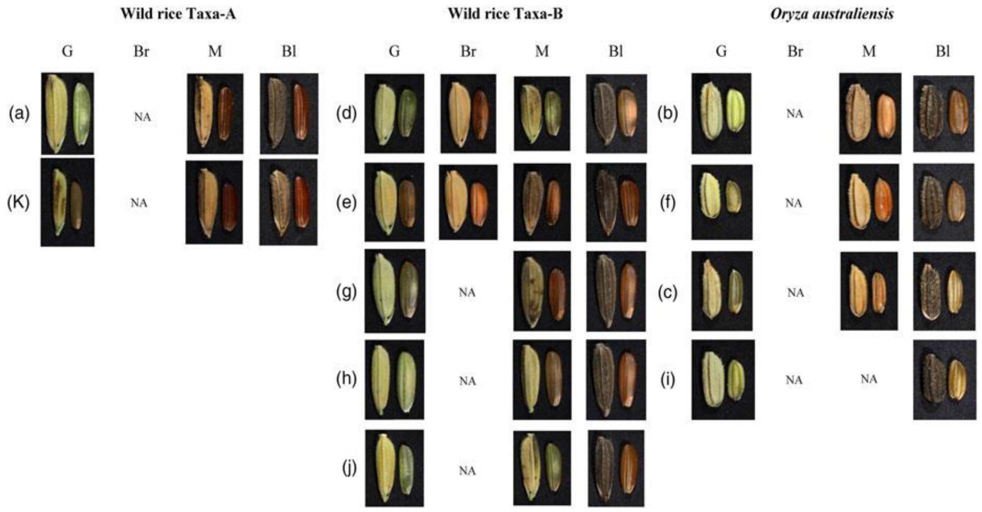


Figure 3 Grain appearance of the Australian wild rices (Tikapunya et al., 2017)

2.6 Genetic diversity analysis

Diversity between creatures is one of the oldest topics argued among researchers. They question why creatures are diverse, how to group them, what the basis for a classification is, and one of the most important questions is, ‘What is the cutting edge between two populations that divides them into two different groups?’ Many scientific researchers have developed various methods of measuring the differences between specific populations in order to organise them into groups to simplify studying them and to find the relationships among them in terms of their evolution, based on morphology or agronomy characteristics or biochemical reactions–and recently, DNA molecular markers. Getting this knowledge allows researchers to better understand the biological system interfaces. In addition, choosing the right parents to hybridise and finding new resources to enhance existing cultivars is important to them. Genetic diversity measurements based on recent advances in technology have become extremely sophisticated (Mondini et al., 2009; Weir, 1996).

2.6.1 Molecular genotyping tools:

Molecular genotyping involves using molecular markers to identify the relationship between two individuals or two populations. This could be used to study the polymorphic rate in the population, the allele numbers to each polymorphic gene and the percentage of heterozygous (Karp et al., 1996). It has been reported that there are 38 molecular techniques (SNP, SSR, AFLP, CAPS, SSCP, etc.) used in assessing plant genetic diversity. They vary in accuracy, sensitivity, cost, time consumption and complexity. A good molecular marker should: be polymorphic; provide clear resolution of the genetic variety; be easy to use and cost effective; need only a small amount of tissue or DNA; be linked to the phenotypic character; and not require previous studies. In fact, there is no molecular marker that has all these features, but markers are selected according to the work requirement in a specific case, depending on the level of the polymorphism, cost, equipment availability etc. (Mondini et al., 2009).

Next generation sequencing (NGS) makes whole genomic sequencing accessible and reliable in terms of the cost and time needed to get the rows of data. The advantage of this technology is that it overcomes all the previous challenges that faced the earlier molecular markers. This is simply because the comparison is grounded in the “original code” or entire genome of the individual, or samples that represent the population. This allows deep study of the differences in the populations constructed on the original DNA sequence of the organism. However, analysing these data is not an easy job and is itself a new challenge. A number of generations of platforms have been developed during the last decade. Competition in terms of the amount of data, cost and speed are the main features of those new generations. For instance, the amount of data about a single cell Hiseq X can cover the wild rice *O. rufipogon* genome approximately 2000 times. This depth of reading will make judgments on variations more confident.

2.7 NGS application in rice genetic diversity analysis

2.7.1 Specific gene sequences

Early, when sequencing was costly, the NGS technique was utilised for specific regions which may or may not have been studied before, according to its classification or function importance, like functional nucleotide polymorphisms (FNPs). (Hollingsworth et al., 2009). For instance, *rbcL* and *matK* chloroplast genes and 20 other regions were sequenced and used effectively as barcodes in order to identify and differentiate rice species. They were also used to associate favorable rice cooking characteristics with functional SNPs in those genes (Kharabian-Masouleh et al., 2012; Schroeder et al., 2012). Another study sequenced 6.4 Kb of the *Rc* genes in Jiangsu weedy rice O. *sativa* f. *spontanea*, (which has red pericarp inhibited), which showed higher nucleotide polymorphism and the segregation proportion of Jiangsu weedy rice than US weedy rice (Li et al., 2014b). The *Gn1a* gene, which controls the cytokinin oxidase dehydrogenase enzyme that regulates the grain number per panicle, has been sequenced and investigated in wild and cultivated samples. Fourteen diverse alleles have been recognized AP1 – AP14, with clear association between them, and spikelet numbers and grain yield, as well as significant diversity, have been recorded. In addition, the AP9 allele was associated with a large panicle and high yield (Wang et al., 2015).

2.7.2 Chloroplast DNA sequencing

To date, more than 850 chloroplast genomes have been deposited in the database (www.ncbi.nlm.nih.gov/genomes). In the *Oryza* genus, 12 different chloroplast species have been released belonging to cultivated and wild rice. Recently, the chloroplast of *O. australiensis* (EE genome) has been released. Researchers have found that chloroplast size in the *O. australiensis* is 135.224 Kbp, which is higher than for all other *Oryza* spp. by approximately 700 bp (Wu and Ge, 2014). Also *O. nivara* chloroplast DNA sequence has been studied and 57, 61 and 159 insertions, deletions and substitutions respectively, were found compared to *O. sativa.* The most substitutions were in the large single copy LSC (68) and (10) in the small single copy SSC. On the other hand, most of the insertions and deletions were in the coding regions of the inverted repeats (Shahid Masood et al., 2004). Moreover, (Tong et al., 2015) evaluated the differences between 30 Korean accessions and five wild and cultivated rice: *O. nivara, O. meridionalis, O. australiensis, O. sativa japonica* and *O. sativa indica.* In total, 180 SNPs and 41 INDELs located in 63 genes and 153 intergenic regions were found. The phylogeny result supported the independent origin theory of domesticated rice *O. sativa indica* and *japonica*. Interestingly, inconsistent and ambiguous results were found when the researchers compared the phylogeny tree of the chloroplast to the nuclear phylogeny from the same study of 1.6 million SNPs.

In Australia, (Brozynska et al., 2014b; Waters et al., 2012) have shown the relationships among *O. sativa* and other wild Asian *O. rufipogon, O. australiensis* and Australian *O. rufipogon* and *O. meridionalis* relatives. More than 850 SNPs have been detected based on chloroplast DNA sequence levels. The O. *australiensis* was the most distinct species from the others (EE genome). The interesting result was that the Australian *O. rufipogon* was closest to *O. meridionalis* – more so than to the Asian *O. rufipogon* 32 and 68 SNP respectively. This suggests the Australian *O. rufipogon* is different from the Asian *O. rufipogon* and could be a new species. Therefore, it has been suggested that *O. rufipogon* could be a perennial form of *O. meridionalis*. This has the potential to be a novel gene pool for improving cultivated rice.

2.7.3 Whole genome sequencing

Whole genome sequencing allows the development of accurate, specific markers which are linked to favorable traits. Furthermore, whole genome sequencing allows the design of markers within wide flanking regions, allowing the tracking of changes and re-combinations in regions surrounding genes in cross breeding systems (Duitama et al., 2015). The completed sequence of O. *sativa japonica* Nipponbare was finalised in 2005 by the International Rice Genome Sequencing Project (IRGSP). They estimated the error rate at less than one per 10 Kb (Kawahara et al., 2013). Then, two genomic assemblies were produced, the first one by the Rice Genome Annotation Project (RGAP) and the second one by the Rice Annotation Project (RAP). It has been noted there were slight differences between both of them, but this confused the rice community when a reference was needed. Therefore, another two individuals of O. *sativa japonica* and Nipponbare were sequenced to correct the previous sequence and to compare the allele diversity among the individuals from the same population (Kawahara et al., 2013). The resequencing project reduced the error rate to 0.15 per 10 Kb, which was a decrease of 85% on the errors in the previous reference. The average allele frequency was 0.20 per 10 Kb, which should be taken into account when comparing diversity among individuals (Kawahara et al., 2013).

An enormous recent project has sequenced 3000 rice accessions of *O. sativa* to represent a wide spread of diversity back to 89 countries with 14 X genome coverage on average. The seeds of all accessions are accessible from the International Rice Genebank Collection (IRGC). Both the sequence data and the source of these data (seeds) constitute valuable repositories for developing and improving cultivated varieties (Li et al., 2014a). In another project, (Huang et al., 2012) sequenced the whole genome of 1,083 varieties of cultivated rice (both O. *sativa indica* and *japonica)* and 446 accessions of the wild rice *O. rufipogon,* the progenitor of the cultivated rice O *sativa*. SNP analysis supports the single event theory of domesticated rice.

In China, 517 from 50 000 accessions, different morphologically, geographically and genetically, have been chosen for sequencing (with around 1 X coverage) to study their agronomical traits. Three point six million SNPs were recorded, approximately one SNP per 9.32 Kb. Those SNPs have been successfully linked to agronomical traits, as shown in (Table 4) (Huang et al., 2010).

Table 4. SNPs and their impact on related agronomical traits (Huang et al., 2010)

Trait	Chromo-some	Position (IRGSP 4)	Major allele	Minor allele	Gene loci
Tiller number	4	3760194	A	T	-
Grain width	5	4907158	C	G	-
Grain length	3 5	17371398 5343949	G A	C G	GS3 qSW5
Gelatinization temperature	6	6726252	C	T	ALK
Amylose content	6	1770929	T	C	Waxy
Apiculus color	6	5335519	A	G	OsC1
Pericarp color	2	27066598	A	G	-
Hull color	6 9	10378142 7366211	T T	C C	- Ibf
Heading date	2	1439288	G	A	-
Drought tolerance	1	5536395	G	T	-
Degree of seed shattering	2	25025325	C	T	-

Another five Korean rice accessions (Dongjin, Korean japonican cultivar and three other culture lines – HY-08, HY-04 and BLB – and their progenitor Hwayeong) have been sequenced with a coverage yield of 61 X. In total, 1,154,063 variations were found: 1,024,202; 53,180 and 76,681 SNPs, insertions and deletions respectively. The largest differences were in the coding regions of five genes that control important functions like ATP binding, signal transduction and the phosphorylation of protein / amino acid. Associating these SNPs with favorable functions will provide valuable sources from which to select SNP(s) which regulate a specific trait (Jeong et al., 2013). Another 94 varieties of O. *sativa* and 10 wild species were sequenced at 2.87-64.83X. 23 million variants were identified: 80% were in the repetitive element, which is extreme. However, changing analysis strategy led to reducing these variants to 4.4 million with 80% of them genotyped (Duitama et al., 2015). Further, 1483 accessions of O. *sativa (sub spp. indica, aus, tropical japonica* and *temperate japonica)* were sequenced at low coverage with approximately 1-3 X. The aim was to assemble individuals at low coverage and not ignore the variation among individuals; for instance, important genes like *GW5, Sub1A* and *Pikm-1* which are absent in the reference O. *sativa* Nipponbare, were found in other cultivars (Marroni et al., 2014; Yao et al., 2015).

In addition, many other studies used the entire genome sequence analysis of both wild and domesticated rice to measure polymorphism levels and genetic diversity. For example, the polymorphism between the O. *sativa ssp. indica* cv. Guangluai-4 and O. *sativa japonica ssp*. cv. Nipponbare has around 1.6 million SNPs with an average 6.9 SNPs per Kb. In addition, about 80,000 and 92,000 insertions and deletions were found, respectively. These SNPs have been distributed across 32 gene families, coding/ non-coding regions, stop codons / prevent stop codons. Likewise, 194 high rate SNPs genes with more than 100 SNPs/ genes, considered as hotspot genes, have been identified. Additionally, more details for several loci which are associated with the important traits *S5, Sub1, LRK Pup1* for hybrid sterility, submergence tolerance, yield improvement and phosphorus deficiency loci respectively, have been provided. Another two million SNPs identified between Korean rice cv. Tongil and *O. sativa japonica* cv. Nipponbare with an average 5.77 SNP/ Kb., showing 91.8% of the total cv. Tongil genome goes back to *O. indica* and 7.9 % comes from *O. japonica* parents (Hu et al., 2014; Kim et al., 2014a; Schatz et al., 2014; Srivastava et al., 2014).

The wild African rice *Oryza brachyantha* (FF genome) has also been sequenced and assembled using Short Oligonucleotide Analysis Package (SOAP) *de novo*. It has been annotated and 32,038 coding genes and a total sequence of 261Mb were reported. *Oryza brachyantha* has a very compact genome compared to other *Oryza* species. It has 22,185 genes which belong to 18,020 families; in contrast, O. *sativa* has 28,830 genes belonging to 20,177 families. In other words, it has shared 17076 and lost 2157 gene families in comparison to O. *sativa*. Besides, 30 % of these shared genes are located in different positions to those in O. *sativa*. These differences could prove important in the ways they can inform efforts to improve cultivated rice and evolutionary research (Chen et al., 2013).

2.8 Starch related genes

Starch, at around 90% of dry rice grain weight, has vital importance as a direct source of energy in the human diet and in the food industry that requires different properties in its products to meet the market's necessities. Recent increases in health problems like obesity, and developing type-2 diabetes or colon disease due to lifestyle changes have led to a rethinking of starch properties such as resistant starch, RS, which could be the solution to the new health threats (Zhou et al., 2016). Starch consists of two kinds of polysaccharide, mainly amylose 15-30 % and amylopectin 65-85%. Amylose has the structure of a linear chain, produced by bonding α 1,4 D-glucose units; while the amylopectin is a highly branched molecule composed of α 1,4 D-glucose units and α 1,6 D-glucose units that are responsible for the branching. The amylose / amylopectin ratio has great impact on the physical and chemical properties of the starch that are reflected in cooking processes. High amylose content tends to fluffy single grains, whereas low amylose tends to glossy when cooked (Dobo et al., 2010; Pérez and Bertoft, 2010; Yan et al., 2009; Yu et al., 2011; Zhang et al., 2014).

Many genes are involved in the starch synthesis pathway, mainly: granule-bound starch synthase I (*GBSSI*), starch synthases *SSI, SSII, SSIII, SSIV,* starch branching enzyme *SBE,* starch debranching enzyme *DBE,* and isoamylase *ISA.* However, the granule-bound starch synthase GBSS-I gene (waxy), which expresses mainly in storage tissue like endosperms, has an important impact on amylose content (Cheng et al., 2012; Dian et al., 2003; Yu et al., 2011).

The multiplicity of genes that are involved in the starch synthesis process makes understanding this pathway very complicated. In Arabidopsis for example, the SS-II deficiency mutant causes an increase in total amylose and in the amylose/amylopectin ratio. On the other hand, the double mutant deficiency in SS-II and SS-III causes sluggish plant growth and decreased starch content (Zhang et al., 2008). Chain length distribution analysis shows mainly independent functionality in SSI, BEI and BEIIb genes; however, BEIIb deficiency reduces the short chain ratio in the amylopectin, and the be2b mutant has more amylose than the wild–probably because of amylopectin synthesis reduction (Abe et al., 2014).

The PUL function to some extent overlaps with that of ISA1, but deficiency in ISA1 has more impact on amylopectin synthesis than PUL (Fujita et al., 2009). Also, (Fujita et al., 2011) suggested just SSI or SSIIIa is essential for starch biosynthesis and remarkably, found 30-33% amylose in high SSI activity and recessive SSIII, while (Kharabian-Masouleh et al., 2012) identified 66 functional SNPs in 18 starch biosynthesis related genes. Thirty-one SNPs were found associated with cooking quality. Other studies have shown resistant starch properties as the result of a deficiency of SSIIIa genes and high expression of waxy genes (Zhou et al., 2016). There is one amino acid substitution on the product of the SBE3 gene, Leucine, in the wild that changed to Proline in the mutant, and this resulted in resistant starch in rice (Yang et al., 2012).

The sequencing of the Swarna rice cultivar that has a low Glycemic Index (GI), showed nearly 1.1 million SNPs and 0.1 million InDels, the majority of them in chromosome 1. The Starch Synthesis Related Genes (SSRGs), except *BEIIa*, have been found polymorphic in Swarna, compared to *O. sativa* Nipponbare (Table 5) (Rathinasabapathi et al., 2015).

Table 5. Starch analysis genes SNPs and InDels in Swarna cultivar modified (Rathinasabapathi et al., 2015)

No	Gene Name	Gene	Non- Coding SNPs	Non- coding InDels
1	ADP- glucose pyrophosphorylase (small unit)	*AGPS2b*	14	1
2	Alpha 1,4- glucan phosphorylase	*SPHOL*	9	1
3	Glucose 6-phosphate-translocator	*GPT1*	9	2
4	Granule-bound starch synthase I	*GBSSI*	10	4
5	Granule-bound starch synthase II	*GBSSII*	82	7
6	Starch synthase I	*SSI*	59	7
7	Starch synthase IIa	*SSIIa*	16	0
8	Starch synthase IIb	*SSIIb*	14	3
9	Starch synthase IIIa	*SSIIIa*	20	2
10	Starch synthase IIIb	*SSIIIb*	13	3
11	Starch synthase IVa	*SSIVa*	11	1
12	Starch synthase IVb	*SSIVb*	17	0
13	Branching enzyme I	*BEI*	9	1
14	Branching enzyme IIa	*BEIIa*	0	0
15	Branching enzyme IIb	*BEIIb*	23	5
16	Debranching enzyme -isoamylase 1	*ISA1*	9	1
17	Debranching enzyme -isoamylase 2	*ISA2*	1	0
18	Debranching enzyme -Pullulanase	*PUL*	47	3

In rice cultivars, three different alleles have been identified in GBSS-I. These alleles vary in the number of CT repeats in the 5′-UTR, as well as in the SNPs in the splicing site of the first intron, exons 4, 6 and 10. This relates to a huge variation in the mRNA expression level of up to 10 times, which is associated with the amount of amylose (Cai et al., 1998; Chen et al., 2008b; Dobo et al., 2010; Hirano et al., 1996; Hirose and Terao, 2004; Isshiki et al., 1998; Larkin and Park, 2003, 1999; Mikami et al., 2008).

(Chen et al., 2017) reported shifting in the exon intron splicing region of *SSII-1* gene, that caused alternative transcript by adding 28 bp fragment to the mature mRNA. Up to ten nucleotides of the edges of the introns and exons (exon, intron splicing enhancer and silencer) have extreme importance, as the edge on intron exon can be shaped the transcriptome. Any change in these regions might influence the protein sequence (Jian et al., 2013; Prathepha, 2007).

Starch has been strongly selected throughout the evolutionary history of rice and is strongly linked to consumer preferences. Wild rice does not have sticky starch, which trait was carefully chosen for rice varieties only after domestication; and development of glutinous rice may have happened in many stages. A SNP in *GBSSI* gene G to A was responsible for decreasing the granule-bound starch synthase activity that changes wild rice to glutinous rice. This mutation first arose in Southeast Asia then spread to the temperate *japonica* varieties. The study of the WAXY gene suggests that this mutation is very rare in the wild species and that it most possibly arose by innovative mutation (Meyer and Purugganan, 2013).

Evolutionary study of the GBSS-I shows two main and six minor GBSS-I haplotypes have been found in wild and domesticated rice. H2 was the most ancient one with 89% of the accessions. In domesticated rice, the GBSSS-I gene had three independent paths in its own evolutionary history. *aus* has the oldest evolutionary path, which agrees with the theory of three independent origins for domesticated rice (Civáň et al., 2015; Kim et al., 2016; Singh et al., 2015; Singh et al., 2017). GBSSI gene variation was less in the wild compared to the cultivated rice, which means different selection pressures have been applied to domesticated rice to meet the demands of different consumer requirements throughout the history of rice domestication (Cheng et al., 2012; Singh et al., 2017; Vaughan et al., 2008).

Australian wild rice has high amylose content and has a different amylose and amylopectin structure from domesticated rice varieties as well as pasting properties and a fine molecular structure, all of which suggests it has an alternative biosynthesis mechanism that can lead to new rice products and the development of new cultivars with a low glycemic index, which is important for diabetic rice (Calingacion et al., 2014; Tikapunya et al., 2017b).

Chapter 3

3 Chloroplast phylogeography of AA genome rice species

3.1 Abstract

Whole chloroplast genome sequence analysis of 59 wild and domesticated rice samples was used to investigate their phylogeny providing more detail on the biogeography of the major groups of wild A genome rices globally. An optimized chloroplast assembly method was developed and applied to extracting high quality whole chloroplast genome sequences from shot gun whole DNA sequencing data. Forty complete high quality chloroplast genome sequences were assembled (including; temperate japonica, tropical *japonica* and *aus*). South American, African wild rice relationship were conformed. The Australian chloroplast type was found to extend north to the Philippines. The remainder could be divided into an African (*O. barthii* and the domesticated *O. glaberrima*) clade and the Asian taxa. The Asian taxa could be placed in two distinct clades including the domesticated O. *sativa ssp. indica* and *O. sativa ssp. japonica* respectively. These two groups of wild rices had substantially overlapping distributions with the *O. sativa japonica* group extending further west into India. The aromatic rices had *japonica* chloroplasts as expected. A polyphyletic maternal genome origin of the cultivated *aus* group of rices was suggested by the identification of *aus* accessions in both the indica and *japonica* clades. The current distribution of the chloroplast types appears to differ significantly to that of the nuclear genome diversity suggesting a complex evolutionary history of the rice progenitors leading to the domestication of rice.

Keywords: Asian wild rice, chloroplast sequence, phylogenetic analysis, *Oryza* AA genome, de novo assembly, mapping assembly

3.2 Introduction

The *Oryza* genus belongs to the Poaceae (grass) family and has 26 species two of which (*Oryza sativa* with two sub species *japonica* and *indica* are Asian in origin and *Oryza glaberrima* which is African in origin,(Wambugu et al., 2013)) were domesticated thousands years ago and 24 of which are wild species (Appendix 2 Table 16). The wild species are morphologically distinct and many display significant genetic diversity. The wild species, in particular the AA genome group of close inter-fertile relatives, have been utilized as genetic resources to improve cultivated rice (Brozynska et al., 2015; Sanchez et al., 2013).

Oryza sativa was domesticated around 8000-9000 years ago based on the archeological evidence (Gross and Zhao, 2014). There have been two distinct theories for the origin of domesticated rice: The first involves a single origin which suggest that *O. sativa ssp. japonica* and *O. sativa ssp. indica* were derived from a common domestication of the Asian wild rice *O. rufipogon.* (Flowers et al., 2012; Molina et al., 2011; Tong et al., 2016; Vaughan et al., 2008). The second theory is multiple domestication events in which the main sub species are domesticated at around the same time in separate areas (He et al., 2011; Sang and Ge, 2007). A common version of the first theory suggests that *japonica* was domesticated first and then subjected to introgression of wild germplasm to form *indica*. This hypothesis is supported by evidence of common domestication alleles in both *japonica* and *indica* (Huang et al., 2012). The second theory proposes multiple independent domestications (Choi et al., 2017; Kumagai et al., 2016). This is attractive due to the significant genetic distance between *japonica* and *indica* clades estimated to be around 1 million years. (Feltus et al., 2004; Xu, 2010)

Substantial recent research has addressed this issue. (Huang et al., 2012) analyzed the SNPs variation (around 8 millions) of 1083 varieties of *O. sativa* subsp. *indica* and *japonica* as well as 446 geographically isolated accessions of *O. rufipogon* from the Asian continent. This study of the whole genome supported the single event theory and divided *O. rufipogon* into three groups (*Or-I, Or-II* and *Or-III*). In contrast, (Civáň et al., 2015) re-analyzed the SNPs variation and identified evidence for domestication of rice in three separate regions. They trace the origins of domestication of *japonica* to populations of wild rice in the Yangtze valley of Southern China and, *indica* to populations in Indochina and the Brahmaputra valley and *aus* to central India and Bangladesh. Aromatic rice was attributed to a hybridization between *japonica* and *aus*.

Recent reports show that Australian wild rice is distinct from other wild rice populations. These populations are different morphologically and genetically and may represent distinct taxa (Brozynska et al., 2014b; Kim et al., 2015; Sotowa et al., 2013; Wambugu et al., 2015). The genetic value of the Australian wild rice populations is enhanced due to their isolation from domesticated rice reducing the potential for contamination by gene flow from domesticated populations and keeping intact the genepool of wild diversity as a reservoir of genes for rice improvement (Henry et al., 2010).

The chloroplast which is a highly conserved maternally inherited organelle in plants, not involved in recombination, has been used as an important tool for analysis of evolutionary relationships and to estimate genetic distance among plant species. *Oryza* chloroplast genomes have a narrow range of sizes around 135 kb and have been used to study relationships within the group (Appendix 2 Table 17). (Ravi et al., 2008; Wambugu et al., 2015).

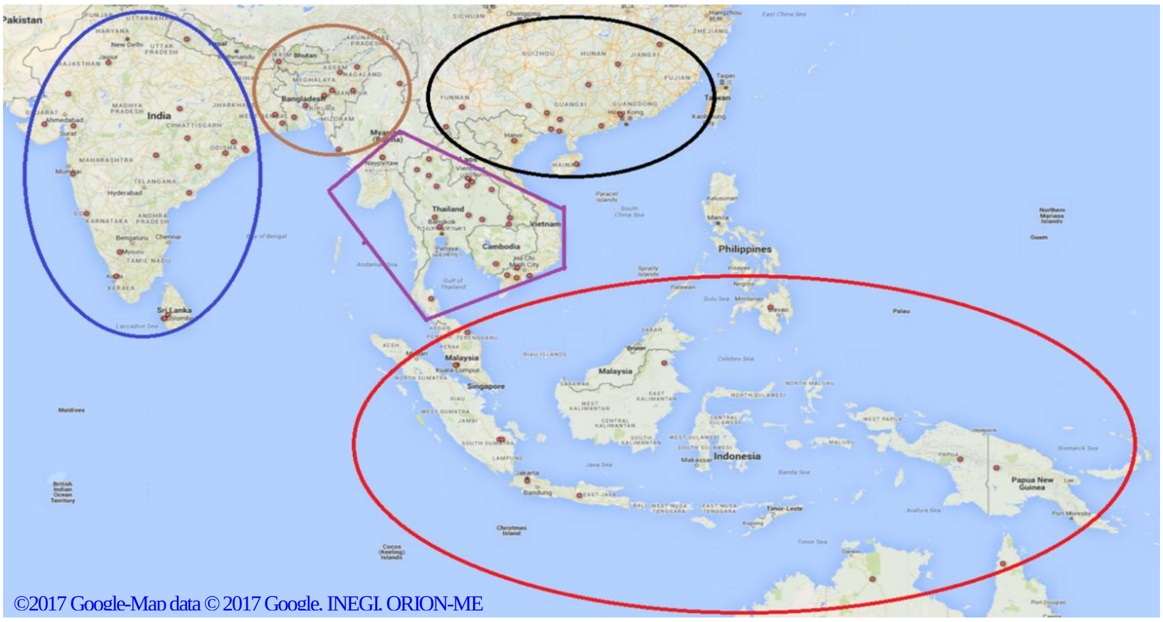
The aim of this study was to assemble and analyze the whole chloroplast genomes from wild populations of AA genome rice and use this to determine the genetic relationships with their geographical distribution, especially between the closest relatives of domesticated rice from Asia and Australia.

3.3 Materials and methods

Raw sequence data for the sequences of Asian *O. rufipogon* and *O. sativa* were obtained from the EMBL website using the links provided by (Huang et al., 2012) . The *O. rufipogon* collections included both perennial and annual *O. rufipogon* germplasm maintained in China and Japan. The whole genome coverage of Illumina sequence reads was between 0.21X and 6.92 X. Samples with sequence coverage between 0.9X and 6.75X were selected. Assuming this coverage will be enough to cover all chloroplast sequence, as there are numerous copies per cell (for instance 1000 -1700 copy of chloroplast genome per cell in *Arabidopsis* leaf) (Zoschke et al., 2007). The locations from which these samples were sourced was examined on a map (Figure 4), and grouped into 5 major geographic zones: (Z1: India, Z2 India and Burma, Z3 China, Z4 Thailand, Vietnam and Cambodia, Z5 Oceania Australia, Papua new Guinea, Indonesia, Malaysia and Singapore). Samples within each zone were grouped further and 6-9 samples were chosen to represent each zone (Appendix 2 Figure 14-16)

3.3.1 Chloroplast genome assembly

Next Generation sequencing (NGS) reads were analyzed using CLC Genomic workbench software and Clone Manager Professional 9, to assemble the chloroplast sequence (Kim et al., 2015). A quality check (QC) was applied to all raw data. Based on the results of the QC report, reads were trimmed to obtain PHRED score above 25. Chloroplast genome sequence for each of the selected accession was assembled using a Chloroplast Assembly Pipeline (CAP) (Appendix 3). Essentially, the method is comprised of a Mapping assembly component (M-component) and a de novo assembly component (D-component). Both the M- and the D-components have two sub-processes designed to reduce errors in the chloroplast sequences



©2017 Google-Map data © 2017 Google, INEGI, ORION-ME

Figure 4. Distribution of 79 Asian wild rice accessions. The accessions were divided into those from 5 different geographic zones for comparison. Map sourced from Google maps.

derived from each of these assembly components. The chloroplast sequences from the M- and the D-components were assembled, mismatches identified and errors resolved by manual curation by observing reads mapped to the mismatch positions (Appendix 3).

3.3.2 Phylogenetic analysis

The assembled chloroplast sequences and chloroplast sequences were analysed using Geneious V 9.1.3 software (BioMatters, USA). Sequences were aligned using the plugin MAFFT (Katoh et al., 2002). The alignment file was inspected physically. Maximum Likelihood ML, Maximum Parsimony MP using , MrBayes (Huelsenbeck and Ronquist, 2001), PHYLM (Carbonell-Caballero et al., 2015; Guindon and Gascuel, 2003), Fast Tree(Price et al., 2009), RAxML (Stamatakis, 2006), Garli(Gutell and Jansen, 2006) methods were used to analyse the evolutionary relationships. (Appendix 2 Table 18)

3.3.3 Genome annotation

All chloroplast sequences were annotated using the CpGAVAS website (http://www.herbalgenomics.org/0506/cpgavas/analyzer/home) with the default parameters. The outcome was imported directly to Geneious software for comparison with the reference *O. sativa japonica* NC_001320 to obtain the functional nucleotide polymorphisms (FNPs). Thereafter one chloroplast sequence was used to draw the chloroplast map using OGDraw v1.2 (Lohse et al., 2013). Manual editing was used to identify the polymorphic genes in all chloroplast in this study.

3.4 Results

3.4.1 Raw data

The available raw sequence data was first assessed to identify samples with good genome coverage. Only 79 of 446 samples (17%) that had a whole genome coverage at or above 0.9 X (0.9 - 7.0 X) were selected for analysis. These 79 samples were randomly distributed and covered a wide area that was divided into five major zones in Asia. Some samples were located very close to others so further selection was used to obtain 6-9 samples per zone with whole genome coverage ≥ 0.9X (Table 6).

3.4.2 Chloroplast assembly

A well-developed dual pipe line (Appendix 3) was used for chloroplast assembly. High quality of 40 new chloroplast sequence (31 wild rice *O. rufipogon* and 9 of domesticated rice was achieved. Mapping reads to a reference and de novo assembly procedures were the core of this pipe line, allowing successful assembly of all major regions of the chloroplasts, large single copy and inverted repeat A and small single copy and inverted repeat B. The output of the analyses was subjected to additional steps which further reduced errors significantly to limit manual correction. The sequence coverage was the limiting factor preventing some samples passing through this pipeline (Table 6 and Table 7) some accessions could not be resolved and failed to deliver a complete consensus sequence because of low coverage and gaps in some regions, although the whole genome coverage of W3091 and W2331 was around 2X.

Twelve samples had no differences between the two pipe lines while 21 samples had just 1-3 differences and 7 had 4-7 differences (Appendix 2 Table 18). Finally, manual inspection was used to check all the gaps and differences to identify the correct call. Some of these differences were found to be due to low coverage and assembly errors and some were real differences compared to the reference. The average coverage of the whole chloroplast for all 40 accessions was 775 X. Five accessions had no coverage for some regions based on the mapping procedure, however the *de novo* procedure had enough coverage to resolve them through manual inspection of the mapped reads (Table 7).

3.4.3 Chloroplast alignment

Fifty nine chloroplast genomes were aligned in (Geneious software V 9.1.3) using MAFFT plugin tool. The Alignment sequence was 135702 bp. The number of identical sites was (97.6 %) while the number of variable sites was (2.4%). The minimum and maximum lengths were 134116bp and

Table 6 Geographic origin of wild rice *O. rufipogon* accessions. The location (latitude and longitude) of collection, ecotype, sequence coverage (whole genome basis) and total number of sequence reads are provided for each accession.

Zone	Accession ID	Original producing area	Latitude	Longitude	Ecotype according to (Huang et al., 2012)	whole genome	
						Sequencing coverage	Total reads
Z1 IND	W1743	India	26.92	75.82	Or-I	1.09	3,839,420
Z1 IND	W1998	India	22.2	73.2	Or-III	2.24	7,875,088
Z1 IND	W1782	India	12.31	76.64	Or-III	3.48	12,259,790
Z1 IND	W1777	India	19.95	79.3	Or-III	4.84	17,025,200
Z1 IND	W1683	India	20.1	84.48	Or-II	6.75	23,695,210
Z1 IND	W2066	Nepal	28.6	81.6	Or-III	1.66	5,845,360
Z1 IND	W1804	Sri Lanka	6.93	79.95	Or-II	3.99	14,037,346
Z2 InB	W0634	Burma	25.38	97.39	Or-II	1.13	3,979,158
Z2 InB	W0628	Burma	20.4	92.85	Or-II	2.31	8,113,788
Z2 InB	W1083	India	27	88.4	Or-I	1.37	4,853,498
Z2 InB	W0153	India	22.4	88.66	Or-III	2.54	8,927,010
Z2 InB	W1126	India	24.86	92.36	Or-II	2.85	9,991,228
Z2 InB	W1096	India	26.2	92.94	Or-II	4.84	16,981,922
Z3 CHI	W3085	China	23.6	102.01	Or-III	1.18	4,133,106
Z3 CHI	W3091	China	26.8	113.55	Or-II	1.81	6,346,610
Z3 CHI	W3002	China	22.19	112.31	Or-III	2.95	10,342,360
Z3 CHI	W3052	China	23.73	106.91	Or-III	3.73	15,348,634
Z3 CHI	W3065	China	19.25	110.46	Or-III	4.02	16,574,456
Z3 CHI	W2331	Vietnam.	21.03	105.85	Or-I	2.1	7,390,804
Z4 TCV	W0626	Burma	19.77	96.11	Or-I	2.03	7,170,788

Z4 TCV	W2308	Laos	17.57	102.38	Or-II	2.47	8,692,358
Z4 TCV	W1939	Thailand	8.54	99.73	Or-II	1.61	5,610,368
Z4 TCV	W1554	Thailand	15.09	99.99	Or-II	3.1	10,883,160
Z4 TCV	W1870	Thailand	15.23	102.5	Or-II	4.18	14,547,246
Z4 TCV	W1854	Thailand	19.64	99.52	Or-II	4.73	16,484,594
Z4 TCV	W2316	Vietnam.	10.39	107.02	Or-I	3.75	13,193,968
Z5 OCE	W1236	New Papua Guinea	-5.31	141.61	Or-II	0.91	3,200,720
Z5 OCE	W1230	New Papua Guinea	-4.63	138.93	Or-I	0.97	3,426,114
Z5 OCE	W2078	Australia	-14.3	132.4	Or-III	1.18	4,163,924
Z5 OCE	W2108	Australia	-13.07	142.07	Or-III	2.22	7,803,528
Z5 OCE	W1975	Indonesia	-2.99	104.76	Or-II	2.74	9,650,252
Z5 OCE	W1977	Indonesia	-6.4	106.82	Or-II	3.98	13,998,776
Z5 OCE	W2024	Indonesia	3.29	117	Or-II	4.38	15,418,262
Z5 OCE	W0576	Malaysia	5.8	102.38	Or-II	3.69	12,940,976
Z5 OCE	W1214	Philippine	7.86	124.86	Or-III	2.92	10,232,274
Z3 CHI	HP483_*indica*	China	28.30	109.71	Domesticated	2.76	12,466,512
Z3 CHI	HP179_*indica*	China	27.68	120.55	Domesticated	3.01	13,623,130
Z3 CHI	HP49_temperate_japonica	China	33.55	109.91	Domesticated	2.15	9,718,138
Z3 CHI	HP46_temperate_japonica	China	26.89	109.20	Domesticated	0.55	2,475,544
	GP715_aus	Bengal	NA	NA	Domesticated	1.81	8,194,072
	GP706_tropical_japonica	Ivory Coast	NA	NA	Domesticated	2.21	9,997,788
	GP294_aromatic	Pakistan	NA	NA	Domesticated	2.75	12,419,702
	GP285_aus	Pakistan	NA	NA	Domesticated	2.27	10,253,590
	GP284_aromatic	Pakistan	NA	NA	Domesticated	2.64	11,923,052
	GP629_tropical_japonica	Indonesia	NA	NA	Domesticated	2.04	9,231,004

134911bp respectively. All characters weighed equally and 134573 characters were constant. Parsimony-informative characters were 308 and parsimony-uninformative characters were 821.

The number of differences between the reference O. *sativa* subsp. *japonica* Nipponbare GU592207.1 and wild accessions totalled 4975. These differences were distributed between deletions, tandem repeat deletions, insertions, tandem repeat insertions, single nucleotide polymorphism transitions, single nucleotide polymorphism transversions and substitutions with the number of differences reflecting the genetic distance among the species (Table 8)

3.4.4 Phylogenetic analysis

Five software tools were used to analyse the sequences using Maximum likelihood, Maximum Parsimony and Bayesian approaches. All analyses gave identical phylogenetic trees in regard to the main clades and sub clades. However there were some minor differences at the end of some clades or a lack of resolution (Appendix 2 Table 19).

The phylogeny of the fifty nine accessions (Figure 5) followed largely their geographical distribution (Appendix 2, Figure 23-25). *O. glumipatula* (South America) and *O. longistaminata* (Africa) were the first distinct group within the A genome species, this clade was reported by (Kim et al., 2015; Wambugu et al., 2015). The Australian clade including *O. meridionalis* and other accessions from Australia and one from further north in the Philippines was the next distinct clade identified.

The rest of the accessions divided into two main clades, the African rice species, *O. barthii* and O. *glaberrima* and the Asian species. The Asian accessions divided into two big clades: an *indica* group including *O. sativa* subsp. *indica*, *O. nivara* group, *aus* (GP-285) as one clade, and a *japonica* group including O. *sativa* subsp. Japonica, aromatic rices (GP-284, GP-294), temperate *japonica* (HP-46 and HP-49), tropical *japonica* (GP-706) and *aus* (GP-715) as the second big clade. The *indica* grouping could be further divided into two clades with *O. nivara* in a distinct grouping. The geographical distributions of the Asian clades were overlapping. However, accessions in the *O. sativa japonica* sub clade extended further west into India, while the *O. sativa* subsp. *indica* were more abundant further to the south and east. (Civáň et al., 2015; Garris et al., 2005; Kim et al., 2014b; Tong et al., 2015; Tong et al., 2016)

Table 7 Chloroplast sequence analysis by mapping, the chloroplast coverage and the number of gaps following mapping is given for each accession.

Zone	Accession ID	Chloroplast genome (based on mapping procedure)				Final chloroplast obtained
		Minimum coverage	Maximum coverage	Average coverage	Number of gaps \regions with no coverage	
Z1 IND	W1743	0	867	338.07	1	No
Z1 IND	W1998	11	2488	1154.15	0	Yes
Z1 IND	W1782	21	2508	1624.43	0	Yes
Z1 IND	W1777	51	2649	1641.65	0	Yes
Z1 IND	W1683	110	4984	2885.86	0	Yes
Z1 IND	W2066	20	808	407.07	0	Yes
Z1 IND	W1804	119	2505	1514.29	0	Yes
Z2 InB	W0634	1	1434	491.92	0	Yes
Z2 InB	W0628	34	1582	786.89	0	Yes
Z2 InB	W1083	30	686	369.88	0	Yes
Z2 InB	W0153	31	1898	1003.02	0	Yes
Z2 InB	W1126	28	1362	819.7	0	Yes
Z2 InB	W1096	46	3350	1738	0	Yes
Z3 CHI	W3085	8	376	241.04	0	Yes
Z3 CHI	W3091	0	618	397.98	7	No
Z3 CHI	W3002	4	327	160.72	0	Yes
Z3 CHI	W3052	16	772	436.59	0	Yes
Z3 CHI	W3065	23	486	301.37	0	Yes
Z3 CHI	W2331	0	1794	393.66	1	No
Z4 TCV	W0626	0	1324	440.71	1	Yes
Z4 TCV	W2308	2	1751	766.29	0	Yes

Z4 TCV	W1939	68	823	452.87	0	Yes
Z4 TCV	W1554	0	4081	2099.6	1	Yes
Z4 TCV	W1870	261	2260	1508.46	0	Yes
Z4 TCV	W1854	32	2936	1397.8	0	Yes
Z4 TCV	W2316	106	1587	992.9	0	Yes
Z5 OCE	W1236	0	720	289.63	1	No
Z5 OCE	W1230	4	934	368.03	0	Yes
Z5 OCE	W2078	88	960	503.22	0	Yes
Z5 OCE	W2108	162	1806	966	0	Yes
Z5 OCE	W1975	49	1341	762.79	0	Yes
Z5 OCE	W1977	58	1991	1242.43	0	Yes
Z5 OCE	W2024	0	2008	1033.65	21	Yes
Z5 OCE	W0576	57	1821	1182.77	0	Yes
Z5 OCE	W1214	73	1652	947.93	0	Yes
Z3 CHI	HP483_*indica*	0	997	500.37	1	Yes
Z3 CHI	HP179_*indica*	2	1160	574.11	0	Yes
Z3 CHI	HP49_temperate_japonica	42	1114	463.69	0	Yes
Z3 CHI	HP46_temperate_japonica	4	167	80.54	0	Yes
-	GP715_aus	12	413	270.56	0	Yes
-	GP706_tropical_japonica	39	301	177.07	0	Yes
-	GP294_aromatic	28	849	461.06	0	Yes
-	GP285_aus	0	337	174.87	1	Yes
-	GP284_aromatic	20	772	418.43	0	Yes
-	GP629_tropical_japonica	0	415	98.05	1	No

Table 8 Variants among AA chloroplast genomes. Deletion, Insertions SNPs when compared with O. sativa subsp. *japonica* Nipponbare GU592207.1 Del: deletion, Del.T.R. : deletion tandem repeat, Ins.: insertion, Ins.T.R. : insertion tandem repeat, SNP Tr.: SNP tran transition, SNP Trv. :SNP transversion and Subs. : substitution.

No.	Name/ code and origin	Deletions	Del.T.R.	Insertions	Ins. T.R.	SNP Tr.	SNP Trv.	Subs.	Total variation	Density /Kb	Base pair
1	Australian taxa A	10	7	4	10	47	43	4	125	0.929	134557
2	Australian taxa B	6	7	6	11	53	50	2	135	1.003	134557
3	*O.barthii1*	4	7	11	10	33	30	5	100	0.743	134674
4	*O.barthii2*	4	7	6	11	33	30	7	98	0.728	134603
5	*O.barthii3*	4	7	6	8	35	32	4	96	0.713	134596
6	*O.barthii4*	5	7	8	8	35	33	6	102	0.758	134640
7	*O.glaberrima*	4	7	6	9	33	30	7	96	0.713	134606
8	*O.glumipatula*	7	10	9	8	62	41	4	141	1.048	134583
9	*O.longistaminata1*	8	9	9	8	68	36	3	141	1.048	134567
10	*O.longistaminata2*	8	10	9	8	59	39	3	136	1.011	134563
11	*O.meridionalis*	6	6	4	14	45	44	3	122	0.907	134558
12	*O.nivara*	6	11	6	7	35	28	10	103	0.766	134494
13	*O.officinalis*	25	33	35	35	317	201	24	670	4.966	134911
14	*O.rufipogon* Asian1	3	6	0	5	18	17	6	55	0.409	134537
15	*O.rufipogon* Asian2	3	13	2	6	28	25	3	80	0.595	134544
16	*O.sativa.indicaJN861109.1*	9	37	5	4	25	19	8	107	0.796	134448
17	*O.sativa.indicaNC_008155.1*	7	8	5	6	26	18	6	76	0.565	134496
18	*O.sativa.japonicaNC_001320.1*	36	18	34	15	22	32	15	172	1.279	134525
19	W0153 Z2 India	6	7	4	8	28	21	9	83	0.617	134484
20	W0576 Z5 Malaysia	7	8	5	7	25	18	8	78	0.58	134502
21	W0626 Z4 Burma	6	8	4	7	25	26	9	85	0.632	134456

22	W0628 Z2 Burma	3	5	2	5	14	12	2	43	0.32	134583
23	W0634 Z2 Burma	6	8	5	6	24	23	7	79	0.587	134511
24	W1083 Z2 India	1	3	0	5	6	2	3	20	0.149	134537
25	W1096 Z2 India	1	3	0	4	6	2	2	18	0.134	134536
26	W1126 Z2 India	7	8	3	7	24	17	8	74	0.55	134494
27	W1214 Z5 Philippine	12	7	6	11	50	44	2	132	0.981	134549
28	W1230 Z5 Papua New Guinea	6	7	6	8	24	23	8	82	0.61	134521
29	W1554 Z4 Thailand	7	7	4	6	25	17	8	74	0.55	134495
30	W1683 Z1 India	1	3	0	4	6	2	3	19	0.141	134536
31	W1777 Z1 India	1	3	0	4	6	2	3	19	0.141	134536
32	W1782 Z1 India	3	7	3	3	18	16	5	55	0.409	134595
33	W1804 Z1 Sri Lanka	3	5	2	4	14	13	3	44	0.327	134582
34	W1854 Z4 Thailand	7	2	1	5	6	4	4	29	0.216	134116
35	W1870 Z4 Thailand	6	8	5	8	25	24	10	86	0.639	134516
36	W1939 Z4 Thailand	7	8	4	6	24	17	7	73	0.543	134494
37	W1975 Z5 Indonesia	7	8	3	7	24	17	7	73	0.543	134495
38	W1977 Z5 Indonesia	7	9	3	7	36	27	7	96	0.714	134508
39	W1998 Z1 India	3	8	3	3	15	16	4	52	0.386	134595
40	W2024 Z5 Indonesia	7	8	4	7	24	17	8	75	0.558	134520
41	W2066 Z1 Nepal	6	7	8	6	28	24	8	87	0.647	134542
42	W2078 Z5 Australia	10	7	6	12	44	45	3	127	0.944	134553
43	W2108 Z5 Australia	12	7	4	11	48	42	4	128	0.951	134542
44	W2308 Z4 Laos	2	2	1	4	5	4	4	22	0.164	134553
45	W2316 Z4 Vietnam	2	4	0	0	3	3	2	14	0.104	134556
46	W3002 Z3 China	7	7	4	7	23	18	7	73	0.543	134501

47	W3052 Z3 China	6	8	5	9	26	25	8	87	0.647	134516
48	W3065 Z3 China	9	6	6	5	32	24	9	91	0.676	134539
49	W3085 Z3 China	6	8	5	9	27	25	10	90	0.669	134517
50	HP483_*indica*	7	8	5	7	25	18	8	78	0.58	134502
51	HP179_*indica*	7	6	3	7	25	18	7	73	0.543	134496
52	HP49 temperate japonica	0	0	0	0	1	0	0	1	0.007	134551
53	HP46 temperate japonica	0	1	0	2	1	0	1	5	0.037	134553
54	GP715 aus	1	5	0	4	7	2	4	23	0.171	134534
55	GP706 tropical japonica	0	3	0	5	3	1	1	13	0.097	134556
56	GP294 aromatic	1	4	0	4	8	2	3	22	0.164	134532
57	GP285 aus	6	7	4	7	26	18	7	75	0.557	134540
58	GP284 aromatic	1	4	0	4	8	2	3	22	0.164	134532
59	Total differences	352	444	283	418	1497	1763	1379	4975		

3.4.5 SNPs and FNPs variation

Further analysis was preformed based on grouping the accessions within the main clades. The total number of variations relative to the reference *O. sativa. japonica NC_001320.1* were 4975 in total with 3478 SNPs and 1497 InDels. The clade related to *O. nivara* had the highest number of SNPs (704) and InDels. (318), while the *indica* related clade had the second highest number of variants, in total (769). However, these numbers are completely changed when we look at the variants per accession in the clade to overcome the effect of sample size in each clade. This clearly shows that the lowest variation per accession was in the *japonica* related clades (36 per accession) while the highest were in the South American and Australian related clades at 139 and 128 respectively. (Table 9).

A total of 80 genes were annotated in the 40 chloroplasts with 13 of them having functional variations (Figure 6) (*atpB, atpI, ccsA, cemA, clpP, matK, ndhF, ndhK, psaA, psbB, rpoC1*, *rpoC2,* and *rps18*). The total number of functional nucleotide polymorphisms (FNPs) in all chloroplasts was 36 and 12 of them were found to be common in all accessions (6 genes) 4 FNPs in *psaA*, , 2 in *psbB*, one in *clpP* , *ndhK*, *atpB*, *rps18*, and 2 in hypothetical protein (Table 10). The number of FNPs varied from 12 to 19. The lowest FNPs/SNP proportion was 12.0 % in W2078 Z5 Australia, while the highest was 20.7% in W1998 Z1 India. There were no unique FNPs in 13 accessions, while the highest number of unique FNPs was 7 with the proportion of unique FNPs at 37% in W1214 Z5 Philippine (Table 11).

We found around 265 (SNPs / InDels) that could be used as markers to discriminate at the clade level. These could be used to screen wild accessions to identify novel genetic resources for rice breeding and track the evolutionary relationships of the wild accessions (Appendix 2 Table 20 and Table 21).

3.5 4. Discussion

This analysis of the complete sequence of the 40 new chloroplast genomes of wild and domesticated rice population contributes to our understanding of the evolutionary relationships in *Oryza* species and will facilitate better use of wild rice in rice breeding (Daniell et al., 2016; Matsuoka et al., 2002; Tang et al., 2004). The well-developed assembly pipeline used in this study was critical in efficiently obtaining an accurate whole chloroplast genome sequence

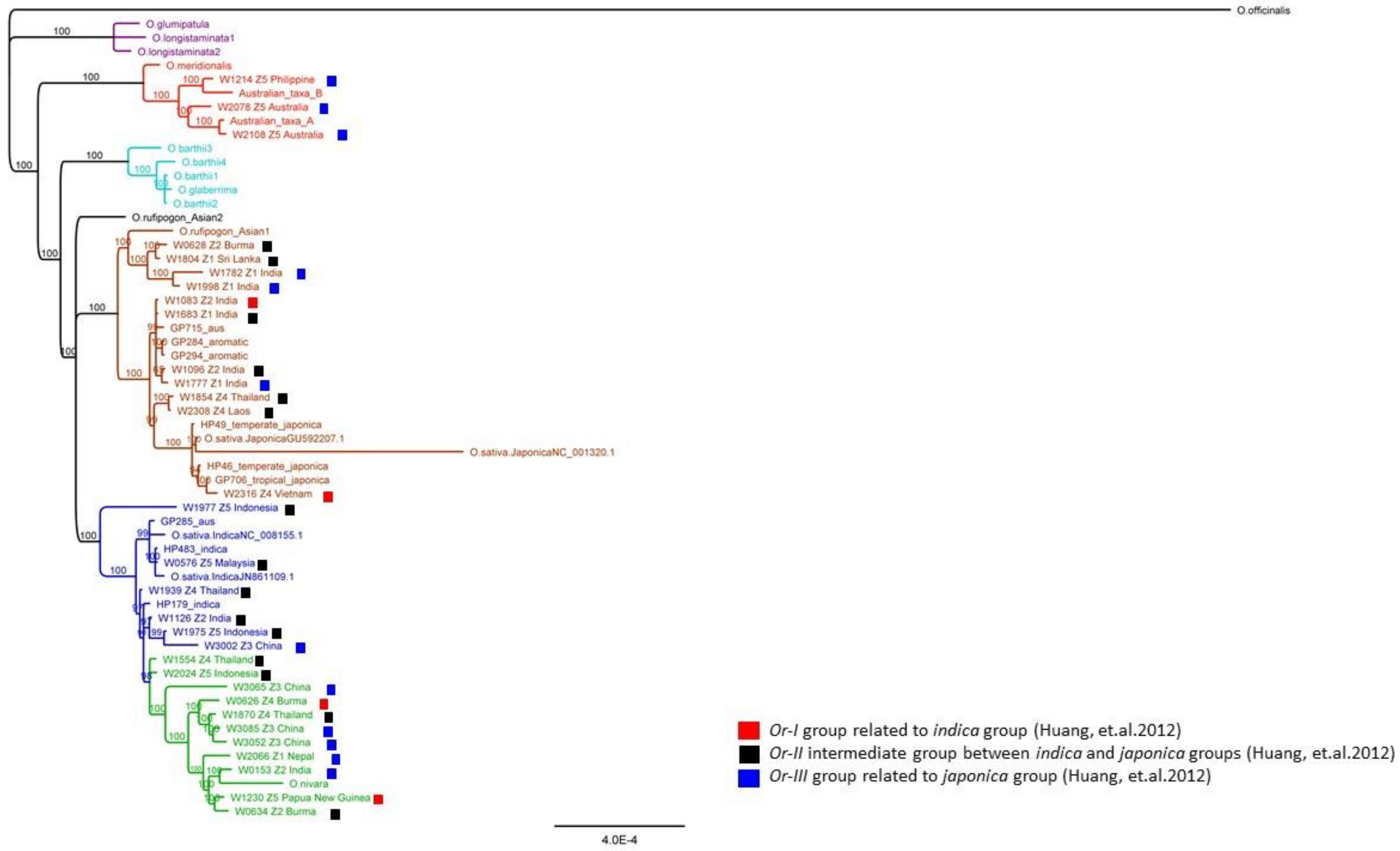


Figure 5 Phylogenetic relationship of *Oryza* chloroplast AA Genome. Analysis using MrBayes GTR model with 2000 bootstraps and O. australiensis as an out group. Numbers on branches refer to probability percentage.

despite variable coverage. The complementation of the two procedure (mapping reads to reference and *de novo* assembly) eliminates many errors which might have been considered as a real differences in the past. The geographically separated African and South American wild rices were found to be genetically distinct from the Asian domesticated rice, and Asian/Australian wild rice (Figure 5 and Figure 7) in agreement with earlier studies based on fewer samples (Brozynska et al., 2017; Wambugu et al., 2015).

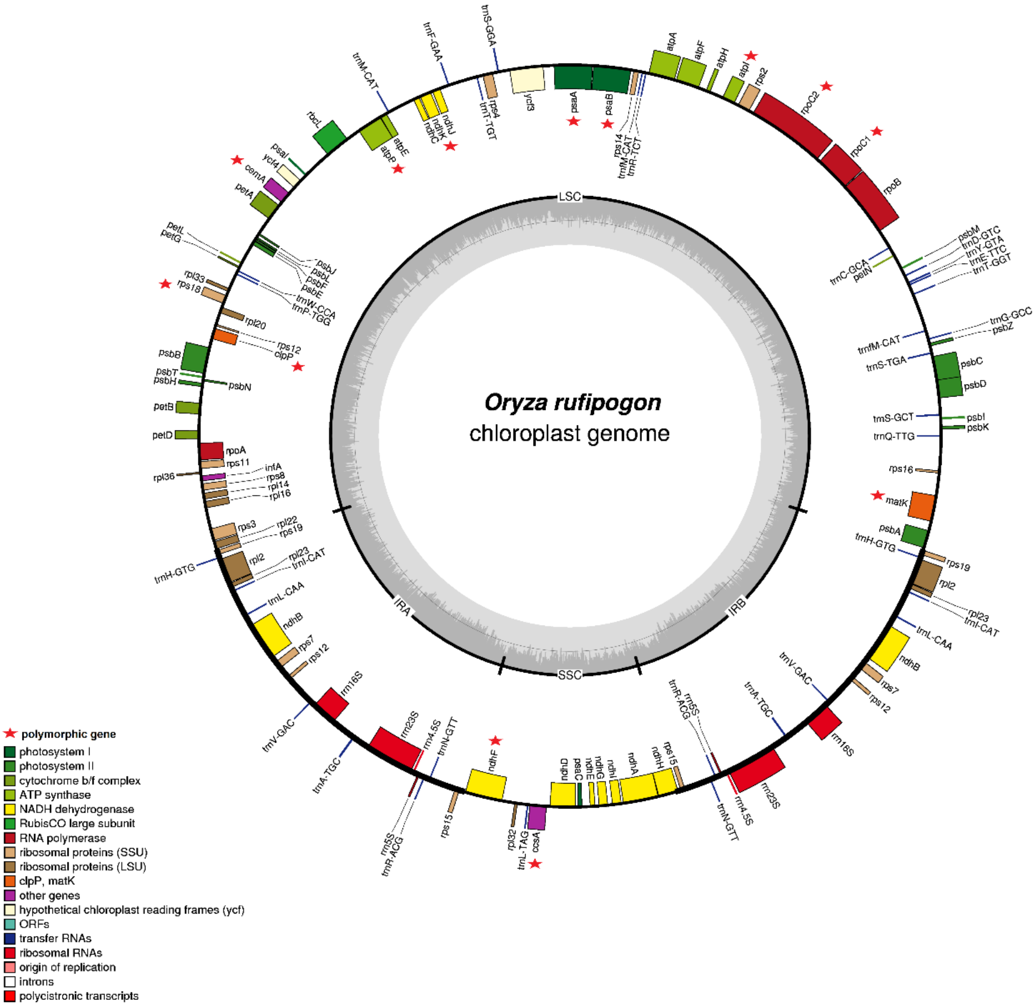
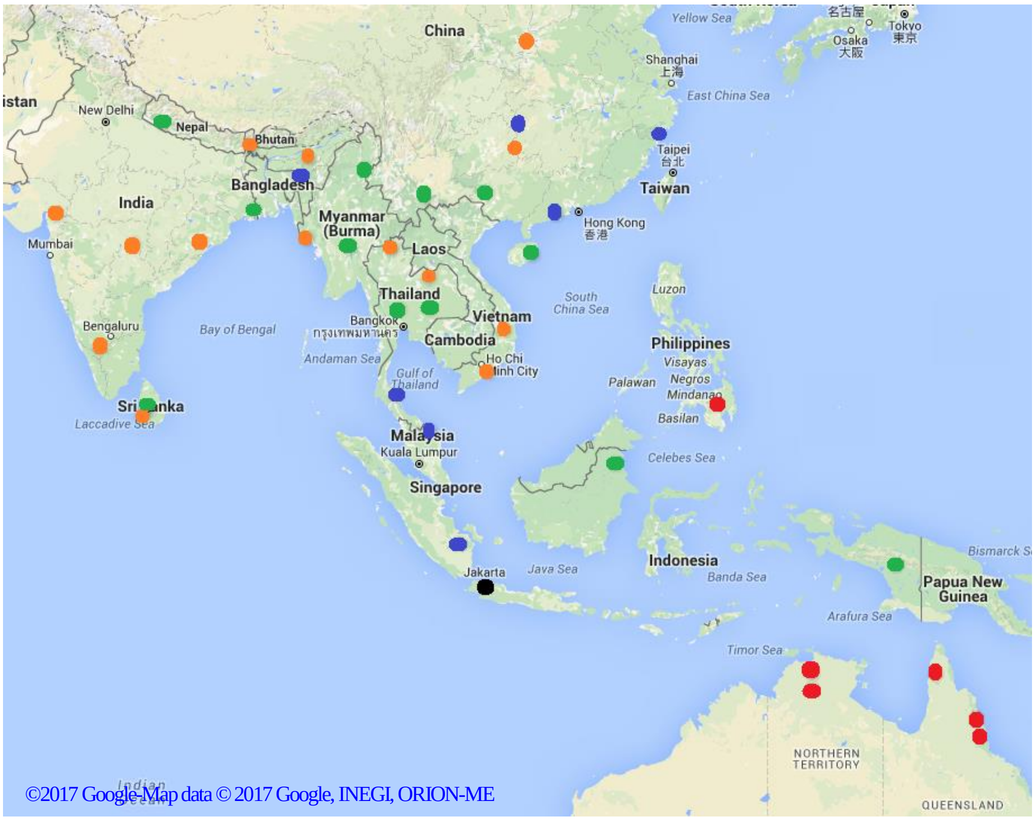


Figure 6 Chloroplast gene map. Polymorphic genes are marked with *. The inner circle represents the four chloroplast regions LSC, IRB, SSC and IRA. The GC content is shown in the grey area



©2017 Google-Map data © 2017 Google, INEGI, ORION-ME

Figure 7 Phylogeographic distribution of diversity in *Oryza* spp. AA chloroplast genomes. The *O. sativa* spp. *indic*a and *O. nivara* clade group are represented by blue and green dots respectively. The yellow dots represent the clade related to *O. sativa subsp.* Japonica. The Australian clade is marked with red dots. The black dot represents W1977 which was an out group relative to the two sub clades including *O. nivara* and *O. sativa* subsp. *indica*. *Asian and Australian accession positions were based on collection site GPS locations. Map sourced from Google maps.

Table 9 Polymorphisms between the clades defined by the chloroplast phylogeny SNPs, InDel and Deletions between the clades as defined in Figure 4

	1	Ratio*	2	Ratio*	3	Ratio*	4	Ratio*	5	Ratio*	6	Ratio*	7	Ratio*	Total 58 accessions	Ratio per accession in Total
SNP	423	21.15	568	51.64	573	95.5	704	58.67	353	70.6	315	105	542	542	3478	59.97
InDel	305	15.25	308	28	196	32.67	318	26.5	139	27.8	103	34.33	128	128	1497	25.81
Total	728	36.4	876	79.64	769	128.17	1022	85.17	492	98.4	418	139.33	670	670	4975	85.78

1-*japonica* clade (20 accessions) 2-*indica* clade (11 accessions) 3-Australian clade (6 accessions) 4-Nivara clade (12 accessions) 5-African clade (5 accessions) 6-South American clade (3 accessions) 7-O. officinalis (1 accessions) * Ratio per accession

Table 10 Functional variation in chloroplast genome sequences. FNPs location, amino acid substitution, codon changed and polymorphism type.

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
1	2,603	matK	maturase K	NP_039361.2	I -> F	matK CDS	201	601	1	T -> A	ATT -> TTT	SNP (transversion)	Substitution
2	8,415		hypothetical protein	NP_039365.1	R -> S	hypothetical protein CDS	23	67	1	C -> A	CGC -> AGC	SNP (transversion)	Substitution
3	8,538		hypothetical protein	NP_039365.1	L -> V	hypothetical protein CDS	64	190	1	C -> G	CTT -> GTT	SNP (transversion)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
4	8,599		hypothetical protein	NP_039365.1	G -> E	hypothetical protein CDS	84	251	2	G -> A	GGG -> GAG	SNP (transition)	Substitution
5	8,622		hypothetical protein	NP_039365.1	S -> P	hypothetical protein CDS	92	274	1	T -> C	TCC -> CCC	SNP (transition)	Substitution
6	22,488	rpoC1	RNA polymerase beta' subunit	NP_039374.1	Q -> E	rpoC1 CDS	4	10	1	C -> G	CAA -> GAA	SNP (transversion)	Substitution
7	24,178	rpoC1	RNA polymerase beta' subunit	NP_039374.1	N -> S	rpoC1 CDS	567	1,700	2	A -> G	AAT -> AGT	SNP (transition)	Substitution
8	24,756	rpoC2	RNA polymerase beta" subunit	NP_039375.1	Q -> H	rpoC2 CDS	10	30	3	G -> T	CAG -> CAT	SNP (transversion)	Substitution
9	25,379	rpoC2	RNA polymerase beta" subunit	NP_039375.1	R -> K	rpoC2 CDS	218	653	2	G -> A	AGA -> AAA	SNP (transition)	Substitution
10	25,835	rpoC2	RNA polymerase beta" subunit	NP_039375.1	D -> G	rpoC2 CDS	370	1,109	2	A -> G	GAT -> GGT	SNP (transition)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
11	25,897	rpoC2	RNA polymerase beta" subunit	NP_039375.1	H -> D	rpoC2 CDS	391	1,171	1	C -> G	CAT -> GAT	SNP (transversion)	Substitution
12	26,188	rpoC2	RNA polymerase beta" subunit	NP_039375.1	R -> G	rpoC2 CDS	488	1,462	1	A -> G	AGA -> GGA	SNP (transition)	Substitution
13	28,019	rpoC2	RNA polymerase beta" subunit	NP_039375.1	W -> L	rpoC2 CDS	1,098	3,293	2	G -> T	TGG -> TTG	SNP (transversion)	Substitution
14	28,336	rpoC2	RNA polymerase beta" subunit	NP_039375.1	C -> G	rpoC2 CDS	1,204	3,610	1	T -> G	TGT -> GGT	SNP (transversion)	Substitution
15	29,113	rpoC2	RNA polymerase beta" subunit	NP_039375.1	N -> D	rpoC2 CDS	1,463	4,387	1	A -> G	AAC -> GAC	SNP (transition)	Substitution
16	30,548	atpI	ATP synthase CF0 A subunit	NP_039377.1	D -> E	atpI CDS	16	48	3	T -> G	GAT -> GAG	SNP (transversion)	Substitution
17	40,251	psaA	photosystem I P700	NP_039383.1	R -> G	psaA CDS	334	1,000	1	G -> C	CGC -> GGC	SNP (transversion)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
			chlorophyll a apoprotein A1										
18	40,482	psaA	photosystem I P700 chlorophyll a apoprotein A1	NP_039383.1	R -> G	psaA CDS	257	769	1	G -> C	CGA -> GGA	SNP (transversion)	Substitution
19	40,684	psaA	photosystem I P700 chlorophyll a apoprotein A1	NP_039383.1	H -> Q	psaA CDS	189	567	3	A -> T	CAT -> CAA	SNP (transversion)	Substitution
20	40,839	psaA	photosystem I P700 chlorophyll a apoprotein A1	NP_039383.1	S -> T	psaA CDS	138	412	1	A -> T	TCC -> ACC	SNP (transversion)	Substitution
21	49,212	ndhK	NADH dehydrogenase subunit K	NP_039387.2	R -> T	ndhK CDS	12	35	2	C -> G	AGA -> ACA	SNP (transversion)	Substitution
22	53,201	atpB	ATP synthase CF1 beta subunit	NP_039390.1	R -> P	atpB CDS	37	110	2	C -> G	CGG -> CCG	SNP (transversion)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
23	56,134		hypothetical protein	NP_039393.1	N -> K	hypothetical protein CDS	59	177	3	C -> G	AAC -> AAG	SNP (transversion)	Substitution
24	56,770		acetyl-CoA carboxylase beta subunit	NP_039394.1	S -> C	acetyl-CoA carboxylase beta subunit CDS	73	218	2	C -> G	TCC -> TGC	SNP (transversion)	Substitution
25	56,776		acetyl-CoA carboxylase beta subunit	NP_039394.1	Q -> L	acetyl-CoA carboxylase beta subunit CDS	75	224	2	A -> T	CAG -> CTG	SNP (transversion)	Substitution
26	59,000	cemA	envelope membrane protein	NP_039398.1	L -> F	cemA CDS	108	324	3	G -> T	TTG -> TTT	SNP (transversion)	Substitution
27	66,104	rps18	ribosomal protein S18	NP_039408.1	T -> N	rps18 CDS	155	464	2	C -> A	ACC -> AAC	SNP (transversion)	Substitution
28	67,982	clpP	ATP-dependent Clp protease	NP_039410.1	P -> A	clpP CDS	103	307	1	G -> C	CCG -> GCG	SNP (transversion)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
			proteolytic subunit										
29	69,349	psbB	photosystem II 47 kDa protein	NP_039411.1	A -> V	psbB CDS	184	551	2	C -> T	GCG -> GTG	SNP (transition)	Substitution
30	70,278	psbB	photosystem II 47 kDa protein	NP_039411.1	A -> T	psbB CDS	494	1,480	1	G -> A	GCA -> ACA	SNP (transition)	Substitution
31	70,281	psbB	photosystem II 47 kDa protein	NP_039411.1	I -> F	psbB CDS	495	1,483	1	A -> T	ATC -> TTC	SNP (transversion)	Substitution
32	84,369		hypothetical protein	NP_039431.1	Q -> E	hypothetical protein CDS	125	373	1	C -> G	CAA -> GAA	SNP (transversion)	Substitution
33	102,760	ndhF	NADH dehydrogenase subunit 5	NP_039441.1	F -> C	ndhF CDS	293	878	2	A -> C	TTC -> TGC	SNP (transversion)	Substitution
34	105,906	ccsA	cytochrome c biogenesis protein	NP_039443.1	Y -> S	ccsA CDS	224	671	2	A -> C	TAT -> TCT	SNP (transversion)	Substitution

	Sequence location	Gene	Gene product	Protein ID	AA Change	CDS	CDS Codon number	CDS position	CDS Position within codon	Change	Codon change	Polymorphism type	Effect on protein
35	124,775		hypothetical protein	NP_039456.1	M -> L	hypothetical protein CDS	34	100	1	A -> C	ATG -> CTG	SNP (transversion)	Substitution
36	130,749		hypothetical protein	NP_039460.1	Q -> E	hypothetical protein CDS	125	373	1	G -> C	CAA -> GAA	SNP (transversion)	Substitution

- Blue FNPs are found in all accessions relative to the reference

Table 11 Summary of variants identified for all Asian wild rice samples analysed.

	accession	SNP	FNP	Common FNPs	unique FNPs	unique FNPs ratio
1	W0153 Z2 India	102	13	12	1	7.69
2	W0576 Z5 Malaysia	96	15	12	3	20
3	W0626 Z4 Burma	104	14	12	2	14.29
4	W0628 Z2 Burma	80	16	12	4	25
5	W0634 Z2 Burma	100	14	12	2	14.29
6	W1083 Z2 India	62	12	12	0	0
7	W1096 Z2 India	62	12	12	0	0
8	W1126 Z2 India	94	15	12	3	20
9	W1214 Z5 Philippine	148	19	12	7	36.84
10	W1230 Z5 Papua New Guinea	100	15	12	3	20
11	W1554 Z4 Thailand	95	15	12	3	20
12	W1683 Z1 India	62	12	12	0	0
13	W1777 Z1 India	62	12	12	0	0
14	W1782 Z1 India	90	18	12	6	33.33
15	W1804 Z1 Sri Lanka	81	16	12	4	25
16	W1854 Z4 Thailand	64	12	12	0	0
17	W1870 Z4 Thailand	102	14	12	2	14.29
18	W1939 Z4 Thailand	94	14	12	2	14.29
19	W1975 Z5 Indonesia	94	15	12	3	20
20	W1977 Z5 Indonesia	116	18	12	6	33.33
21	W1998 Z1 India	87	18	12	6	33.33
22	W2024 Z5 Indonesia	94	15	12	3	20
23	W2066 Z1 Nepal	105	16	12	4	25
24	W2078 Z5 Australia	142	17	12	5	29.41
25	W2108 Z5 Australia	143	18	12	6	33.33
26	W2308 Z4 Laos	63	12	12	0	0
27	W2316 Z4 Vietnam	60	12	12	0	0
28	W3002 Z3 China	94	15	12	3	20
29	W3052 Z3 China	104	14	12	2	14.29
30	W3065 Z3 China	109	15	12	3	20
31	W3085 Z3 China	105	14	12	2	14.29
32	HP483_*indica*	51	6	12	0	0
33	HP179_*indica*	50	9	12	3	20
34	HP49_temperate_japonica	1	6	12	0	0
35	HP46_temperate_japonica	2	6	12	0	0
36	GP715_aus	13	6	12	0	0
37	GP706_tropical_japonica	5	6	12	0	0
38	GP294_aromatic	13	6	12	0	0
39	GP285_aus	51	9	12	3	20
40	GP284_aromatic	13	9	12	3	20

The phylogenetic tree shows clearly that the Australian clade is distinct from all others. However, this clade extends north from Australia (to the Philippines) overlapping with an Asian clade including accessions form Papua New Guinea (Figure 4 and Figure 5). Other Australian plant species have been found to have relationships with plants in the Philippines(Simpson, 1977; Yap, 2010). The Philippines is at the boundary of regions having an Australia association or origin and those with an Asian link.

The analysis divided the Asian wild and domesticated accessions into two main clades, one related to O. *sativa* spp. *japonica* and the other to O. *sativa* spp. *indica* which in turn divided into two sub clades related to *O. sativa* spp. *indica* and *O. nivara* respectively. This supports the view that these lineages were separated some time ago (0.99 million years,(Brozynska et al., 2017; Kumagai et al., 2016; Liu et al., 2015)) and that the much more recent domestication was from distinct gene pools (Brozynska et al., 2017; Civáň et al., 2015). The overlap of the Australian and *indica* clades supports a recent phylogeny study (Brozynska et al., 2017; Fuchs et al., 2016) based on the nuclear gene analysis which shows greater introgression between the Australian wild rice and the *nivara/indica* group than between the Australian and *japonica* group. The analysis shows that chloroplast diversity is greater further south and east being higher in the clade related to *indica* and highest in Australia.

The *aromatic, tropical* and *temperate japonica* are much closer to *O. sativa japonica* which agrees with previous study apart from the discovery that *aus* appears in both clades *japonica* and *indica*. This suggest that the maternal genomes of *aus* come from two different origins.(Kumagai et al., 2016) (Cicáň et al., 2015; Kim et al., 2015; Tong et al., 2015; Tong et al., 2016)

Despite the existence of distinct clades based upon chloroplast sequence the accessions did not show a strong geographic isolation being spread widely across the south and east of Asia. Divergence may have been caused by a past period of geographic isolation creating distinct populations that became the progenitors for domestication of *indica* and *japonica* rice. These populations have now been widely distributed across the entire region in Asia with the Australian populations retaining more geographic distribution. The populations may have accumulated useful mutations in response to the selective pressure of different environments during periods of geographic separation.

The nuclear genome diversity in these wild rices does not follow the same pattern as the chloroplast genomes (Figure 4, (Cicáň et al., 2015; Huang et al., 2012)). This suggests that the evolution of the wild progenitors of domesticated rice followed a complex path probably involving many dispersal events and chloroplast capture. Interestingly the majority of the accessions in the chloroplast clade including *O. nivara* had *japonica* like nuclear genomes while the majority of the chloroplast clades related to *japonica* and *indica* were intermediate in nuclear genome (Huang et al., 2012).

The chloroplast is not just an energy factory for the cell but has an impact on intracellular signalling and may regulate the whole cells response to the surround environment. (Bobik and Burch-Smith, 2015; Daniell et al., 2016; Sun and Guo, 2016). The extent to which adaptation has shaped the evolution of these distinct chloroplast genomes is not yet clear. The 36 FNPs distributed over 13 genes (*atpB, atpI, ccsA, cemA, clpP, matK, ndhF, ndhK, psaA, psbB, rpoC1, rpoC2* and *rps18*) and hypothetical proteins could provide adaptation to specific environments. Especially as they control vital biological processes in the plant cell like ATP synthesis, envelope membrane protein, NADH dehydrogenase, photosystem I and II, ribosomal protein S18 , RNA polymerase. Any variation in these chloroplast genes may also affect nuclear gene expression and led to dramatic changes in plant performance in normal conditions or under biotic / abiotic stress (Table 10 and Figure 6). (Brozynska et al., 2015; Dal Bosco et al., 2003; Inaba and Schnell, 2008; Li, 2012; Sun and Guo, 2016; Wang et al., 2014; Xu et al., 2005a; Zheng et al., 2016). Variation in maternal genomes has been shown to have a dramatic impact on human phenotype (Wallace, 2016). Maternal genome variation in rice might also offer significant adaptation to environment. Only two chloroplast types seem to have been introduced into domestication of *japonica* and *indica* rice. The wider range of chloroplast types revealed in this study might represent an untapped resource for rice genetic improvement. Twelve of the 36 FNPs which were found to be common in all accessions (Table 10) (Appendix 2, Figure 22) may represent domestication related variation between O. *sativa japonica* NC_001320 and all these wild rices. These may have resulted from selection pressure in cultivation and may include some accumulated mutations that could be harmful in the wild and would not survive outside of the domesticated gene pool.

Rice passed through the bottle neck of the domestication process with human selection that focused on specific characters like seed shattering, uniform maturing and yield and led to loss of other important alleles which might have a role in biotic / abiotic stress resistance and adapt to environment changes. The wild FNPs identified in this study represent the original gene pool before domestication and may be useful in developing rice genotypes for cultivation in future environments (Andersson et al., 2010; Brozynska et al., 2015; Hajjar and Hodgkin, 2007; Henry, 2009; Song et al., 2005; Xu et al., 2012). Further study of these FNPs is required to determine their significance. Analysing chloroplast genomes provides a useful tool for conserving and utilizing the genetic resources in the A genome genepool of *Oryza* species and supporting food security.

Chapter 4

4 Diversity and Evolution of Rice Progenitors in Australia

4.1 Abstract

In the thousands of years of rice domestication in Asia, many useful genes have been lost from the gene pool. Wild rice is a key source of diversity for domesticated rice. Genome sequencing has suggested that the wild rice populations in northern Australia may include novel taxa, within the AA genome group of close (inter-fertile) wild relatives of domesticated rice that have evolved independently due to geographic separation and been isolated from the loss of diversity associated with gene flow from the large populations of domesticated rice in Asia. Australian wild rice was collected from 27 sites from Townsville to the northern tip of Cape York. Whole chloroplast genome sequences and 4555 nuclear gene sequences (more than 8Mbp) were used to explore genetic relationships between these populations and other wild and domesticated rices. Analysis of the chloroplast and nuclear data showed very clear evidence of distinctness from other AA genome *Oryza* species with significant divergence between Australian populations. Phylogenetic analysis suggested the Australian populations represent the earliest-branching AA genome lineages and may be critical resources for global rice food security. Nuclear genome analysis demonstrated that the diverse *O. meridionalis* populations were sister to all other AA genome taxa while the Australian *O. rufipogon*-like populations were associated with the clade that included domesticated rice. Populations of apparent hybrids between the taxa were also identified suggesting ongoing dynamic evolution of wild rice in Australia. These introgressions model events similar to those likely to have been involved in the domestication of rice.

keywords: Australian wild rice, nuclear genes, chloroplast sequence, phylogenetic analysis

4.2 Introduction

Rice (*Oryza* sativa L.) is a critically important cereal crop being a key source of carbohydrates (calories) and an important source of many other nutrients for more than half of the world’s people(Civáň et al., 2015; Huang et al., 2012). The wild relatives of rice represent a valuable resource for rice improvement and adaptation to meet the needs of a growing human population in a changing environment.(Henry, 2016; Henry et al., 2010; Mickelbart et al., 2015).

Wild *Oryza* species are widespread in northern Australia(Henry et al., 2010). This is an area without a long history of rice cultivation, implying that the wild populations have remained largely isolated from the impacts of gene flow from domesticated crops that has apparently been widespread in Asia (Brozynska et al., 2017). The AA genome species of rice include cultivated species and their close relatives(Choi et al., 2017). Draft genome sequences of the AA genome populations from Australia have recently been reported indicating that these populations may be an important genetic resource for rice because of their high diversity and phylogenetic relationship to domesticated rice(Brozynska et al., 2015; Brozynska et al., 2014b; Brozynska et al., 2017; Sotowa et al., 2013; Wambugu et al., 2015).

We now report on an analysis of the genomes of rice collected from sites over a wide area in northeastern Australia allowing analysis of the diversity and relationships within and between these wild populations.

4.3 Material and methods

4.3.1 Field collections

Samples and data were collected during May 2015, 2016 and 2017, from north eastern Queensland, Australia. Collections ranged from south of Townsville to the most northerly parts of Cape York Peninsula (Figure 8). Seeds and vegetative material were collected from 29 sites. GPS coordinates, observations of plant spike form, awn length, an herbarium voucher, and photographs of flowers (where possible) were obtained at each site (Appendix 4, Table 27,).

4.3.2 Morphological measurement

Anther and awn measurements were recorded in the field. For anther length, 4 to 8 flowers from 3 to 6 immature panicles were selected at random from each population, photographed against a standard background with a scale, and measurements obtained later in the laboratory using Image-Pro Plus software (Media Cybernetics, MD, USA, http://www.mediacy.com/index.aspx?page=IPP). The awn length was measured for ten different plants from each population selected at random.

4.3.3 DNA extraction and sequencing

Vegetative tissue from 29 samples (representing each of the collection sites) was prepared and

DNA extracted as described by Furtado (Furtado, 2014). Three approaches were used to assess the quality and quantity of the extracted DNA: Nano Drop (Thermo Fisher Scientific), agarose gel electrophoresis, and Qubit (Thermo Fisher Scientific). Multiplex sequencing of the 29 wild rice samples was conducted using a Hiseq 4000 (Illumina) using 2X 150 paired end technique, aiming to produce approximately 10 X whole genome coverage on average. Reference chloroplast genome sequences were obtained as described in (Appendix 4, Table 28).

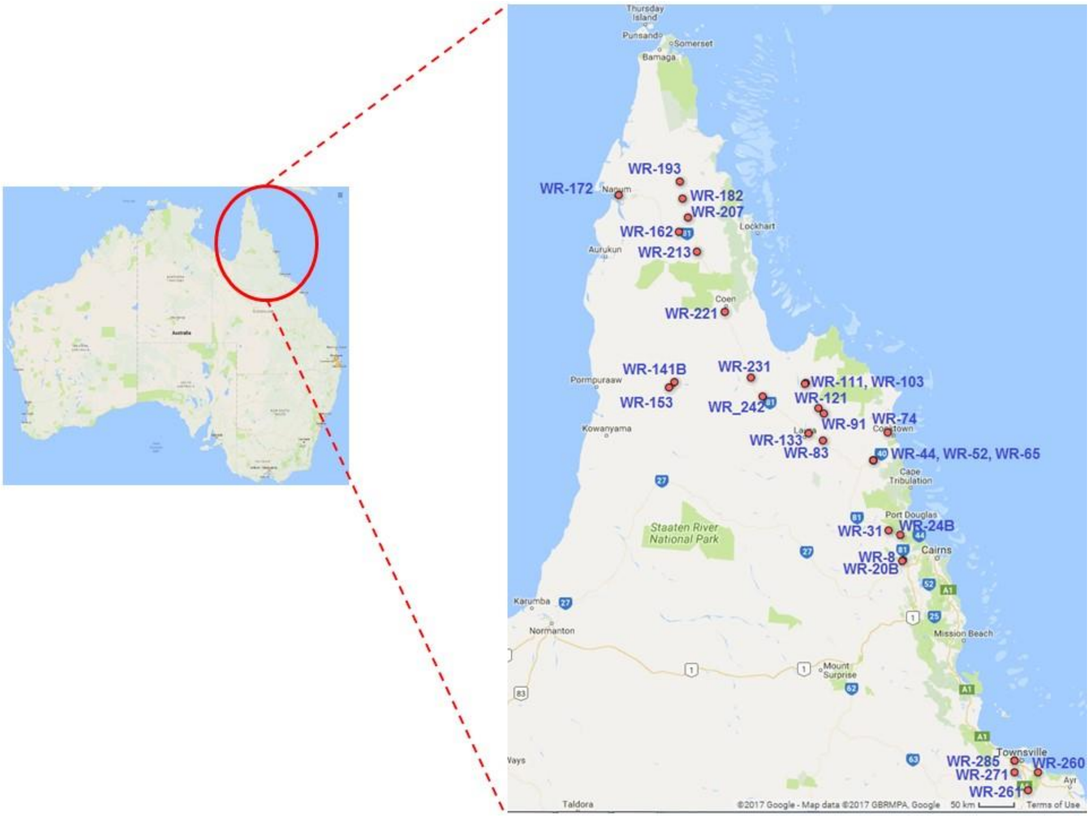


Figure 8 Australian wild rice collection sites. Red dots indicate collection sites.

4.3.4 Chloroplast genome assembly

The sequence reads were analyzed using CLC Genomic workbench V.9, Geneious V.9.1.5 and Clone Manager Professional 9, (Kim et al., 2015). A quality check (QC) was applied to all raw data. Based on the results of the QC report, reads were trimmed. A dual pipeline approach was used to assemble the chloroplast genome sequences: mapping reads to reference, and *de novo* assembly. The outputs of both pipelines were combined and all discrepancies were resolved and corrected manually.

4.3.5 Chloroplast phylogenetic analysis

The assembled chloroplast genome sequences together with those that were obtained from earlier studies (a total of 42), were analysed using Geneious V 9.1.5 (geneious.com). Chloroplast genomes were aligned using the MAFFT (MAFFT v7.308 Algorithm: auto, scoring matrix: 1PAM / k=2 gap open penalty:1.53 offset value:0.123) plugin tool (Katoh et al., 2002). The alignment file was inspected physically. Bayesian Inference (BI), Maximum Likelihood (ML), and Maximum Parsimony (MP) approaches, using the software packages MrBayes (Huelsenbeck and Ronquist, 2001), PHYLM(Carbonell-Caballero et al., 2015; Guindon and Gascuel, 2003), PAUP(Swofford, 2003) respectively were utilized to infer the evolutionary relationships. (Appendix 4, Table 32). Genetic diversity for the whole chloroplast calculated using DnaSP software (Rozas et al., 2003)

4.3.6 Chloroplast genome annotation

All chloroplast sequences were annotated using the CpGAVAS website (http://www.herbalgenomics.org/0506/cpgavas/analyzer/home), using the default parameters as recommended. The outcome was imported directly into Geneious software to allow comparison with the reference *O. sativa japonica* NC_001320 to identify polymorphisms.

4.3.7 Phylogenetic analysis of nuclear genes

Phylogenetic analysis was based upon a set of 4643 genes that were found in all include *Oryza* species (Brozynska et al., 2017). These sequences were obtained from the sequence data pool for each field sample and reference genome using the software packages FastQC, BWA, Samtools, bcftools and MUMmer. The accession identifiers of the reference samples used were: *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 and O. *punctata* BB GCA_000573905.1. A total of 4555 genes were obtained from all samples and references. These genes were divided into groups based upon the chromosomal location in *O. sativa japonica*. Multiple sequence alignment was performed at the gene level using MAFFT (Katoh et al., 2002). Following this individual gene alignment files were concatenated into single alignment for each chromosome, then all chromosomes were combined into a whole genome alignment of 8,179,015 base pairs (Figure 10 B).

Phylogenetic trees were reconstructed using three analytical approaches: maximum likelihood (ML), maximum parsimony (MP) and Bayesian inference (BI). For the ML analysis. PHYML version 20131022 was used with the following settings: Tree topology search: NNIs, Initial tree= parsimony, model of nucleotide substitution= GTR (Guindon and Gascuel, 2003). For the MP analysis PAUP 4.0 was used with the following setting: stepwise taxon addition with random seed, heuristic tree search strategy, and 1000 bootstrap (Swofford, 2003). For the BI analysis MrBayes was used with same as reported in (Brozynska et al., 2017).

4.4 Results and Discussion

Wild AA genome rice was collected from 27 sites in north Queensland, Australia (Figure 8 and Appendix 4, Table 27). Plants were found around the margins of lakes and creeks (Appendix 4,) where for the most part, water was available to support their growth. Wild rice was not located on Cape York north of the Jardine River (-11.103665, 142.283901) or on the Islands of Torres Strait, consistent with Herbarium records (AVH, accessed 30/06/2017). Although the cause of this distributional gap, and its temporal dynamics, is unclear, it may represent a contemporary barrier to gene flow with populations to the north in New Guinea and South East Asia.

Wild plants in the field showed significant morphological variation (Appendix 4, Table 27), particularly in spike morphology, awn length and anther length. Awn length varied more than 3 fold between sites with the open panicle types (*O. rufipogon*-like, Taxon A) having shorter awns than the closed panicle types (*O. meridionalis*-like, Taxon B). The shortest anthers (c. 1.5 mm) were found in plants resembling *O. meridionalis* or taxon B. In contrast, the longest anthers (4.5 mm) were found in plants resembling *O. rufipogon* or taxon A. Both awn and anther length showed highly significant ($P < 0.01$) differences between sites. The results agree with previous studies of these Australian populations. (Brozynska et al., 2014b; Sotowa et al., 2013; Waters et al., 2012).

All regions of the chloroplasts were successfully sequenced. The high sequence coverage ensured a complete genome sequence was obtained for all sites in the assembly pipeline that was used. The average coverage of the total chloroplast for all samples was 683 X while the highest and lowest coverages were 2063X and 10X respectively (Appendix 4, Table 28). Compared to the reference sequence an average of 129.6 variants (deletions, insertions, and SNPs) per sample were found (Appendix 4, Table 29), which agrees with the results reported by (Brozynska et al., 2014b). A total of 18 functional polymorphisms were found in the chloroplasts with six of them common to all samples (Appendix 4, Table 30 and Table 31).

The aligned sequence comprised 135,532 bp. Of the variable sites 227 were parsimony-informative and 661 were uninformative (427 were unique). The phylogenetic trees constructed using different approaches (Appendix 4, Table 32) were highly congruent (Brozynska et al., 2014b; Kim et al., 2015; Wambugu et al., 2015). As in earlier work (Wambugu et al., 2015), a clade including *O. glumipatula* and *O. longistaminata* was sister to all other AA genome rices which were divided into an Australian clade, and a clade with Asian and African taxa including the two domesticated species. The Australian clade contained two main clades: a small clade (7 populations) containing Taxon A and a much larger clade (20 populations) containing the majority of the samples including Taxon B and *O. meridionalis*. This result confirms that the chloroplast genome of Taxon A is not closely related to that of Asian *O. rufipogon* despite the plants having a similar appearance. Eight unique chloroplast molecular makers were found in all members of the clade that includes Taxon A (Appendix 4, Table 33) (Kim et al., 2015). The chloroplasts of the different Australian AA genome taxa showed significant genetic differences (Figure 9).The concatenated alignment of 4555 nuclear genes comprised 8,179,015 bp of which 44.1% were invariant. The minimum and maximum lengths were 5,916,081 bp and 7,013,653 bp respectively, slightly longer than reported previously (Brozynska et al., 2017). The nuclear analysis (as one full length sequence and by chromosomes) grouped the Australian samples into two main clades. One of these included Taxon A and the other much larger group (27 samples) included Taxon B and *O. meridionalis* types (Appendix 4, Table 34 and Figure 10). This analysis confirmed the nuclear genomes of the diverse *O. meridionalis* group including Taxon B are sister to those of all other AA genome taxa. However, four other Australian samples including Taxon A grouped within the clade that includes all other AA genome species as suggested by the single genome analysis (Brozynska et al., 2017). The phylogeny based upon individual chromosomes (Appendix 4, Figure 33-35) shows that these populations were a sister to all Asian and African rices (chromosomes 4,5,6,7,8) or the Asian rices (chromosome 9,10), *O. indica/O. nivara* (1,2,3,11) or Australian (12) clades indicating significant introgression between the different populations of wild rice.

The chloroplast genomes of Taxon B are diverse and include a small number (populations WR-44, WR- 52. WR-153, WR-162) that showed close relationships to the chloroplast genome found in the plants with an A genome. These included the most divergent B types (eg WR-44, WR-52 and WR-162). Some of these were from sites where morphological traits were somewhat intermediate between the Taxon A and Taxon B types. For example, the populations found on the Lakeland-Cooktown road had large anthers and panicles that varied from open to closed. The divergent B nuclear genome and A chloroplast genome suggests plants in these populations may be hybrids. Population WR-65 had a B type chloroplast but an A type nuclear genome.

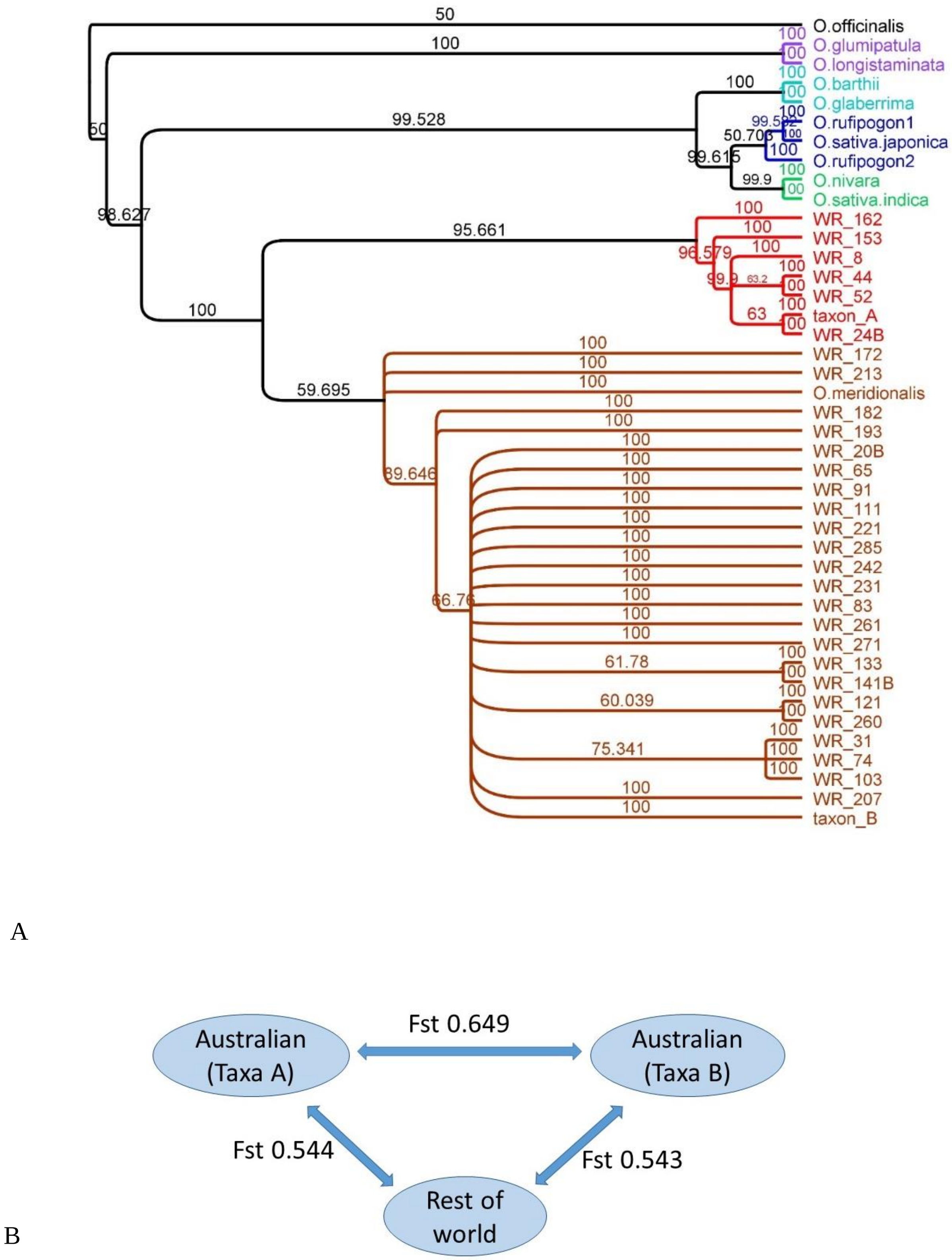


Figure 9 Diversity of chloroplast genomes A, Phylogenetic tree based on MP analysis of whole chloroplast genome sequences Colours relate to the main clades. red and brown clades are from Australia. Bootstrap values (MP 1000 replicates) are shown on the branches; B, Genetic distances between populations in Australia and elsewhere

Both chloroplast and nuclear gene analysis suggest a high diversity of AA genome wild rice in Australia. This supports the view that Australia might be a centre of diversity for the AA genome clade. The populations with a morphology similar to *O. meridionalis* are diverse and may include both annual and perennial types (Brozynska et al., 2014b; Sotowa et al., 2013). These populations could all be considered part of one diverse species, *O. meridionalis*. The nuclear genome analysis of the *O. rufipogon*-like (Taxon A) populations places them in the Asian clade together with domesticated rices. This suggests these Australian populations should be considered as a distinct, undescribed taxon (Brozynska et al., 2017). Analysis of the chloroplast genomes placed Australian plants with *O. rufipogon*-like morphology in the Australian clade, distant from the Asian *O. rufipogon* which were placed in the Asian clade. Some populations with a nuclear genome similar to *O. meridionalis* had a chloroplast genome that was closer to the *O. rufipogon*-like plants (Taxon A) suggesting that their evolutionary history involved some introgression or hybridization and chloroplast capture (Brozynska et al., 2014b; Brozynska et al., 2017; Wambugu et al., 2015). One example of chloroplast capture in the other direction was also detected (WR-65). This illustrates a dynamic state of evolution of wild *Oryza* in Australia. This type of ongoing introgression is demonstrated by the analysis of the individual chromosomes in these populations and similar events may explain the domestication of wild *indica* by introgression of domestication alleles from domesticated japonica (Civáň et al., 2015). Extensive evidence shows distinct wild progenitors populations for *indica* and *japonica* rice that require separate domestication (Civáň et al., 2015) while the presence of common domestication related alleles suggests a single domestication event (Huang et al., 2012). The discovery of natural hybrids between taxa with greater divergence than *indica* and *japonica* demonstrates the potential for similar hybridization events to be associated with the transfer of domestication related alleles during rice domestication.

Further research should determine the diversity of useful alleles in these populations that might be incorporated into domesticated rice to improved stress tolerance and grain quality. The need for increased efforts to conserve these species *in situ* and *ex situ* is suggested by the very limited collection of this material in seed collections and the more limited distribution of the *O. rufipogon* like populations in the wild in locations that may be threatened by the incursion of weeds.

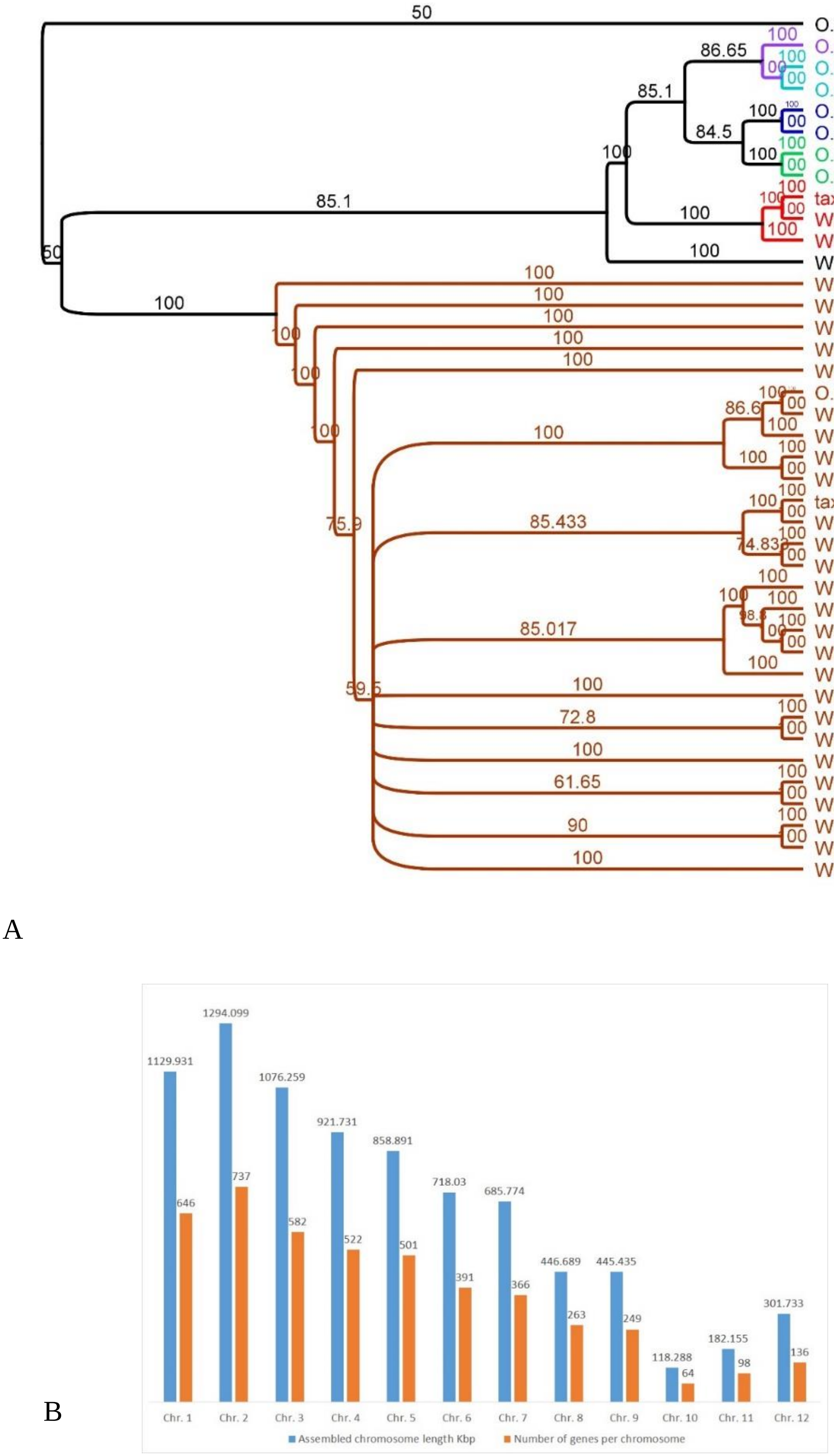


Figure 10 Individual chromosome analysis showing diversity of nuclear genomes A, Phylogenetic tree based on MP analysis of the concatenated alignment of all nuclear genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap values (Maximum Parsimony, 1000 replicates) are shown on the branches; B, Individual chromosome length and number of genes per chromosome.

Chapter 5

5 Starch gene diversity in Australian wild rice

5.1 Abstract

Starch quality and quantity are crucial for rice consumers or industry. Starch properties have been linked directly to human health. Many genes are associated with starch properties. The relationship between the starch related genes: *ISA2, ISA3, PUL, SBE1, SBE3, SBE4, SSI, SSII-1, SSII-2, SSII-3, SSIII, SSIV and GBSSI* in the Australian wild rice population of Cape York were studied. The results showed that the populations previously described as taxa A, grouped with domesticated rice; while taxa B was in a different clade. Interestingly two accessions, WR-65 and WR-44, had an in between position, suggesting hybridisation between these populations. Many SNPs/FNPs were recorded in the UTRs and exonic region of these genes that could possibly impact on their expression. CDS prediction of the *GBSSI* gene showed an extra 120bp. This was due to a change in the predicted splicing site that would lead to intron retention and add 40 amino acid to the predicted protein. It seems that this addition would not affect protein structure and the active site; however, this may explain the different starch properties of this taxa reported previously. Australian wild rice populations have potential as a novel source of starch related genes which may help improve the health of rice consumers.

Keywords: Rice, starch genes, starch genes phylogenetic, gene splicing, GBSSI, intron retention

5.2 Introduction

Starch is around 90% of the dry rice grain weight and has vital importance as a direct source of energy in the human diet; but the food industry requires different rice properties to meet market requirements. Recently, increasing concerns about health problems like obesity, developing type-2 diabetes and colon disease due to lifestyle and diet changes have led to evaluation of starch properties like resistant starch (RS) that could help address these health challenges (Zhou et al., 2016). Starch consists of two kinds of polysaccharide: amylose 15-30 % and amylopectin 65-85%. Amylose is a linear chain produced by linking glucose α 1,4, while the amylopectin is a highly branched molecule composed of α 1,4 linked glucose chains with α 1,6 links that are responsible for the branching. The amylose / amylopectin ratio has great impact on the physical and chemical properties of the starch that impact on the cooking process. Rices with high amylose content tend to give fluffy single grains; on the other hand, low amylose rice tends to be glossy when cooked (Dobo et al., 2010; Pérez and Bertoft, 2010; Yan et al., 2009; Yu et al., 2011; Zhang et al., 2014).

Many genes are involved in the starch synthesis pathway, mainly granule-bound starch synthase I (*GBSSI*), starch synthase *SSI, SSII, SSIII, SSIV,* starch branching enzyme (*SBE),* starch debranching enzyme (*DBE)* and isoamylase (*ISA).* However, the *GBSSI* gene (waxy) which is expressed mainly in storage tissue such as the endosperm, has a major influence on the amylose content (Cheng et al., 2012; Dian et al., 2003; Yu et al., 2011).

The large number of genes that are involved in the starch synthesis process make understanding and manipulating this pathway much more difficult. In Arabidopsis for example, an *SSII* deficient mutant causes an increase in total amylose and amylose/amylopectin ratio; on the other hand, a double mutant deficient in *SSII* and *SSIII* gives sluggish plant growth and decreased starch content (Zhang et al., 2008). Chain length distribution analysis shows mainly independent functionality of the *SSI, BEI* and *BEIIb* genes. However a *BEIIb* deficiency reduces the short chain ratio in the amylopectin, and a be2b mutant has more amylose compared with the wild type, probably because of a reduction in amylopectin synthesis (Abe et al., 2014). While *PUL* function to some extent overlaps with *ISA1*, deficiency of *ISA1* has more impact on amylopectin synthesis than *PUL* (Fujita et al., 2009). Fujita et al. (2011) suggested *SSI* or *SSIIIa* alone were essential for starch biosynthesis, and remarkably, found 30-33 % amylose with high *SSI* activity and recessive *SSIII*. (Kharabian-Masouleh et al., 2012) identified 66 functional SNPs in 18 starch biosynthesis related genes. Of these, 31 SNP were found to be associated with cooking quality. Other studies have shown resistant starch properties as a result of deficiency of the *SSIIIa* gene and high expression of the waxy gene (Zhou et al., 2016), whereas, a single amino acid substitution in the *SBE3* gene (leucine in the wild changed to Proline in the mutant) resulted in resistant starch in rice (Yang et al., 2012).

In rice cultivars, three different alleles have been identified in *GBSSI,* based on the number of CT repeats in the 5′-UTR as well as SNPs in the splicing site of the first intron, exons 4, 6 and 10. These variants are associated with a huge variation in the mRNA expression level of up to 10 times, which is in turn is associated with the amylose content (Cai et al., 1998; Chen et al., 2008b; Dobo et al., 2010; Hirano et al., 1996; Hirose and Terao, 2004; Isshiki et al., 1998; Larkin and Park, 2003, 1999; Mikami et al., 2008). Other researchers have reported changes in the exon intron splicing region of *SSII-1* gene, that cause an alternative transcript leading to the addition of a 28 bp fragment to the mature mRNA (Chen et al., 2017). The sequences of up to ten nucleotides on the edges of the introns and exons (exon, intron splicing enhancer and silencer) have extreme importance, as they can shape the transcriptome by influencing splicing and expression. Any change in these regions might influence the expression level or protein sequence (Jian et al., 2013; Prathepha, 2007).

Starch traits have been under strong selection throughout the history of rice domestication, as they are directly linked to consumer preferences. Wild rice does not have sticky starch, stickiness being one starch trait, as stickiness was carefully selected for only after domestication; and the development of glutinous rice, may have occurred over many stages.

Evolutionary study of *GBSSI* shows two major and six minor haplotypes in wild and domesticated rice. The H2 allele was the most ancient one found in 89% of the accessions. In domesticated rice the *GBSSI* gene has had three independent paths in rice evolutionary history. *aus* rice has the oldest one. This agrees with the theory of three independent origins of the domesticated rice (Civáň et al., 2015; Kim et al., 2016; Singh et al., 2015; Singh et al., 2017). *GBSSI* gene variation was found to be less in the wild than in cultivated rice, which demonstrates that selection pressure has been applied it to meet the demands of different consumers during domestication (Cheng et al., 2012; Singh et al., 2017; Vaughan et al., 2008).

Alternative splicing events are well known in plants and impact on post transcriptional regulation and may result in protein diversity. Alternative splicing provides ability to adjust the transcriptome according to the environment, and can be divided in to exon skipping , intron retention, alternative donor and alternative acceptor changes (Cooper et al., 2009; Wang and Brendel, 2006). Arabidopsis and rice have been used as models in studies of alternative splicing. In rice, for instance, around 20% of the expressed genes showed nearly 14500 alternative splicing events, 53.5% of which were intron retention and 13.8% exon skipping; whereas, in human, 58% of alternative splicing was reported as exon skipping and intron retention was just 5%. In Arabidopsis, 40 % of the genes have alternative splicing events shared with rice, suggesting that there is a conserved mechanism regulating this process and involved in plant evolution (Kiegle et al., 2018; Wang and Brendel, 2006). In rice more than 50% of genes have splicing events responsive to stress in the environment (Zhiguo et al., 2013).

Australian wild rice has a very high amylose content and has a different amylose and amylopectin structure as well as pasting properties and fine molecular structure, suggesting an alternative biosynthesis mechanism that can lead to new rice products. This may allow development of new cultivars with low glycemic index, which is important for diabetic rice (Calingacion et al., 2014; Tikapunya et al., 2017b).

The aim of this study is: 1, to understand the diversity of starch genes in the Australian wild rice population. 2, determine the functional variation in these genes (nominate synonymous and non-synonymous SNPs in the coding region as well as the variation in the exon-intron splicing enhancer and silencer that have potential impact on the transcriptome). This study aims to better understand the variation in starch properties of these taxa and their potential utility in rice breeding and production.

5.3 Materials and methods

5.3.1 Australian wild rice collection

Samples were collected during May 2015 and 2016 from north eastern Queensland, Australia. Locations ranged from south of Townsville to the most northerly parts of Cape York Peninsula (B). Vegetative material was collected from 29 sites. At each site, GPS coordinates and phenotypic characteristics were recorded. DNA was extracted as described by (Furtado, 2014). The extracted DNA was subjected to quality and quantity checks. Thereafter samples were sequenced with a Hiseq 4000 (Illumina), using a 2X 150 paired end technique, with an aim to produce approximately 10 X whole genome coverage on average. See Chapter 4for GPS locations and other details (Moner et al., 2018).

5.3.2 Starch related gene sequence

Raw sequence data were imported into CLC genomic workbench V.10, and mapped to the *Oryza sativa japonica* Group (assembly Build 4.0) as a reference. Gene loci (Table 12) and A) were extracted using CLC extraction tools. Thereafter, all sequences were imported into Geneious V9.1.5 (geneious.com) and aligned using the MAFFT plugin tool (Katoh et al., 2002). The alignment file was inspected physically for any errors or misaligning. SNP finding and annotation tools were used to identify synonymous and non-synonymous nucleotides and amino acid substitutions.

Table 12 Details of thirteen starch related genes in rice reference gene name, ID and size are shown.

	Gene name	Size bp	Gene ID NCBI database
1	*ISA2*	2724	4338695
2	*ISA3*	11317	4347328
3	*PUL*	13139	4335042
4	*SBE1*	7644	4342117
5	*SBE3*	11571	4329532
6	*SBE4*	3309	4335763
7	*SSI*	7746	9269493
8	*SSII-1*	8015	4348711
9	*SSII-2*	5006	4330709
10	*SSII-3*	4976	4340567
11	*SSIII*	7943	4337056
12	*SSIV*	8082	4331077
13	*GBSSI*	5065	4340018

ISA: starch-debranching enzyme isoamylase, *PUL*: starch-debranching enzymes pullulana, *SBE:* starch branching enzyme, *SS*: soluble starch synthesis enzyme, *GBSS*: granule-bound starch synthesis

5.3.3 Phylogenetic analysis

Bayesian Inference (BI), Maximum Likelihood (ML), and Maximum Parsimony (MP) approaches, using the software packages MrBayes (Huelsenbeck and Ronquist, 2001), RAxML (Stamatakis, 2006; Stamatakis et al., 2008) and PAUP (Swofford, 2003) respectively were utilised to infer the evolutionary relationships. The phylogenetic analysis was done based on two levels: individual genes and all genes combined in one alignment file.

5.3.4 CDS prediction

Full *GBSSI* gene sequences were uploaded to the GENSCAN web server: http://genes.mit.edu/GENSCAN.html. for analysis, organism module: Arabidopsis, with suboptimal exon cutoff =1. Print option: predicted CDS and peptides (Burge and Karlin, 1997; Burge and Karlin, 1998; Salzberg et al., 1998).

5.3.5 Protein model

Predicted amino acid was used to find the best homology model through SWISS-MODEL server: https://swissmodel.expasy.org/ (Arnold et al., 2006; Biasini et al., 2014; Kiefer et al., 2008)

5.3.6 Protein alignment and 3D structure

The protein 3D structure was obtained by upload the protein model file.pdb to the FATCAT server: http://fatcat.sanfordburnham.org (Ye and Godzik, 2003).

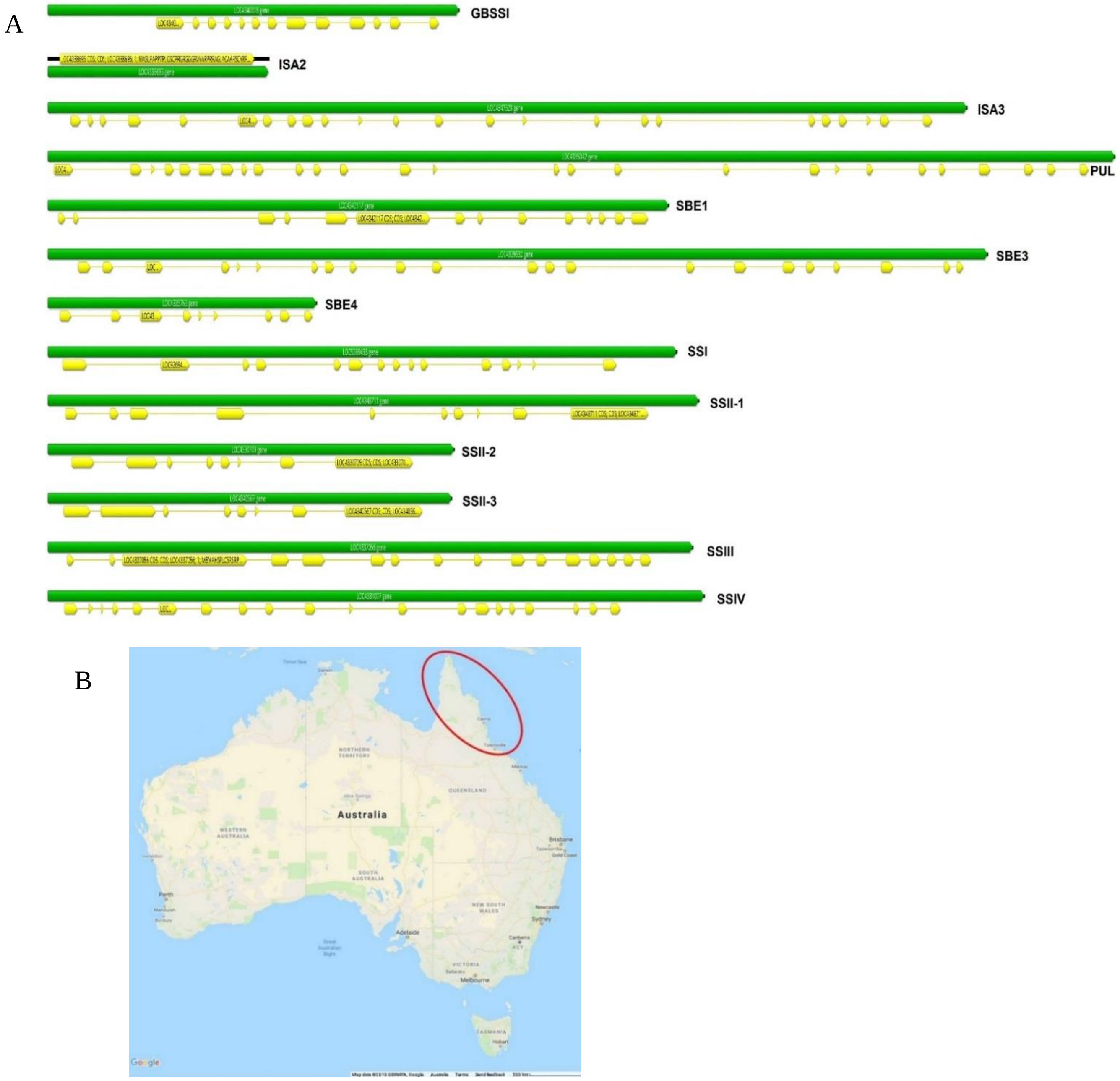


Figure 11 A. Gene structure of 13 starch related genes. Green bars are complete gene sequences, yellow bars are exons. B. Australian wild rice collection sites North of Queensland

5.4 Results

A phylogenetic tree of the 13 starch related genes, clearly shows two main clades. The populations described earlier as Taxa A have grouped with the domesticated rice reference (*O. sativa* japonica). Accessions of the other populations, Taxa B, all grouped together in a separate clade. Interestingly, two accessions WR-65 and WR-44 were intermediate between these clades (Figure 12 and Figure 45 -57). WR-65 and WR-44 were examined further due to their location in the phylogenetic tree. The alignment file shows two types of reads in both accessions for some of these genes. These variants seem to reflect the heterozygous nature of these plants (Figure 13). This suggests that they have resulted from hybridisation between these populations in agreement with our overall analysis of the nuclear genes (Moner et al., 2018).

Individual starch related genes (*ISA2, ISA3, PUL, SBE1, SBE3, SBE4, SSI, SSII-1, SSII-2, SSII-3, SSIII, SSIV and GBSSI*) were not all intermediate. Five genes (*SBE3, SSI, SSII-1, SSIII and SSIV*) have different associations jumping between clades A and B for these two accessions (WR-65 and WR-44). Moreover, some of these genes (*ISA2, PUL, SBE1, SBE3, SSI, SSII-1, SSII-3 and SSIV*) divide into at least two main sub clades in the Taxa B population (Figure 45 -57). The GBSSI gene phylogenetic tree shows that Australian wild rice can be grouped into the three different groups previously reported in the evolutionary history of this gene (Figure 12 and Figure 45-57) (Singh et al., 2017).

Nucleotide variation (synonymous and nonsynonymous) showed some differences in each gene (Table 13). The highest SNPs/ FNPs were in the *ISA2*, *SSII-2*, and *SSIII* genes respectively, while the lowest were in *SSI* and *GBSSI*. Some of these SNPs/FNPs were highly specific to either Taxa A or B. (Table 35 and Supplementary File 1). Interestingly, overlaying these differences with annotation information showed that many of these variations were located in the UTR and exons intron boundaries. Because very high amylose content had been recorded in these populations and *GBSSI* has the main role in amylose biosynthesis, the large number of variations in the intron exon boundary of this gene were investigated and the likely sequences of cDNAs were predicted. The full length sequences of the *GBSSI* gene for Taxa A, B, *O. rufipogon* Asian populations and *O. sativa japonica* as validation reference, were predicted. Several SNPs were recorded in these accessions, but these did not affect the length of the transcripts. However, Taxa B had a large insertion of 120bp (Figure 14 and Figure 15) that could provide an explanation of the high amylose content in this taxon.

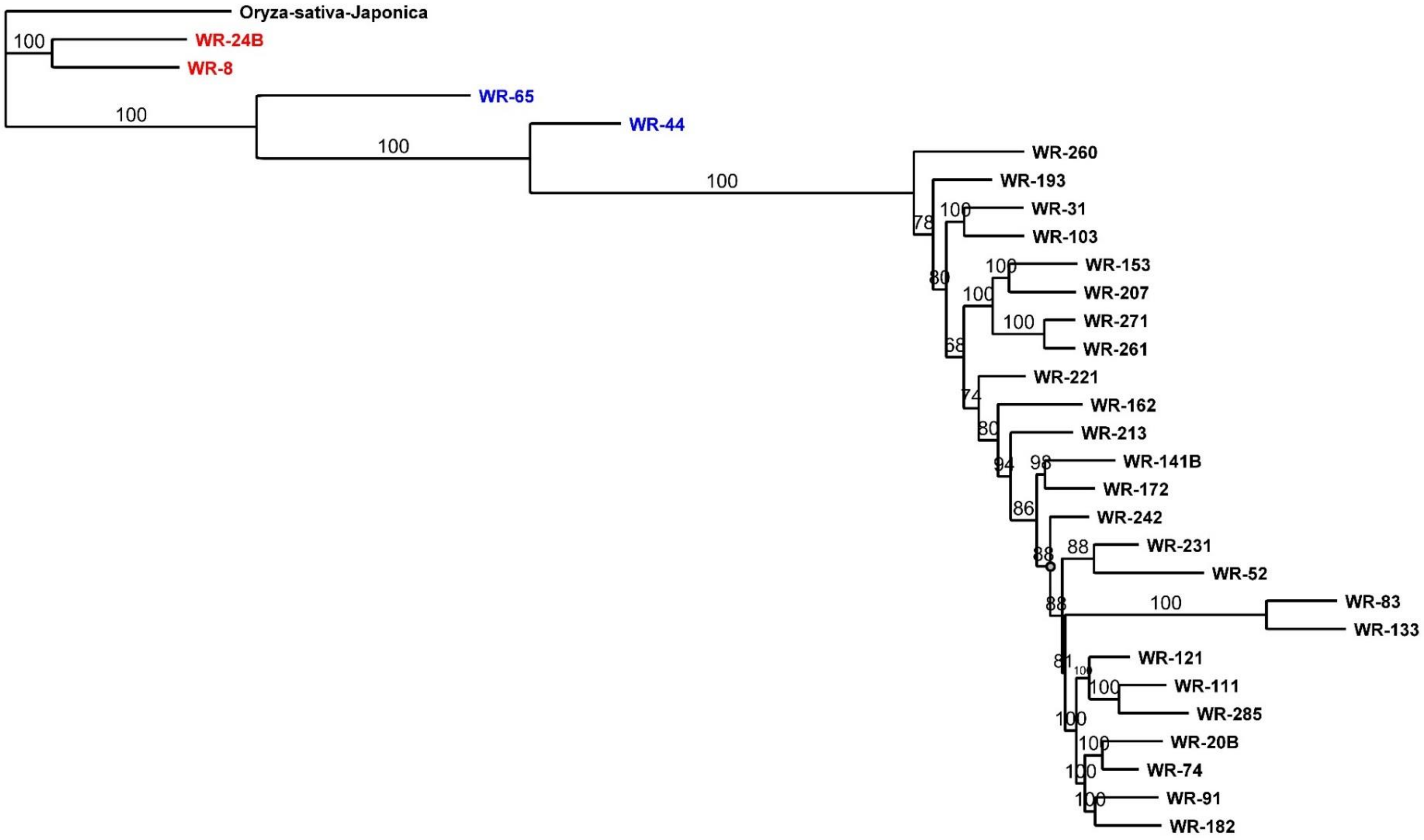


Figure 12 Phylogenetic tree based on maximum likelihood and Bayesian analysis (both agreed in topology) of 13 starch gene sequences. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids

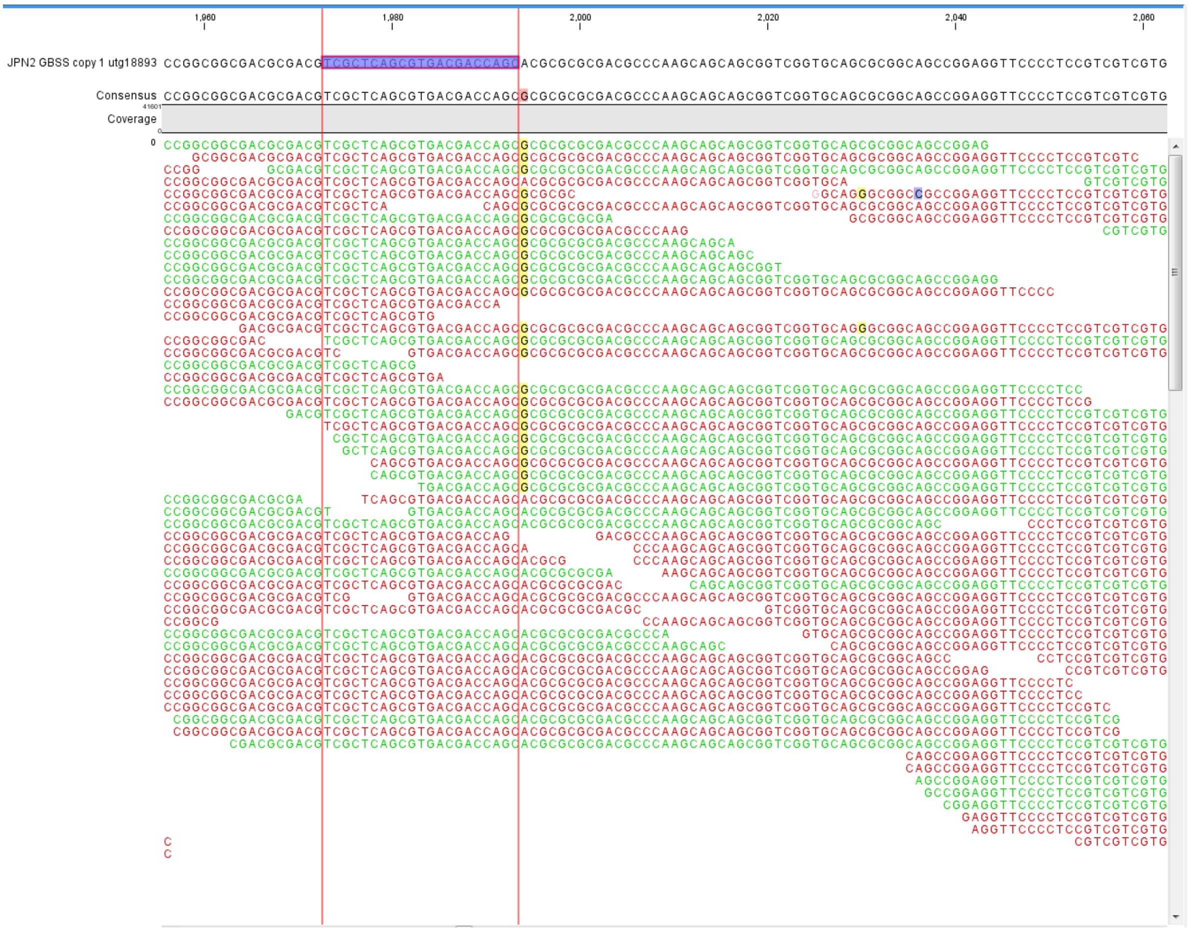

1,960
1,980
2,000
2,020
2,040
2,060
JPN2 GBSS copy 1 utg18893 CCGGCGGCGACGCGACGTCGCTCAGCGTGACGACCAGCACGCGCGCGACGCCCAAGCAGCAGCGGTCGGTGCAGCGCGGCAGCCGGAGGTTCCCCTCCGTCGTCGTG
Consensus CCGGCGGCGACGCGACGTCGCTCAGCGTGACGACCAGCGCGCGCGCGACGCCCAAGCAGCAGCGGTCGGTGCAGCGCGGCAGCCGGAGGTTCCCCTCCGTCGTCGTG
41601
Coverage
0
0

Figure 13. Two type of reads as evidence of hybridisation in Australian wild rice population from North Queensland.

The results suggest 12 exons in taxa B whereas 13 were predicted for the others as shown previously. Exon 11 in the taxa B was predicted to be 336 bp which is equivalent to exon 11 and 12 and the insert of 120bp. The intron between exon 11 and 12 equal 120bp. This suggest that the whole intron remained and was not removed during the predicted splicing process. This led to an additional 40 amino acid in the predicted protein but kept the sequence in frame. One T/A SNP in the intron 11 splicing enhancer is possibly responsible for this intron retention. The 3D structure comparison between taxa B and the reference shows significant differences in the linking region as well as the beta sheet (Figure 16).

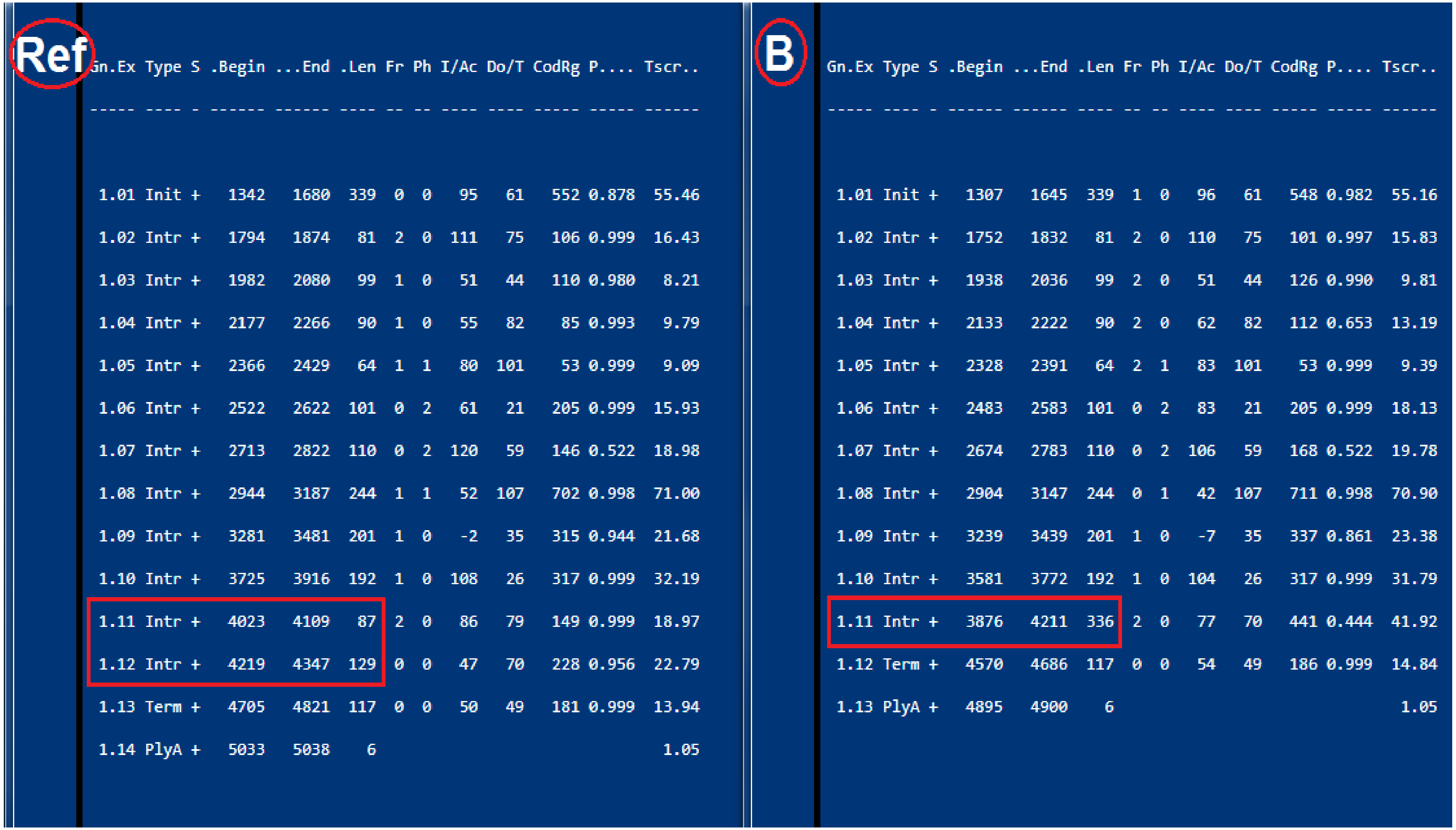
5.5 Discussion

The Australian wild rice populations of the Cape York have unique characteristics (Brozynska et al., 2017). Starch analysis of these populations shows in general high amylose content compared the domesticated cultivars (Tikapunya et al., 2017b).

Table 13. SNPs/FNP summary for starch related genes compared to the *O. sativa japonica* assembly Build 4.0

Genes	SNPs	per accession	FNPs	per accession
ISA2	495	17.07	44	1.52
ISA3	235	8.11	23	0.8
PUL	346	11.94	40	1.38
SBE1	108	3.73	14	0.49
SBE3	90	3.11	9	0.32
SBE4	172	5.94	21	0.73
SSI	57	1.97	7	0.25
SSII-1	315	10.87	18	0.63
SSII-2	512	17.66	35	1.21
SSII-3	327	11.28	33	1.14
SSIII	585	20.18	36	1.25
SSIV	157	5.42	27	0.94
GBSSI	12	0.42	10	0.35

Phylogenetic analysis of starch related genes indicated that the Australian wild rice, Taxa B, accessions were well differentiated from domesticated rice. Their starch related genes may explain their different starch structure and content, especially their high amylose (Tikapunya et al., 2017b). Starch related genes in general were subjected to selection over the course of domestication and breeding to enhance the cultivar to suit human use and taste. GBSSI and SBE genes in particular were under strong selection pressure due to requirements to meet the demand of different consumers. This led to the loss of important allele from those genes and also other starch related genes (Yu et al., 2011). Australian wild rice as an intact population can deliver varieties of alleles to develop new cultivars with specific starch properties for consumption of healthy rice with low glycemic index or even for industry requirements (Brozynska et al., 2015; Henry et al., 2010).



```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Init +   1342   1680  339  0  0   95   61   552 0.878  55.46
 1.02 Intr +   1794   1874   81  2  0  111   75   106 0.999  16.43
 1.03 Intr +   1982   2080   99  1  0   51   44   110 0.980   8.21
 1.04 Intr +   2177   2266   90  1  0   55   82    85 0.993   9.79
 1.05 Intr +   2366   2429   64  1  1   80  101    53 0.999   9.09
 1.06 Intr +   2522   2622  101  0  2   61   21   205 0.999  15.93
 1.07 Intr +   2713   2822  110  0  2  120   59   146 0.522  18.98
 1.08 Intr +   2944   3187  244  1  1   52  107   702 0.998  71.00
 1.09 Intr +   3281   3481  201  1  0   -2   35   315 0.944  21.68
 1.10 Intr +   3725   3916  192  1  0  108   26   317 0.999  32.19
 1.11 Intr +   4023   4109   87  2  0   86   79   149 0.999  18.97
 1.12 Intr +   4219   4347  129  0  0   47   70   228 0.956  22.79
 1.13 Term +   4705   4821  117  0  0   50   49   181 0.999  13.94
 1.14 PlyA +   5033   5038    6                                 1.05
```

```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Init +   1307   1645  339  1  0   96   61   548 0.982  55.16
 1.02 Intr +   1752   1832   81  2  0  110   75   101 0.997  15.83
 1.03 Intr +   1938   2036   99  2  0   51   44   126 0.990   9.81
 1.04 Intr +   2133   2222   90  2  0   62   82   112 0.653  13.19
 1.05 Intr +   2328   2391   64  2  1   83  101    53 0.999   9.39
 1.06 Intr +   2483   2583  101  0  2   83   21   205 0.999  18.13
 1.07 Intr +   2674   2783  110  0  2  106   59   168 0.522  19.78
 1.08 Intr +   2904   3147  244  0  1   42  107   711 0.998  70.90
 1.09 Intr +   3239   3439  201  1  0   -7   35   337 0.861  23.38
 1.10 Intr +   3581   3772  192  1  0  104   26   317 0.999  31.79
 1.11 Intr +   3876   4211  336  2  0   77   70   441 0.444  41.92
 1.12 Term +   4570   4686  117  0  0   54   49   186 0.999  14.84
 1.13 PlyA +   4895   4900    6                                 1.05
```

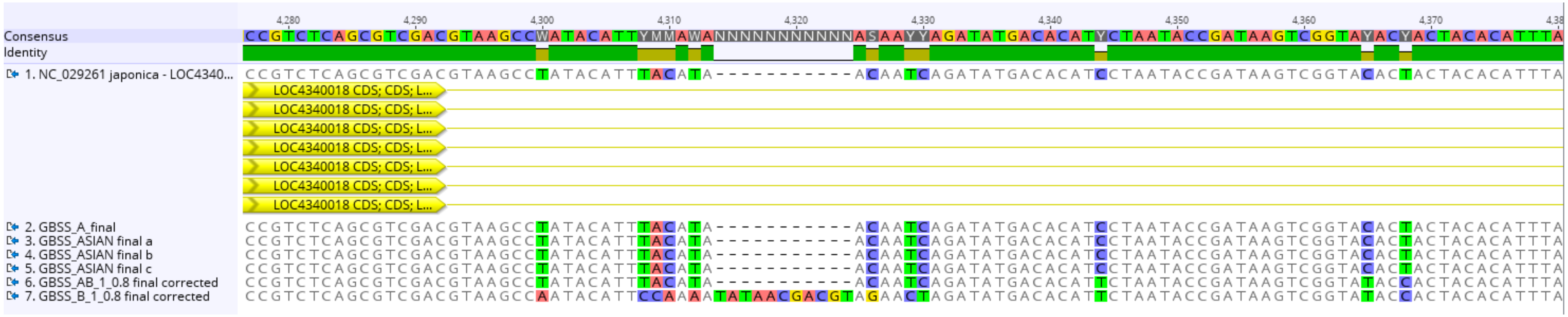
Figure 14. Prediction of the CDS and determined exons boundary in GBSSI taxa B compared to the reference O. sativa japonica. Differences are highlighted by the red rectangle. In Taxa B exon 11 and 12 were combined and included the intron between them.

Intermediate location of those accessions (WR-65 and WR-44) and jumping between clades across all starch related genes was interesting (Figure 12, 13 and Figure 45-57). Read alignment showed two types of reads that are unlikely to be an error and gave strong evidence of hybridisation between Australian wild rice populations taxa A and B (Moner et al., 2018). The degree of exchange of genomic material was not equal in all starch related genes; therefore they were in different positions in the phylogenetic trees.

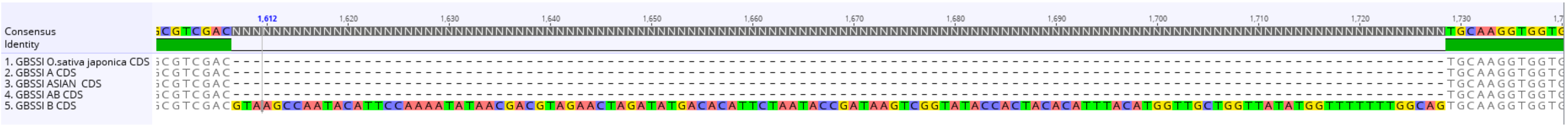
Numbers of SNPs have been identified in those 13 starch related genes (Table 35 and Supplementary File). Their locations were in the 5`UTR, exon and intron boundaries that regulate the expression level and final transcriptome (Srivastava et al., 2018). Specific allele in the UTR and exons of GBSSI influenced the proportion of amylose /amylopectin (Butardo et al., 2016). Splicing regions and their impact on transcription has been well studied in the abundance of human genome resources and plants. In general, several bases up to ten, in either 5` or 3` of the exon-intron boundaries, control this process. Any change in this area impacts on the spliceosome binding site and can cause alternative splicing which can change protein sequences (Jian et al., 2013; Srivastava et al., 2018). Epigenetic mechanisms and co-transcription might be involved in Splicing pre-mature mRNA (Gelfman et al., 2013), by changing chromatin structure and RNA polymerase II elongation, which eventually impact on the spliceosome configuration (Luco et al., 2011; Maor et al., 2015; Ullah et al., 2018; Yearim et al., 2015). All the above might play an important role in the variations in the starch properties that were reported previously in those populations (Tikapunya et al., 2017b).

The *GBSSI* gene in particular, as the key gene associated with amylose synthesis, has many variations in the 5` UTR of the Australian wild rice accessions, which may be associated with regulating the expression level and post translation regulation of this gene, as well as the splicing process (Barrett et al., 2012; Liu et al., 2009; Srivastava et al., 2018; Terada et al., 2000). (Mishra et al., 2016) studied the variation in the 5` UTR of the *OsClpB-C* gene during heat stress and found that it has an essential role in the post-transcriptional control and expression of the *OsClpB-C* gene as well as being involved in ribosomal assembly.

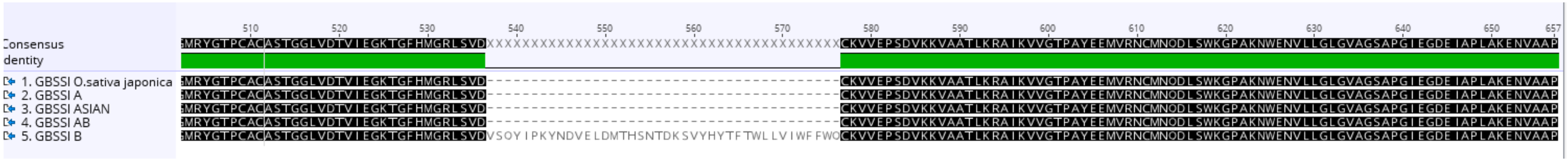
An SNP change from A to G resulting in a change from the negative charged amino acid Aspartic to non-polar amino acid Glycine, did not seem to affect the gene activity *in vitro*, but in fact impacted on starch granule binding and eventually reduced amylose content (Wang et al., 1995; Ayres et al., 1997; Cai et al., 1998). One amino acid change from Cysteine (non- polar) to Valine (non-polar) lead to over expression and a change to an insoluble form,


4,280
4,290
4,300
4,310
4,320
4,330
4,340
4,350
4,360
4,370
4,38
Consensus CCGTCTCAGCGTCGACGTAAGCCWATACATTYMMAWANNNNNNNNNNNASAAYYAGATATGACACATYCTAATACCGATAAGTCGGTAYACYACTACACATTTA
Identity
1. NC_029261 japonica - LOC4340... CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATCCTAATACCGATAAGTCGGTACACTACTACACATTTA
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
LOC4340018 CDS; CDS; L...
2. GBSS_A_final CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATCCTAATACCGATAAGTCGGTACACTACTACACATTTA
3. GBSS_ASIAN final a CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATCCTAATACCGATAAGTCGGTACACTACTACACATTTA
4. GBSS_ASIAN final b CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATCCTAATACCGATAAGTCGGTACACTACTACACATTTA
5. GBSS_ASIAN final c CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATCCTAATACCGATAAGTCGGTACACTACTACACATTTA
6. GBSS_AB_1_0.8 final corrected CCGTCTCAGCGTCGACGTAAGCCTATACATTTACATA-----------ACAATCAGATATGACACATTCTAATACCGATAAGTCGGTATACCACTACACATTTA
7. GBSS_B_1_0.8 final corrected CCGTCTCAGCGTCGACGTAAGCCAATACATTCCAAAATATAACGACGTAGAACTAGATATGACACATTCTAATACCGATAAGTCGGTATACCACTACACATTTA

A


1,612
1,620
1,630
1,640
1,650
1,660
1,670
1,680
1,690
1,700
1,710
1,720
1,730
1,7
Consensus GCGTCGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGCAAGGTGGTG
Identity
1. GBSSI O.sativa japonica CDS GCGTCGAC--TGCAAGGTGGTG
2. GBSSI A CDS GCGTCGAC--TGCAAGGTGGTG
3. GBSSI ASIAN CDS GCGTCGAC--TGCAAGGTGGTG
4. GBSSI AB CDS GCGTCGAC--TGCAAGGTGGTG
5. GBSSI B CDS GCGTCGACGTAAGCCAATACATTCCAAAATATAACGACGTAGAACTAGATATGACACATTCTAATACCGATAAGTCGGTATACCACTACACATTTACATGGTTGCTGGTTATATGGTTTTTTTGGCAGTGCAAGGTGGTG

B


510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
657
Consensus GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXCKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP
Identity
1. GBSSI O.sativa japonica GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVD--CKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP
2. GBSSI A GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVD--CKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP
3. GBSSI ASIAN GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVD--CKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP
4. GBSSI AB GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVD--CKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP
5. GBSSI B GMRYGTPCACASTGGLVDTVIEGKTGFHMGRLSVDVSQYIPKYNDVELDMTHSNTDKSVYHYTFTWLLVIWFFWQCKVVEPSDVKKVAATLKRAIKVVGTPAYEEMVRNCMNQDLSWKGPAKNWENVLLGLGVAGSAPGIEGDEIAPLAKENVAAP

C

Figure 15 A. SNP in the intron 11 splicing enhancer of the GBSSI gene B. intron 11 retention and 120 bp insertion in the CDS C. insertion of extra 40 amino acid as a consequence of the intron retention.

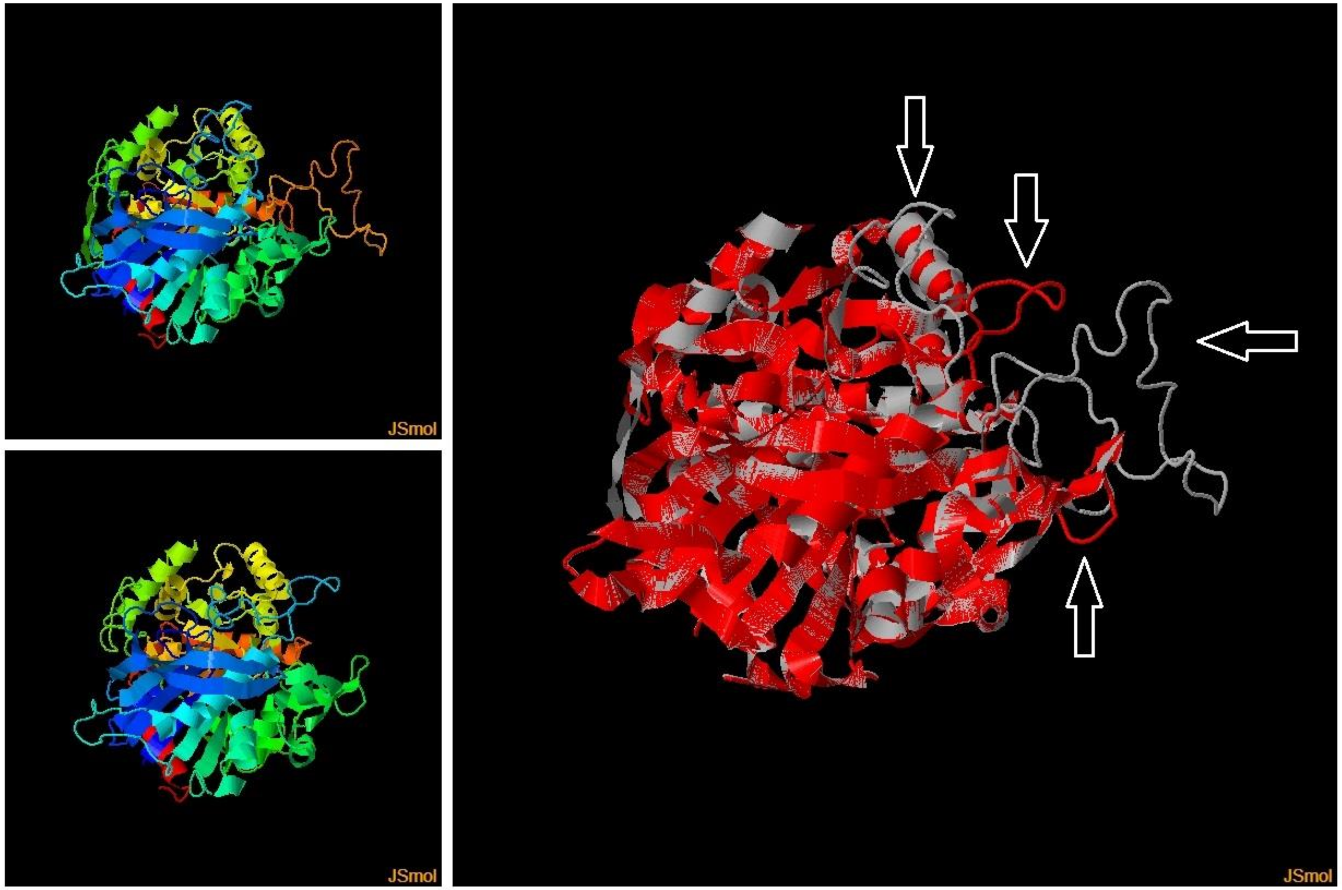


Figure 16. JSmol display of the GBSSI 3D structure alignment (superposition) Taxa B with Reference O.sativa *japonica* (3vue.1.A) using FATCAT. Taxa B and reference are in grey and red respectively. White arrow indicates the difference in the structure between these genes. left top is Taxa B left bottom reference

indicating that a disulfide bond controlled the three dimension stability of the 3D structure and may be very important in maintaining domain arrangement and increasing the efficiency of starch biosynthesis (Momma and Fujimoto, 2012). The number of dinucleotide $(CT)_n$ in the 5` UTR and the first intron splicing junction have been linked with amylose content in some *indica* varieties (Zhu et al., 2003). On the other hand, duplication of 23 bp in the second exon or a SNP in the fourth exon can cause a glutinous trait and loss of binding function between starch granules(Hori et al., 2007; Liu et al., 2009). A combination of several SNPs in exons 6 and 10 led to a change in the amino acid and the splicing site of the first intron, all leading to a range of amylose contents (Chen et al., 2008a; Chen et al., 2008b; Dobo et al., 2010; Hoai et al., 2014). Changing G to T led to incomplete post transcriptional processing of the immature mRNA, giving a glutinous trait (Hirano et al., 1998).

Alternative splicing impacts on gene expression can lead to exon skipping, intron retention or frame shifting that changes or makes nonfunctional the eventual protein (Cartegni et al., 2002). For instance, a G to T SNP in intron 25 of the *DFNA1* gene interrupted the splicing donor site that is responsible for nonsyndromic deafness in humans. This SNP caused a 4 base insertion and frame shift, premature termination and the deletion of 32 amino acids from the protein. (Lynch et al., 1997). As an additional example, a C to T SNP in the seventh exon of the *SMN2* gene results in a truncated protein by changing exon splicing enhancer ESE to exon splicing silencer ESS (Cartegni et al., 2006; Cartegni and Krainer, 2002).

The 40 amino acid insertion reported here as an intron retention event, changed the 3D structure of this protein slightly (Figure 17 - 19). The distance between the nearest residue in the active site, Thr., and the new inserted residue, Phe, was around 15A°, hence it was not likely to affect the active site. The disulfide bond plays an important role in stabilising the protein domain (Figure 18) (Momma and Fujimoto, 2012). The new inserted residues near the disulfide bond also did not appear to affect its function. This was clearly by shown by domain similarity to the reference (Figure 19). However, it impacts on the beta sheet and linking region. This new structure might affect the protein binding or early/ late termination per unit.

In conclusion, a number of variations have been found between domesticated and Australian wild rice starch related genes. These were in critical positions that impact on genes regulation, expression and final transcriptome, which affects the starch properties. More experiments are essential to identify the useful variations, as well as to eliminate the deleterious mutations that might reduce the quantity or harm the quality, and affect how we can employ them to improve existing high quality and healthy rice.

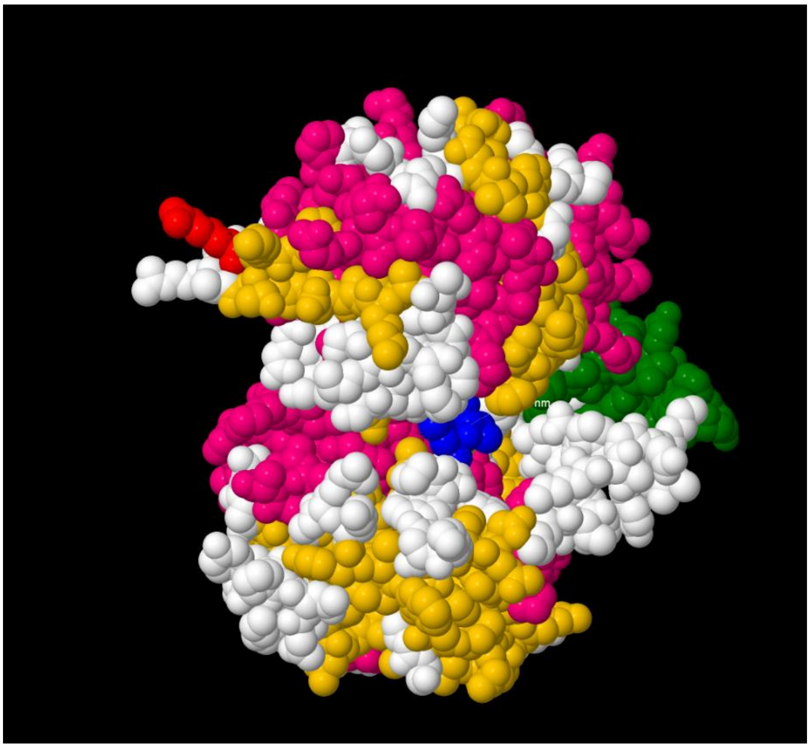


Figure 17. 3D structure filled of the GBSSI gene protein Taxa B blue, green and red colour referring to the pocket KTGGL, 40 amino acid insertion and one amino acid change Ser to Arg

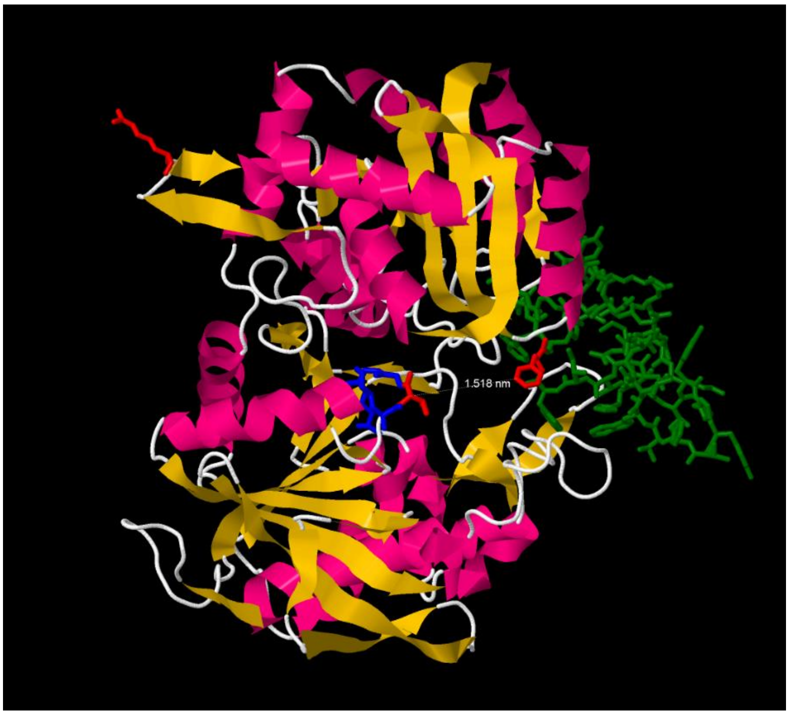


Figure 18. Three dimension structure of the GBSSI gene of Taxa B. The closest distance between the Thr in the active site and the 40 amino acid insertion Phe was 15 A°

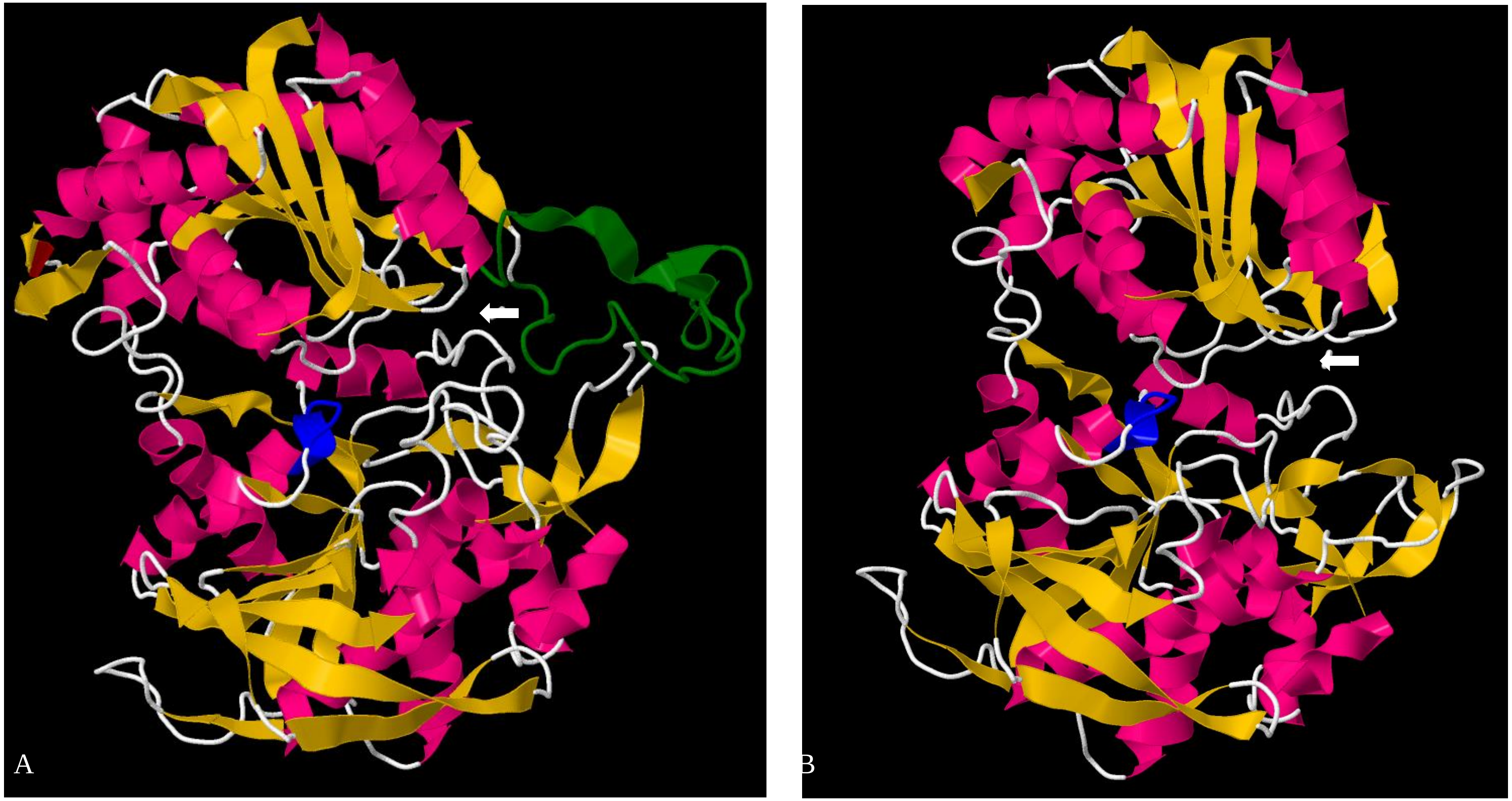


Figure 19 A, GBSSI gene of Taxa B; B, O. sativa japonica. Disulfide bond shown by white arrows

Chapter 6

6 General discussion

6.1 Fulfilment of objectives

This study extends previous studies (Brozynska et al., 2017; Sotowa et al., 2013; Tikapunya et al., 2017a) which reported potentially two new species of wild rice in the North of Queensland. These species are both different from *O. rufipogon,* the closest wild relative of domesticated rice. The first of these studies covered morphological characters with investigation of some genomic loci. The second study was a comprehensive whole genome nuclear and chloroplast assembly and annotation as well as a study of the relationship to other *Oryza* species. However, the study was based on just two individual plants. The third study explored the possibility of consumption of rice from these populations and the grain properties of the rice.

This thesis reports (Chapter 3) the assembly of high quality chloroplast sequences of wild rice populations of Asia as the closest geographic populations to the Australian wild rice, in order to study the phylogeny of these populations. SNPs and other molecular markers were defined to identify and distinguish these populations. Chapter 4 reports an extensive survey of populations from Townsville to the tip of Cape York, with wild rice collected from 27 different sites. This collection showed clearly the two distinct taxa. Chapter 5 was focused on the starch related genes following the report of interesting starch properties, especially the amylose content in these populations (Tikapunya et al., 2017b). The phylogeny of starch related genes was studied individually and together. As *GBSSI* gene has the main role in amylose synthesis, it was studied in more detail. In this chapter, we will discuss the key findings and suggestions for further study of these interesting wild populations.

6.2 Chloroplast genomes of Asian wild rice

The chloroplast is a conserved maternally inherited genome, and has been used as barcode to track the evolution of plant species. A dual pipeline procedure was developed using mapping of reads to a reference and de novo approaches, in order to assemble high quality chloroplast genomes which allowed elimination of assembly errors that may have been counted as a difference previously. Any errors may impact negatively on the analysis of evolutionary relationships and may provide an erroneous assessment to the evolutionary history of the *Oryza* genus. Average coverage was a critical criterion for acceptance of results in the dual pipeline. However, sometimes even with relatively high coverage, there were still some small genomic areas with no coverage due to deletion or chance lack of sequencing. Analysis of 31 wild Asian and 9 domesticated accessions covering South and South East Asia down to the North of Australia gave a perspective on the evolution of these wild populations and how they interact with the surrounding environment.

The phylogenetic tree shows that genetic variation of the wild rice populations is mainly distributed according to geographic origin (based on continent). Interestingly, the Australian type extended to the North (Philippines). Asian populations overlapped and there is no cut off line to separate them, possibly because of the impact of human movement. Two main sub clades representing the origin of the domesticated rice *japonica* and *indica* sub species were identified. The separation of these two subclades supports the multiple domestication theory. Domesticated species of *aus* appeared in both subclades, suggesting that both maternal genomes were involved in this domestication.

The nuclear genome diversity in these wild rices does not follow the same pattern as for the chloroplast genomes (Figure 4) (Civáň et al., 2015; Huang et al., 2012). This suggests that the evolution of the wild progenitors of domesticated rice followed a complex path, probably involving many dispersal events and chloroplast capture. Interestingly, the majority of the accessions in the chloroplast clade, including *O. nivara*, had *japonica* like nuclear genomes; while the majority of the chloroplast clades related to *japonica* and *indica* were intermediate in nuclear genome (Huang et al., 2012).

The chloroplast is not just responsible for photosynthesis but also affects intracellular signaling and performances and responses to the environment. The survival of these populations in the Australian environment mean that these wild plants may have alleles that could contribute to adaptation of this crop to different environments, allowing rice to be grown in new areas. Here we reported 36 nonsynonymous (FNPs) distributed over 13 genes (*atpB, atpI, ccsA, cemA, clpP, matK, ndhF, ndhK, psaA, psbB, rpoC1, rpoC2* and *rps18*) that could provide adaptation to specific environments. especially when they control vital biological processes in the plant cell like ATP synthesis, envelope membrane protein, NADH dehydrogenase, photosystem I and II, ribosomal protein S18 and RNA polymerase.

Maternal genomes, including the chloroplast and mitochondria, have a great impact on the overall phenotype. Just two chloroplast types have domesticated in Asian rice. Other wild chloroplasts failed to pass through the domestication bottle neck and strong selection pressure over thousands of years of the domestication process. Introducing these new wild chloroplast types might help to adapt rice to various environments, or add interesting performance re abiotic / biotic stresses. Analysing chloroplast genomes provides a useful tool for conserving and utilising the genetic resources in the A genome genepool of *Oryza* species and for supporting food security.

6.3 Phylogeny of Australian wild rice populations

We did a comprehensive survey looking for wild rice from Townsville up to the tip of Cape York over two years, 2015 and 2016. Wild rice was found in 27 sites, around creeks and lake margins. Water availability was the key factor in finding these wild plants. Interestingly, there was no wild rice after crossing the Jardine River (-11.103665, 142.283901) up to the tip and to the Islands of Torres Strait. It is unclear why rice does not extend further north on the Cape. As previously reported, wild rice showed significant morphological differences compared to the domesticated rice–mainly by way of long anthers with short awns and open panicles in what was reported as taxa A., and short anthers, very long awns and closed panicles in taxa B, according to previous reports. These morphological traits could be indicators of the evolutionary history of these populations. Long anthers may improve out crossing, while long awns might help seeds attach to animals and enhance distribution. This could be one of the explanations as to why this taxa separates over large areas.

Twenty-nine samples were sequenced successfully with an average coverage of around 10X and overall high quality data. Chloroplast genomes were assembled using the same dual pipeline used for the Asian data, to produce high quality chloroplast genomes for use as reference genomes for these populations in future studies. An average of 129.6 variants were recorded as SNPs, deletions or insertions compared to the reference genome *O. sativa japonica*. Six common nonsynonymous SNPs were identified in all samples, plus another 12 that were not consistent among all samples, possibly including alleles which could be useful for the rice community in improving this important crop.

Chloroplast phylogenetic analysis showed clear distinct clades. Australian wild rices were isolated from all other AA *Oryza* species, with two main subclades corresponding to taxa A and B. Australian wild rice in general was very different from the domesticated rice ancestor *O. rufipogon,* suggesting that it is most likely not the same species, as previously thought. This means Australian wild rice has a repository of new genes that have not been used before, which opens an opportunity to the rice community to add new genetic material to enhance rice varieties. Chloroplast markers that were identified could help in identifying those two main groups in a simple way.

The coding parts of genes (exons) are the key parts of the genome. A set of 4555 genes were compared in order to evaluate these wild populations according to the functionality of the coding sequences. Concatenation of all exons across the 12 chromosomes showed Australian wild rice as a distinct population from all other AA genome species. Taxa A (*O. rufipogon* like) was a sister clade to all domesticated and wild rice, while taxa B (*O. meredionalis* like) was a sister clade to all others (Asian and African *Oryza* AA genome). This indicates that these populations have unique functional material. These genes make them competitive in the Australian environment. Our analysis confirms previous studies of these populations. Individual chromosome phylogenetic analysis shows significant introgression between the different populations of wild rice.

The divergent B nuclear genome and A chloroplast genome suggest plants in some of these populations may be hybrids. Population WR-65 had a B type chloroplast but an A type nuclear genome. Both chloroplast and nuclear gene analysis suggest a high diversity of AA genome wild rice in Australia. This supports the view that Australia might be a centre of diversity for the AA genome clade. Some populations with a nuclear genome similar to *O. meridionalis* had a chloroplast genome that was closer to the *O. rufipogon*-like plants (Taxon A), suggesting that their evolutionary history involved some introgression or hybridisation and chloroplast capture. One example of chloroplast capture in the other direction was also detected (WR-65). This illustrates a dynamic state of evolution of wild *Oryza* in Australia. This type of ongoing introgression is demonstrated by the analysis of the individual chromosomes in these populations and similar events may explain the domestication of wild *indica* by introgression of domesticated alleles from domesticated japonica. The discovery of natural hybrids between taxa with greater divergence than *indica* and *japonica,* demonstrates the potential for similar hybridisation events to be associated with the transfer of domestication related alleles during rice domestication.

6.4 Starch related genes in wild rice populations

Starch analysis of these populations shows high amylose content compared with domesticated cultivars. Therefore, we focused on analysis of starch related genes. Phylogenetic analysis of these genes indicated that the Australian wild rice, Taxa B, accessions were well differentiated from domesticated rice. This may explain why they have different starch structure and content, especially high amylose. Starch related traits were one of the key factors that breeders focused on. As a result, this has been under selection to meet the consumer's requirements. Important alleles from these genes have been lost during domestication. Australian wild rice is an intact population that was not involved in domestication so can deliver novel alleles or possibly new genes to help develop new cultivars with specific starch properties for healthy rice.

Phylogenetic analysis of the starch related genes showed two accessions (WR-65 and WR-44) were in between the main clades and jumped between clades across all starch related genes. This was unexpected and required more investigation. Read alignment showed two types of reads that are unlikely to be an error, that provided strong evidence of hybridisation between Australian wild rice populations, taxa A and B. Important SNPs were identified across all 13 genes in 5`UTR, exon, exon and intron boundaries that regulate the expression level and shape the final transcriptome.

GBSSI has the main role in amylose synthesis, amylose content and amylose /amylopectin ratio; therefore, it was targeted for more attention. Many variations were found in the 5` UTR of the Australian wild rice accessions, which may be associated with regulation of the expression level and post translation regulation of this gene, as well as the splicing process. To confirm the importance of these SNPs, we predict the CDS of this gene and interestingly found one SNP (T to A) in the exon splicing enhancer that had an effect on the splicing process, causing alternative splicing and retention of the whole intron between exons 11 and 12. This intron retention might be responsible for the increased amylose and the distinct starch structure in this population. The transcript of this predicted CDS showed a 40 amino acid insertion without any effects on the translation frame. The 3D structure of this protein showed a slight change in the beta sheet and linking region but no change in the main protein domains. This insertion was also far from the protein active site which retained functionality. This new structure might affect the protein binding or early/ late termination per unit, or speed up the synthesis per time unit.

6.5 Future directions

Crop wild relatives are important genetic material to improve and develop domesticated cultivars. These wild plants represent a vast repository of undiscovered genes. Introducing them into breeding programs adds new alleles that might be the key to planting the crop in new areas which have not been used before. In this study, it was shown clearly that the Australian wild rice population of Cape York was distinct from all other wild and domesticated rice AA genomes. In addition, there is a high probability this includes a new species. This new species should receive much more attention. Priority number one is to protect this population from extinction. This new species was found in limited sites compared to the other populations, which means there is a high potential to lose it due to competition from weeds. Secondly, new species classification should be confirmed, and a new scientific name proposed.

The high quality chloroplast genomes that were assembled in this study as well as the SNPs and FNP markers, could be used to guide any further survey as a simple technique to identify unknown wild rice on a large scale, especially at the pre-selection stage. These markers should clarify the evolutionary linkage and make it easy to select from large collections.

These wild populations should be under intensive study to evaluate and characterise their desirable traits like biotic and abiotic stress, nutritional value and productivity. Further research should determine the diversity of useful alleles in these populations that might be incorporated into domesticated rice to improve stress tolerance and grain quality. Moreover, Cape York populations located in an area isolated from commercial rice fields, provide material suitable to study for other evolutionary relationships and are models for evolutionary studies and for testing new evolutionary hypotheses. These latter include hybridisation events in these populations, that prove rice evolution is a dynamic and ongoing process.

A number of hypotheses could be used to explain why starch properties in these populations showed a different structure and high amylose content. One, reported here, is intron retention in the *GBSSI* gene, which needs to be validated. RNA-seq analysis is essential to confirm this and to determine the expression level of this gene. Further studies are required to study these genes in more depth. This discovery could be the key to the production of high quality and healthy rice with low glycemic index and reduced diabetes risk.

Chapter 7

7 References

Abe, Natsuko, Hiroki Asai, Hikari Yago, Naoko F Oitome, Rumiko Itoh, Naoko Crofts, Yasunori Nakamura and Naoko Fujita 2014. Relationships between starch synthase I and branching enzyme isozymes determined using double mutant rice lines. BMC Plant Biol 14: 80.

Andersson, Stefan, Maarten Ellmer, Tove H Jorgensen and Anna Palmé 2010. Quantitative genetic effects of bottlenecks: experimental evidence from a wild plant species, Nigella degenii. Journal of heredity 101: 298-307.

Arnold, Konstantin, Lorenza Bordoli, Jürgen Kopp and Torsten Schwede 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22: 195-201.

Bal, AR and SK Dutt 1986. Mechanism of salt tolerance in wild rice (Oryza coarctata Roxb). Plant and soil 92: 399-404.

Barrett, Lucy W, Sue Fletcher and Steve D Wilton 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cellular and molecular life sciences 69: 3613-3634.

Biasini, Marco, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni and Lorenza Bordoli 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research 42: W252-W258.

Bobik, Krzysztof and Tessa M Burch-Smith 2015. Chloroplast signaling within, between and beyond cells. Frontiers in plant science 6.

Brozynska, M., A. Furtado and R. J. Henry 2014a. Direct chloroplast sequencing: comparison of sequencing platforms and analysis tools for whole chloroplast barcoding. PLoS ONE 9: e110387. doi: 10.1371/journal.pone.0110387

Brozynska, M., A. Furtado and R. J. Henry 2015. Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol J 14: 1070-1085. doi: 10.1111/pbi.12454

Brozynska, M., E. S. Omar, A. Furtado, D. Crayn, B. Simon, R. Ishikawa and R. J. Henry 2014b. Chloroplast Genome of Novel Rice Germplasm Identified in Northern Australia. Trop Plant Biol 7: 111-120. doi: 10.1007/s12042-014-9142-8

Brozynska, Marta, Dario Copetti, Agnelo Furtado, Rod A Wing, Darren Crayn, Glen Fox, Ryuji Ishikawa and Robert J Henry 2017. Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice. Plant Biotechnol J 15: 765-774.

Burge, Chris and Samuel Karlin 1997. Prediction of complete gene structures in human genomic DNA1. Journal of molecular biology 268: 78-94.

Burge, Christopher B and Samuel Karlin 1998. Finding the genes in genomic DNA. Current opinion in structural biology 8: 346-354.

Butardo, Vito M, Roslen Anacleto, Sabiha Parween, Irene Samson, Krishna de Guzman, Crisline Mae Alhambra, Gopal Misra and Nese Sreenivasulu 2016. Systems genetics identifies a novel regulatory domain of amylose synthesis. Plant Physiol: pp. 01248.02016.

Cai, Xiu-Ling, Zong-Yang Wang, Yan-Yan Xing, Jing-Liu Zhang and Meng-Min Hong 1998. Aberrant splicing of intron 1 leads to the heterogeneous 5′ UTR and decreased expression of waxy gene in rice cultivars of intermediate amylose content. The Plant Journal 14: 459-465.

Calingacion, Mariafe, Alice Laborte, Andrew Nelson, Adoracion Resurreccion, Jeanaflor Crystal Concepcion, Venea Dara Daygon, Roland Mumm, Russell Reinke, Sharifa Dipti and Priscila Zaczuk Bassinello 2014. Diversity of global rice markets and the science required for consumer-targeted rice breeding. PLoS ONE 9: e85106.

Carbonell-Caballero, Jose, Roberto Alonso, Victoria Ibañez, Javier Terol, Manuel Talon and Joaquin Dopazo 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus Citrus. Molecular biology and evolution 32.

Cartegni, Luca, Shern L Chew and Adrian R Krainer 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nature Reviews Genetics 3: 285.

Cartegni, Luca, Michelle L Hastings, John A Calarco, Elisa de Stanchina and Adrian R Krainer 2006. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. The American Journal of Human Genetics 78: 63-77.

Cartegni, Luca and Adrian R Krainer 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nature Genetics 30: 377.

Chen, Chao, Shan Gao, Qing Sun, Yuling Tang, Yuhao Han, Jinkun Zhang and Zhipeng Li 2017. Induced splice site mutation generates alternative intron splicing in starch synthase II (SSII) gene in rice. Biotechnology & Biotechnological Equipment 31: 1093-1099.

Chen, J., Q. Huang, D. Gao, J. Wang, Y. Lang, T. Liu, B. Li, Z. Bai, J. Luis Goicoechea, C. Liang, C. Chen, W. Zhang, S. Sun, Y. Liao, X. Zhang, L. Yang, C. Song, M. Wang, J. Shi, G. Liu, J. Liu, H. Zhou, W. Zhou, Q. Yu, N. An, Y. Chen, Q. Cai, B. Wang, B. Liu, J. Min, Y. Huang, H. Wu, Z. Li, Y. Zhang, Y. Yin, W. Song, J. Jiang, S. A. Jackson, R. A. Wing, J. Wang and M. Chen 2013. Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. Nat Commun 4: 1595. doi: 10.1038/ncomms2596

Chen, Ming-Hsuan, Christine J Bergman, Shannon RM Pinson and Robert G Fjellstrom 2008a. Waxy gene haplotypes: associations with pasting properties in an international rice germplasm collection. Journal of cereal science 48: 781-788.

Chen, Ming-Hsuan, Christine Bergman, Shannon Pinson and Robert Fjellstrom 2008b. Waxy gene haplotypes: Associations with apparent amylose content and the effect by the environment in an international rice germplasm collection. Journal of cereal science 47: 536-545.

Cheng, Jun, Muhammad Awais Khan, Wen-Ming Qiu, Jing Li, Hui Zhou, Qiong Zhang, Wenwu Guo, Tingting Zhu, Junhua Peng and Fengjie Sun 2012. Diversification of genes encoding granule-bound starch synthase in monocots and dicots is marked by multiple genome-wide duplication events. PLoS ONE 7: e30088.

Choi, Jae Young, Adrian E Platts, Dorian Q Fuller, Rod A Wing and Michael D Purugganan 2017. The rice paradox: Multiple origins but single domestication in Asian rice. Molecular biology and evolution 34: 969-979.

Civáň, Peter, Hayley Craig, Cymon J. Cox and Terence A. Brown 2015. Three geographically separate domestications of Asian rice. Nature Plants 1: 15164. doi: 10.1038/nplants.2015.164

Cooper, Thomas A, Lili Wan and Gideon Dreyfuss 2009. RNA and disease. Cell 136: 777-793.

Dal Bosco, Cristina, Lina Lezhneva, Alexander Biehl, Dario Leister, Heinrich Strotmann, Gerd Wanner and Jörg Meurer 2003. Inactivation of the chloroplast ATP synthase γ subunit results in high non-photochemical fluorescence quenching and altered nuclear gene expression in Arabidopsis thaliana. Journal of Biological Chemistry.

Daniell, Henry, Choun-Sea Lin, Ming Yu and Wan-Jung Chang 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome biology 17: 1.

De Vicente, M, C López and T Fulton 2004. Genetic diversity analysis with molecular marker data: Learning module. International Plant Genetic Resources Institute (IPGRI).

Dian, Weimin, Huawu Jiang, Qingshuang Chen, Feiyang Liu and Ping Wu 2003. Cloning and characterization of the granule-bound starch synthase II gene in rice: gene expression is regulated by the nitrogen level, sugar and circadian rhythm. Planta 218: 261-268.

Dobo, Macaire, Nicolas Ayres, Grace Walker and Williams D Park 2010. Polymorphism in the GBSS gene affects amylose content in US and European rice germplasm. Journal of cereal science 52: 450-456.

Doebley, J. F., B. S. Gaut and B. D. Smith 2006. The molecular genetics of crop domestication. Cell 127: 1309-1321. doi: 10.1016/j.cell.2006.12.006

Duan, J. and W. Cai 2012. OsLEA3-2, an abiotic stress induced gene of rice plays a key role in salt and drought tolerance. PLoS ONE 7: e45117. doi: 10.1371/journal.pone.0045117

Duitama, J., A. Silva, Y. Sanabria, D. F. Cruz, C. Quintero, C. Ballen, M. Lorieux, B. Scheffler, A. Farmer, E. Torres, J. Oard and J. Tohme 2015. Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. PLoS ONE 10: e0124617. doi: 10.1371/journal.pone.0124617

FAO 2015. <Rice Market Monitor October 2015. FAO XVIII.

Feltus, F Alex, Jun Wan, Stefan R Schulze, James C Estill, Ning Jiang and Andrew H Paterson 2004. An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res 14: 1812-1819.

Flowers, J. M., J. Molina, S. Rubinstein, P. Huang, B. A. Schaal and M. D. Purugganan 2012. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. Mol Biol Evol 29: 675-687. doi: 10.1093/molbev/msr225

Fu, Qiang, Peijiang Zhang, Lubin Tan, Zuofeng Zhu, Dan Ma, Yongcai Fu, Xinchun Zhan, Hongwei Cai and Chuanqing Sun 2010. Analysis of QTLs for yield-related traits in Yuanjiang common wild rice (Oryza rufipogon Griff.). Journal of Genetics and Genomics 37: 147-157.

Fuchs, Eric J, Allan Meneses Martínez, Amanda Calvo, Melania Muñoz and Griselda Arrieta-Espinoza 2016. Genetic diversity in Oryza glumaepatula wild rice populations in Costa Rica and possible gene flow from O. sativa. PeerJ 4: e1875.

Fujita, Naoko, Rui Satoh, Aki Hayashi, Momoko Kodama, Rumiko Itoh, Satomi Aihara and Yasunori Nakamura 2011. Starch biosynthesis in rice endosperm requires the presence of either starch synthase I or IIIa. Journal of experimental botany 62: 4819-4831.

Fujita, Naoko, Yoshiko Toyosawa, Yoshinori Utsumi, Toshiyuki Higuchi, Isao Hanashiro, Akira Ikegami, Sayuri Akuzawa, Mayumi Yoshida, Akiko Mori and Kotaro Inomata 2009. Characterization of pullulanase (PUL)-deficient mutants of rice (Oryza sativa L.) and the function of PUL on starch biosynthesis in the developing rice endosperm. Journal of experimental botany 60: 1009-1023.

Furtado, A. 2014. DNA extraction from vegetative tissue for next-generation sequencing. Methods Mol Biol 1099: 1-5. doi: 10.1007/978-1-62703-715-0_1

Garaycochea, S., P. Speranza and F. Alvarez-Valin 2015. A strategy to recover a high-quality, complete plastid sequence from low-coverage whole-genome sequencing. Appl Plant Sci 3. doi: 10.3732/apps.1500022

Garris, Amanda J, Thomas H Tai, Jason Coburn, Steve Kresovich and Susan McCouch 2005. Genetic structure and diversity in Oryza sativa L. Genetics 169: 1631-1638.

Gelfman, Sahar, Noa Cohen, Ahuvi Yearim and Gil Ast 2013. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon–intron structure. Genome Res 23: 789-799.

Gnanamanickam, Samuel S. 2009. Rice and Its Importance to Human Life. In Biological Control of Rice Diseases, 1-11. Dordrecht: Springer Netherlands.

Gross, B. L. and K. M. Olsen 2010. Genetic perspectives on crop domestication. Trends Plant Sci 15: 529-537. doi: 10.1016/j.tplants.2010.05.008

Gross, B. L. and Z. Zhao 2014. Archaeological and genetic insights into the origins of domesticated rice. Proc Natl Acad Sci U S A 111: 6190-6197. doi: 10.1073/pnas.1308942110

Groves, R. H., R. S. Hill, E. A. Kellogg, M. Lazarides, H. P. Linder, A. McCusker, T. D. Macfarlane, M. K. Macphail, M. Nightingale, S. Renvoize, B. K. Simon, R. Sinclair, L. Watson, C. M. Weiller and R. D. B. Whalley 2009. Flora of Australia: Volume 44A : Poaceae 2: Australian Biological Resources Study CSIRO Publishing.

Guindon, Stéphane and Olivier Gascuel 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology 52: 696-704.

Gutell, Robin R and Robert K Jansen 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.

Hadiarto, T. and L. S. Tran 2011. Progress studies of drought-responsive genes in rice. Plant Cell Rep 30: 297-310. doi: 10.1007/s00299-010-0956-z

Hajjar, Reem and Toby Hodgkin 2007. The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156: 1-13.

He, Z., W. Zhai, H. Wen, T. Tang, Y. Wang, X. Lu, A. J. Greenberg, R. R. Hudson, C. I. Wu and S. Shi 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet 7: e1002100. doi: 10.1371/journal.pgen.1002100

Henry, Robert J 2016. Genomics strategies for germplasm characterization and the development of climate resilient crops. In Crop Breeding: Bioinformatics and Preparing for Climate Change, 3-10: CRC Press.

Henry, Robert J., Nicole Rice, Daniel L. E. Waters, Shabana Kasem, Ryuji Ishikawa, Yin Hao, Sally Dillon, Darren Crayn, Rod Wing and Duncan Vaughan 2010. Australian Oryza: Utility and Conservation. Rice 3: 235-241. doi: 10.1007/s12284-009-9034-y

Henry, Robert James 2009. Plant resources for food, fuel and conservation: Routledge.

Hirano, Hiro-Yuki, Mitsugu Eiguchi and Yoshio Sano 1998. A single base change altered the regulation of the Waxy gene at the posttranscriptional level during the domestication of rice. Molecular biology and evolution 15: 978-987.

Hirano, HY, M Eiguchi and Y Sano 1996. A point mutation, G to T, causes the differentiation of the Wx b allele from Wx a allele, which is specific to Japonica rice. Rice Genet. Newslett 13: 148-149.

Hirose, Tatsuro and Tomio Terao 2004. A comprehensive expression analysis of the starch synthase gene family in rice (Oryza sativa L.). Planta 220: 9-16.

Hoai, Tran Thi Thu, Hiroaki Matsusaka, Yoshiko Toyosawa, Tran Danh Suu, Hikaru Satoh and Toshihiro Kumamaru 2014. Influence of single-nucleotide polymorphisms in the gene encoding granule-bound starch synthase I on amylose content in Vietnamese rice cultivars. Breeding Science 64: 142-148.

Hollingsworth, Peter M., Laura L. Forrest, John L. Spouge, Mehrdad Hajibabaei, Sujeevan Ratnasingham, Michelle van der Bank, Mark W. Chase, Robyn S. Cowan, David L. Erickson, Aron J. Fazekas, Sean W. Graham, Karen E. James, Ki-Joong Kim, W. John Kress, Harald Schneider, Jonathan van AlphenStahl, Spencer C.H. Barrett, Cassio van den Berg, Diego Bogarin, Kevin S. Burgess, Kenneth M. Cameron, Mark Carine, Juliana Chacón, Alexandra Clark, James J. Clarkson, Ferozah Conrad, Dion S. Devey, Caroline S. Ford, Terry A.J. Hedderson, Michelle L. Hollingsworth, Brian C. Husband, Laura J. Kelly, Prasad R. Kesanakurti, Jung Sung Kim, Young-Dong Kim, Renaud Lahaye, Hae-Lim Lee, David G. Long, Santiago Madriñán, Olivier Maurin, Isabelle Meusnier, Steven G. Newmaster, Chong-Wook Park, Diana M. Percy, Gitte Petersen, James E. Richardson, Gerardo A. Salazar, Vincent Savolainen, Ole Seberg, Michael J. Wilkinson, Dong-Keun Yi and Damon P. Little 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106: 12794-12797. doi: 10.1073/pnas.0905845106

Hori, Y, R Fujimoto, Y Sato and T Nishio 2007. A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (Oryza sativa). Theoretical and Applied Genetics 115: 217-224.

Hu, E. A., A. Pan, V. Malik and Q. Sun 2012. White rice consumption and risk of type 2 diabetes: meta-analysis and systematic review. BMJ 344: e1454. doi: 10.1136/bmj.e1454

Hu, Y., B. Mao, Y. Peng, Y. Sun, Y. Pan, Y. Xia, X. Sheng, Y. Li, L. Tang, L. Yuan and B. Zhao 2014. Deep re-sequencing of a widely used maintainer line of hybrid rice for discovery of DNA polymorphisms and evaluation of genetic diversity. Mol Genet Genomics 289: 303-315. doi: 10.1007/s00438-013-0807-z

Hua, L., D. R. Wang, L. Tan, Y. Fu, F. Liu, L. Xiao, Z. Zhu, Q. Fu, X. Sun, P. Gu, H. Cai, S. R. McCouch and C. Sun 2015. LABA1, a Domestication Gene Associated with Long, Barbed Awns in Wild Rice. Plant Cell 27: 1875-1888. doi: 10.1105/tpc.15.00260

Huang, Wenda, Xueyong Zhao, Xin Zhao, Yulin Li and Jie Lian 2016. Effects of environmental factors on genetic diversity of Caragana microphylla in Horqin Sandy Land, northeast China. Ecology and Evolution 6: 8256-8266.

Huang, X., N. Kurata, X. Wei, Z. X. Wang, A. Wang, Q. Zhao, Y. Zhao, K. Liu, H. Lu, W. Li, Y. Guo, Y. Lu, C. Zhou, D. Fan, Q. Weng, C. Zhu, T. Huang, L. Zhang, Y. Wang, L. Feng, H. Furuumi, T. Kubo, T. Miyabayashi, X. Yuan, Q. Xu, G. Dong, Q. Zhan, C. Li, A. Fujiyama, A. Toyoda, T. Lu, Q. Feng, Q. Qian, J. Li and B. Han 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497-501. doi: 10.1038/nature11532

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, M. Li, D. Fan, Y. Guo, A. Wang, L. Wang, L. Deng, W. Li, Y. Lu, Q. Weng, K. Liu, T. Huang, T. Zhou, Y. Jing, W. Li, Z. Lin, E. S. Buckler, Q. Qian, Q. F. Zhang, J. Li and B. Han 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42: 961-967. doi: 10.1038/ng.695

Huelsenbeck, John P. and Fredrik Ronquist 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

Inaba, Takehito and Danny J Schnell 2008. Protein trafficking to plastids: one theme, many variations. Biochemical Journal 413: 15-28.

Isshiki, Masayuki, Kazuko Morino, Midori Nakajima, Ron J Okagaki, Susan R Wessler, Takeshi Izawa and Ko Shimamoto 1998. A naturally occurring functional allele of the rice waxy locus has a GT to TT mutation at the 5′ splice site of the first intron. The Plant Journal 15: 133-138.

Izawa, T., S. Konishi, A. Shomura and M. Yano 2009. DNA changes tell us about rice domestication. Curr Opin Plant Biol 12: 185-192. doi: 10.1016/j.pbi.2009.01.004

Jeong, I. S., U. H. Yoon, G. S. Lee, H. S. Ji, H. J. Lee, C. D. Han, J. H. Hahn, G. An and T. H. Kim 2013. SNP-based analysis of genetic diversity in anther-derived rice by whole genome sequencing. Rice (N Y) 6: 6. doi: 10.1186/1939-8433-6-6

Jian, Xueqiu, Eric Boerwinkle and Xiaoming Liu 2013. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genetics in Medicine 16: 497-503.

JinHua, Xiao, S. Grandillo, S. N.; McCouch Ahn, S. R., S. D.; Li JiMing Tanksley and Yuan LongPing 1996. Genes from wild rice improve yield. Nature 384: 223-224. doi: 10.1038/384223a0

Joseph, L., P. Kuriachan and G. Thomas 2008. Is Oryza malampuzhaensis Krish. et Chand. (Poaceae) a valid species? Evidence from morphological and molecular analyses. Plant Systematics and Evolution 270: 75-94. doi: 10.1007/s00606-007-0606-2

Karp, Angela, OLE Seberg and Marcello Buiatti 1996. Molecular techniques in the assessment of botanical diversity. Annals of Botany 78: 143-149.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma and Takashi Miyata 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066.

Kawahara, Y., M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell and T. Matsumoto 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6: 4. doi: 10.1186/1939-8433-6-4

Khan, Mudasir Hafiz, Zahoor Ahmad Dar and Sher Ahmad Dar 2015. Breeding Strategies for Improving Rice Yield—A Review. Agricultural Sciences 06: 467-478. doi: 10.4236/as.2015.65046

Kharabian-Masouleh, A., D. L. Waters, R. F. Reinke, R. Ward and R. J. Henry 2012. SNP in starch biosynthesis genes associated with nutritional and functional properties of rice. Sci Rep 2: 557. doi: 10.1038/srep00557

Kiefer, Florian, Konstantin Arnold, Michael Künzli, Lorenza Bordoli and Torsten Schwede 2008. The SWISS-MODEL Repository and associated resources. Nucleic Acids Research 37: D387-D392.

Kiegle, Edward A., Alex Garden, Elia Lacchini and Martin M. Kater 2018. A Genomic View of Alternative Splicing of Long Non-coding RNAs during Rice Seed Development Reveals Extensive Splicing and lncRNA Gene Families. Frontiers in plant science 9. doi: 10.3389/fpls.2018.00115

Kim, Backki, Dong-Gwan Kim, Gileung Lee, Jeonghwan Seo, Ik-Young Choi, Beom-Soon Choi, Tae-Jin Yang, Kwang Soo Kim, Joohyun Lee and Joong Hyoun Chin 2014a. Defining the genome structure of 'Tongil'rice, an important cultivar in the Korean "Green Revolution". Rice 7: 22.

Kim, HyunJung, Eung Gi Jeong, Sang-Nag Ahn, Jeffrey Doyle, Namrata Singh, Anthony J Greenberg, Yong Jae Won and Susan R McCouch 2014b. Nuclear and chloroplast diversity and phenotypic distribution of rice (Oryza sativa L.) germplasm from the democratic people's republic of Korea (DPRK; North Korea). Rice 7: 1.

Kim, HyunJung, Janelle Jung, Namrata Singh, Anthony Greenberg, Jeff J Doyle, Wricha Tyagi, Jong-Wook Chung, Jennifer Kimball, Ruaraidh Sackville Hamilton and Susan R McCouch 2016. Population dynamics among six major groups of the Oryza rufipogon species complex, wild relative of cultivated Asian rice. Rice 9: 56.

Kim, K., S. C. Lee, J. Lee, Y. Yu, K. Yang, B. S. Choi, H. J. Koh, N. E. Waminal, H. I. Choi, N. H. Kim, W. Jang, H. S. Park, J. Lee, H. O. Lee, H. J. Joh, H. J. Lee, J. Y. Park, S. Perumal, M. Jayakodi, Y. S. Lee, B. Kim, D. Copetti, S. Kim, S. Kim, K. B. Lim, Y. D. Kim, J. Lee, K. S. Cho, B. S. Park, R. A. Wing and T. J. Yang 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. Sci Rep 5: 15655. doi: 10.1038/srep15655

Koh, Hee-Jong, Suk-Yoon Kwon and Michael Thomson 2015. Current Technologies in Plant Molecular Breeding: Springer.

Kole, Chittaranjan 2011. Wild Crop Relatives: Genomic and Breeding Resources: Cereals: Springer Science & Business Media.

Krishnan, S. G., D. L. Waters and R. J. Henry 2014. Australian wild rice reveals pre-domestication origin of polymorphism deserts in rice genome. PLoS ONE 9: e98843. doi: 10.1371/journal.pone.0098843

Kumagai, Masahiko, Masaaki Kanehara, Shin'ya Shoda, Saburo Fujita, Shizuo Onuki, Shintaroh Ueda and Li Wang 2016. Rice varieties in archaic East Asia: reduction of its diversity from past to present times. Molecular biology and evolution 33: 2496-2505.

Larkin, Patrick D and William D Park 2003. Association of waxy gene single nucleotide polymorphisms with starch characteristics in rice (Oryza sativa L.). Molecular Breeding 12: 335-339.

Larkin, Patrick D and William D Park 1999. Transcript accumulation and utilization of alternate and non-consensus splice sites in rice granule-bound starch synthase are temperature-sensitive and controlled by a single-nucleotide polymorphism. Plant Mol Biol 40: 719-727.

Li, Changbao, Ailing Zhou and Tao Sang 2006. Rice domestication by reducing shattering. Science 311: 1936-1939.

Li, Dejun, Chuanqing Sun, Yongcai Fu, Cheng Li, Zuofeng Zhu, Liang Chen, Hongwei Cai and Xiangkun Wang 2002. Identification and mapping of genes for improving yield from Chinese common wild rice (O. rufipogon Griff.) using advanced backcross QTL analysis. Chinese Science Bulletin 47: 1533-1537.

Li, J. Y., J. Wang and R. S. Zeigler 2014a. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience 3: 8. doi: 10.1186/2047-217X-3-8

Li, Jin Quan and Peng Zhang 2012. Assessment and Utilization of the Genetic Diversity in Rice (Orysa Sativa L.): INTECH Open Access Publisher.

Li, Nannan 2012. Characterization of two chloroplast envelope membrane proteins, lmu.

Li, Xiao-yan, Sheng Qiang, Xiao-ling Song, Kun Cai, Yi-na Sun, Zhi-hua Shi and Wei-min Dai 2014b. Allele Types of Rc Gene of Weedy Rice from Jiangsu Province, China. Rice Science 21: 252-261. doi: 10.1016/s1672-6308(13)60183-3

Lin, Z., M. E. Griffith, X. Li, Z. Zhu, L. Tan, Y. Fu, W. Zhang, X. Wang, D. Xie and C. Sun 2007. Origin of seed shattering in rice (Oryza sativa L.). Planta 226: 11-20. doi: 10.1007/s00425-006-0460-4

Liu, Linglong, Xiaodong Ma, Shijia Liu, Changlan Zhu, Ling Jiang, Yihua Wang, Yi Shen, Yulong Ren, Hui Dong and Liangming Chen 2009. Identification and characterization of a novel Waxy allele from a Yunnan rice landrace. Plant Mol Biol 71: 609-626.

Liu, Rong, Xiao-Ming Zheng, Lian Zhou, Hai-Fei Zhou and Song Ge 2015. Population genetic structure of Oryza rufipogon and Oryza nivara: implications for the origin of O. nivara. Molecular ecology 24: 5211-5228.

Lohse, Marc, Oliver Drechsel, Sabine Kahlau and Ralph Bock 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Research: gkt289.

Londo, J. P., Y. C. Chiang, K. H. Hung, T. Y. Chiang and B. A. Schaal 2006. Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice, Oryza sativa. Proc Natl Acad Sci U S A 103: 9578-9583. doi: 10.1073/pnas.0603152103

Lu, J., T. Tang, H. Tang, J. Huang, S. Shi and C. I. Wu 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet 22: 126-131. doi: 10.1016/j.tig.2006.01.004

Luco, Reini F, Mariano Allo, Ignacio E Schor, Alberto R Kornblihtt and Tom Misteli 2011. Epigenetics in alternative pre-mRNA splicing. Cell 144: 16-26.

Lynch, Eric D, Ming K Lee, Jan E Morrow, Piri L Welcsh, Pedro E León and Mary-Claire King 1997. Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the Drosophila gene diaphanous. Science 278: 1315-1318.

Maor, Galit Lev, Ahuvi Yearim and Gil Ast 2015. The alternative role of DNA methylation in splicing regulation. Trends in Genetics 31: 274-280.

Marroni, F., S. Pinosio and M. Morgante 2014. Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18: 31-36. doi: 10.1016/j.pbi.2014.01.003

Matsuoka, Yoshihiro, Yukiko Yamazaki, Yasunari Ogihara and Koichiro Tsunewaki 2002. Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. Molecular biology and evolution 19: 2084-2091.

Matsushita, S, T Kurakazu, Doi K Sobrizal and A Yoshimura 2003. Mapping of genes for awn in rice using Oryza meridionalis introgression lines. Rice Genet Newsl 20: 17.

Meyer, R. S. and M. D. Purugganan 2013. Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14: 840-852. doi: 10.1038/nrg3605

Mickelbart, M. V., P. M. Hasegawa and J. Bailey-Serres 2015. Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. Nat Rev Genet 16: 237-251. doi: 10.1038/nrg3901

Mikami, I, N Uwatoko, Y Ikeda, J Yamaguchi, HY Hirano, Y Suzuki and Y Sano 2008. Allelic diversification at the wx locus in landraces of Asian rice. Theoretical and Applied Genetics 116: 979-989.

Mishra, Ratnesh Chandra, Amanjot Singh, Lalit Dev Tiwari and Anil Grover 2016. Characterization of 5′ UTR of rice ClpB-C/Hsp100 gene: evidence of its involvement in post-transcriptional regulation. Cell Stress and Chaperones 21: 271-283.

Molina, J., M. Sikora, N. Garud, J. M. Flowers, S. Rubinstein, A. Reynolds, P. Huang, S. Jackson, B. A. Schaal, C. D. Bustamante, A. R. Boyko and M. D. Purugganan 2011. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A 108: 8351-8356. doi: 10.1073/pnas.1104686108

Momma, Mitsuru and Zui Fujimoto 2012. Interdomain disulfide bridge in the rice granule bound starch synthase I catalytic domain as elucidated by X-ray structure analysis. Bioscience, biotechnology, and biochemistry 76: 1591-1595.

Mondini, Linda, Arshiya Noorani and Mario A. Pagnotta 2009. Assessing Plant Genetic Diversity by Molecular Tools. Diversity 1: 19-35. doi: 10.3390/d1010019

Moner , Agnelo Furtado, Ian Chivers, Glen Fox, Darren Crayn and Robert J. Henry 2018. Diversity and Evolution of Rice Progenitors in Australia Ecology and Evolution accepted

Ng, NQ, JG Hawkes, JT Williams and TT Chang 1981. The recognition of a new species of rice (Oryza) from Australia. Botanical journal of the Linnean Society 82: 327-330.

Nock, Catherine J, Daniel LE Waters, Mark A Edwards, Stirling G Bowen, Nicole Rice, Giovanni M Cordeiro and Robert J Henry 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9: 328-333.

Pérez, Serge and Eric Bertoft 2010. The molecular structures of starch components and their contribution to the architecture of starch granules: A comprehensive review. Starch‐Stärke 62: 389-420.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Sanyal, H. Kim, K. Collura, D. S. Brar, S. Jackson, R. A. Wing and O. Panaud 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res 16: 1262-1269. doi: 10.1101/gr.5290206

Prathepha, Preecha 2007. Identification of variant transcripts of waxy gene in non-glutinous rice (O. sativa L.) with different amylose content. Pakistan Journal of Biological Sciences 10: 2500-2504.

Price, Morgan N, Paramvir S Dehal and Adam P Arkin 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution 26: 1641-1650.

Rathinasabapathi, P., N. Purushothaman, V. L. Ramprasad and M. Parani 2015. Whole genome sequencing and analysis of Swarna, a widely cultivated indica rice variety with low glycemic index. Sci Rep 5: 11303. doi: 10.1038/srep11303

Ravi, V., J. P. Khurana, A. K. Tyagi and P. Khurana 2008. An update on chloroplast genomes. Plant Systematics and Evolution 271: 101-122. doi: 10.1007/s00606-007-0608-0

Reed, David H. and Richard Frankham 2003. Correlation between Fitness and Genetic Diversity

Correlación entre Adaptabilidad y Diversidad Genética. Conservation Biology 17: 230-237. doi: 10.1046/j.1523-1739.2003.01236.x

Rozas, Julio, Juan C Sánchez-DelBarrio, Xavier Messeguer and Ricardo Rozas 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

Salzberg, SL, DB Searls and S Kasif 1998. Modeling dependencies in pre-mRNA splicing signals. Computational methods in molecular biology 32: 129.

Sanchez, Paul L., Rod A. Wing and Darshan S. Brar 2013. The Wild Relative of Rice: Genomes and Genomics. 9-25. doi: 10.1007/978-1-4614-7903-1_2

Sang, T. 2009. Genes and mutations underlying domestication transitions in grasses. Plant Physiol 149: 63-70. doi: 10.1104/pp.108.128827

Sang, Tao and Song Ge 2007. The puzzle of rice domestication. Journal of Integrative Plant Biology 49: 760.

Schatz, Michael C, Lyza G Maron, Joshua C Stein, Alejandro H Wences, James Gurtowski, Eric Biggers, Hayan Lee, Melissa Kramer, Eric Antoniou and Elena Ghiban 2014. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome biology 15: 506.

Schroeder, H., A. M. Hoeltken and M. Fladung 2012. Differentiation of Populus species using chloroplast single nucleotide polymorphism (SNP) markers--essential for comprehensible and reliable poplar breeding. Plant Biol (Stuttg) 14: 374-381. doi: 10.1111/j.1438-8677.2011.00502.x

Shahid Masood, M., T. Nishikawa, S. Fukuoka, P. K. Njenga, T. Tsudzuki and K. Kadowaki 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. doi: 10.1016/j.gene.2004.06.008

Simpson, George Gaylord 1977. Too many lines; the limits of the Oriental and Australian zoogeographic regions. Proceedings of the American Philosophical Society 121: 107-120.

Singh, Nisha, Pawan Kumar Jayaswal, Kabita Panda, Paritra Mandal, Vinod Kumar, Balwant Singh, Shefali Mishra, Yashi Singh, Renu Singh and Vandna Rai 2015. Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. Scientific reports 5.

Singh, Nisha, Balwant Singh, Vandna Rai, Sukhjeet Sidhu, Ashok K Singh and Nagendra K Singh 2017. Evolutionary Insights Based on SNP Haplotypes of Red Pericarp, Grain Size and Starch Synthase Genes in Wild and Cultivated Rice. Frontiers in plant science 8: 972.

Song, Zhiping, BO Li, Jiakuan Chen and BAO-RONG LU 2005. Genetic diversity and conservation of common wild rice (Oryza rufipogon) in China. Plant Species Biology 20: 83-92.

Sotowa, M., K. Ootsuka, Y. Kobayashi, Y. Hao, K. Tanaka, K. Ichitani, J. M. Flowers, M. D. Purugganan, I. Nakamura, Y. I. Sato, T. Sato, D. Crayn, B. Simon, D. L. Waters, R. J. Henry and R. Ishikawa 2013. Molecular relationships between Australian annual wild rice, Oryza meridionalis, and two related perennial forms. Rice (N Y) 6: 26. doi: 10.1186/1939-8433-6-26

Srivastava, Ashish Kumar, Yuming Lu, Gaurav Zinta, Zhaobo Lang and Jian-Kang Zhu 2018. UTR-Dependent Control of Gene Expression in Plants. Trends in plant science: 248-259.

Srivastava, S. K., P. Wolinski and A. Pereira 2014. A strategy for genome-wide identification of gene based polymorphisms in rice reveals non-synonymous variation and functional genotypic markers. PLoS ONE 9: e105335. doi: 10.1371/journal.pone.0105335

Stamatakis, Alexandros 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Stamatakis, Alexandros, Paul Hoover and Jacques Rougemont 2008. A rapid bootstrap algorithm for the RAxML web servers. Systematic biology 57: 758-771.

Sun, Ai-Zhen and Fang-Qing Guo 2016. Chloroplast Retrograde Regulation of Heat Stress Responses in Plants. Frontiers in plant science 7.

Sun, CQ, XK Wang, ZC Li, A Yoshimura and N Iwata 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.

Swofford, David L 2003. {PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.}.

Tang, Jiabin, Hong'ai Xia, Mengliang Cao, Xiuqing Zhang, Wanyong Zeng, Songnian Hu, Wei Tong, Jun Wang, Jian Wang and Jun Yu 2004. A comparison of rice chloroplast genomes. Plant Physiol 135: 412-420.

Terada, Rie, Midori Nakajima, Masayuki Isshiki, Ron J Okagaki, Susan R Wessler and Ko Shimamoto 2000. Antisense waxy genes with highly active promoters effectively suppress waxy gene expression in transgenic rice. Plant and Cell Physiology 41: 881-888.

Thalapati, S., A. K. Batchu, S. Neelamraju and R. Ramanan 2012. Os11Gsk gene from a wild rice, Oryza rufipogon improves yield in rice. Funct Integr Genomics 12: 277-289. doi: 10.1007/s10142-012-0265-4

Tikapunya, Tiparat, Glen Fox, Agnelo Furtado and Robert Henry 2017a. Grain physical characteristic of the Australian wild rices. Plant Genetic Resources 15: 409-420.

Tikapunya, Tiparat, Wei Zou, Wenwen Yu, Prudence O Powell, Glen P Fox, Agnelo Furtado, Robert J Henry and Robert G Gilbert 2017b. Molecular structures and properties of starches of Australian wild rice. Carbohydrate Polymers 172: 213-222.

Tong, Wei, Qiang He, Xiao-Qiang Wang, Min-Young Yoon, Won-Hee Ra, Fengpeng Li, Jie Yu, Win Htet Oo, Sun-Kyung Min and Bu-Woong Choi 2015. A chloroplast variation map generated using whole genome re-sequencing of Korean landrace rice reveals phylogenetic relationships among Oryza sativa subspecies. Biological Journal of the Linnean Society.

Tong, Wei, Tae-Sung Kim and Yong-Jin Park 2016. Rice Chloroplast Genome Variation Architecture and Phylogenetic Dissection in Diverse Oryza Species Assessed by Whole-Genome Resequencing. Rice 9: 57.

Ullah, Fahad, Michael Hamilton, Anireddy SN Reddy and Asa Ben-Hur 2018. Exploring the relationship between intron retention and chromatin accessibility in plants. BMC genomics 19: 21.

Utani, Dwinita W, Sugiono Moeljopawiro, Hajrial Aswidinnoor, Asep Setiawan and Ida Hanarida 2008. Blast resistance genes in wild rice Oryza rufipogon and rice cultivar IR64 [online]. Indonesian Journal of Agriculture. 1: 71-76.

Vaughan, Duncan A., Bao-Rong Lu and Norihiko Tomooka 2008. The evolving story of rice evolution. Plant Science 174: 394-408. doi: 10.1016/j.plantsci.2008.01.016

Wallace, Douglas C. 2016. Genetics: Mitochondrial DNA in evolution and disease. Nature 535: 498-500. doi: 10.1038/nature18902

Wambugu, P. W., M. Brozynska, A. Furtado, D. L. Waters and R. J. Henry 2015. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep 5: 13957. doi: 10.1038/srep13957

Wambugu, Peterson W, Agnelo Furtado, Daniel LE Waters, Desterio O Nyamongo and Robert J Henry 2013. Conservation and utilization of African Oryza genetic resources. Rice 6: 1.

Wang, Bing-Bing and Volker Brendel 2006. Genomewide comparative analysis of alternative splicing in plants. Proceedings of the National Academy of Sciences 103: 7175-7180.

Wang, J., H. Xu, N. Li, F. Fan, L. Wang, Y. Zhu and S. Li 2015. Artificial Selection of Gn1a Plays an Important role in Improving Rice Yields Across Different Ecological Regions. Rice (N Y) 8: 37. doi: 10.1186/s12284-015-0071-4

Wang, Yu, Hongping Chang, Shuai Hu, Xiutao Lu, Congying Yuan, Chen Zhang, Ping Wang, Wenjun Xiao, Langtao Xiao and Gang-Ping Xue 2014. Plastid casein kinase 2 knockout reduces abscisic acid (ABA) sensitivity, thermotolerance, and expression of ABA-and heat-stress-responsive nuclear genes. Journal of experimental botany: eru190.

Waters, D. L., C. J. Nock, R. Ishikawa, N. Rice and R. J. Henry 2012. Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. Ecol Evol 2: 211-217. doi: 10.1002/ece3.66

Weir, Bruce S. 1996. Genetic data analysis II: methods for discrete population genetic data. Sunderland, Mass: Sinauer Associates.

Wu, Z. and S. Ge 2014. The whole chloroplast genome of wild rice (Oryza australiensis). Mitochondrial DNA 27: 1062-1063. doi: 10.3109/19401736.2014.928868

Wurm, P, Campbell, L, Batten, GD, Bellairs, SM 2012. Australian native rice: A new sustainable wild food enterprise. Research Project No PRJ000347/Publication No 10/175.

Xie, X., J. Molina, R. Hernandez, A. Reynolds, A. R. Boyko, C. D. Bustamante and M. D. Purugganan 2011. Levels and patterns of nucleotide variation in domestication QTL regions on rice chromosome 3 suggest lineage-specific selection. PLoS ONE 6: e20670. doi: 10.1371/journal.pone.0020670

Xu, Changcheng, Jilian Fan, John E Froehlich, Koichiro Awai and Christoph Benning 2005a. Mutation of the TGD1 chloroplast envelope protein affects phosphatidate metabolism in Arabidopsis. The Plant Cell 17: 3094-3110.

Xu, Jian-Hong, Nori Kurata, Masahiro Akimoto, Hisako Ohtsubo and Eiichi Ohtsubo 2005b. Identification and characterization of Australian wild rice strains of Oryza meridionalis and Oryza rufipogon by SINE insertion polymorphism. Genes & genetic systems 80: 129-134.

Xu, Xun, Xin Liu, Song Ge, Jeffrey D Jensen, Fengyi Hu, Xin Li, Yang Dong, Ryan N Gutenkunst, Lin Fang and Lei Huang 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature biotechnology 30: 105.

Xu, Yunbi 2010. Molecular plant breeding: Cabi.

Yan, Hong-Bo, Xiao-Xue Pan, Hua-Wu Jiang and Guo-Jiang Wu 2009. Comparison of the starch synthesis genes between maize and rice: copies, chromosome location and expression divergence. Theoretical and Applied Genetics 119: 815-825.

Yang, Ruifang, Chunlong Sun, Jianjiang Bai, Zhixiang Luo, Biao Shi, Jianming Zhang, Wengui Yan and Zhongze Piao 2012. A putative gene sbe3-rs for resistant starch mutated from SBE3 for starch branching enzyme in rice (Oryza sativa L.). PLoS ONE 7: e43026.

Yao, W., G. Li, H. Zhao, G. Wang, X. Lian and W. Xie 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16: 187. doi: 10.1186/s13059-015-0757-3

Yap, Sandra L 2010. Phylogeography and Demography of Common Plant Species from the Philippine Islands, The University of Michigan.

Ye, Yuzhen and Adam Godzik 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19: ii246-ii255.

Yearim, Ahuvi, Sahar Gelfman, Ronna Shayevitch, Shai Melcer, Ohad Glaich, Jan-Philipp Mallm, Malka Nissim-Rafinia, Ayelet-Hashahar S Cohen, Karsten Rippe and Eran Meshorer 2015. HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell reports 10: 1122-1134.

Yoon, D. B., K. H. Kang, H. J. Kim, H. G. Ju, S. J. Kwon, J. P. Suh, O. Y. Jeong and S. N. Ahn 2006. Mapping quantitative trait loci for yield components and morphological traits in an advanced backcross population between Oryza grandiglumis and the O. sativa japonica cultivar Hwaseongbyeo. Theor Appl Genet 112: 1052-1062. doi: 10.1007/s00122-006-0207-4

Yu, G., K. M. Olsen and B. A. Schaal 2011. Molecular evolution of the endosperm starch synthesis pathway genes in rice (Oryza sativa L.) and its wild ancestor, O. rufipogon L. Mol Biol Evol 28: 659-671. doi: 10.1093/molbev/msq243

Zhang, Fantao and Jiankun Xie 2014. Genes and QTLs Resistant to Biotic and Abiotic Stresses from Wild Rice and Their Applications in Cultivar Improvements. doi: 10.5772/56825

Zhang, L. B., Q. Zhu, Z. Q. Wu, J. Ross-Ibarra, B. S. Gaut, S. Ge and T. Sang 2009. Selection on grain shattering genes and rates of rice domestication. New Phytol 184: 708-720. doi: 10.1111/j.1469-8137.2009.02984.x

Zhang, Qifa and Rod A Wing 2013. Genetics and genomics of rice: Springer.

Zhang, Xiaoli, Nicolas Szydlowski, David Delvallé, Christophe D'Hulst, Martha G James and Alan M Myers 2008. Overlapping functions of the starch synthases SSII and SSIII in amylopectin biosynthesis in Arabidopsis. BMC Plant Biol 8: 96.

Zhang, Yachuan, Curtis Rempel and Qiang Liu 2014. Thermoplastic starch processing and characteristics—a review. Critical reviews in food science and nutrition 54: 1353-1370.

Zheng, Mengdi, Xiayan Liu, Shuang Liang, Shiying Fu, Yafei Qi, Jun Zhao, Jingxia Shao, Lijun An and Fei Yu 2016. Chloroplast translation initiation factors regulate leaf variegation and development. Plant Physiol: pp. 02040.02015.

Zhiguo, E, Lei Wang and Jianhua Zhou 2013. Splicing and alternative splicing in rice and humans. BMB reports 46: 439.

Zhou, Hongju, Lijun Wang, Guifu Liu, Xiangbing Meng, Yanhui Jing, Xiaoli Shu, Xiangli Kong, Jian Sun, Hong Yu and Steven M Smith 2016. Critical roles of soluble starch synthase SSIIIa and granule-bound starch synthase Waxy in synthesizing resistant starch in rice. Proceedings of the National Academy of Sciences 113: 12844-12849.

Zhu, Youyong, Hairu Chen, Jinghua Fan, Yunyue Wang, Yan Li, Jianbing Chen, JinXiang Fan, Shisheng Yang, Lingping Hu, Hei Leung, Tom W. Mew, Paul S. Teng, Zonghua Wang and Christopher C. Mundt 2000. Genetic diversity and disease control in rice. Nature 406: 718. doi: 10.1038/35021046

Zhu, Z., L. Tan, Y. Fu, F. Liu, H. Cai, D. Xie, F. Wu, J. Wu, T. Matsumoto and C. Sun 2013. Genetic control of inflorescence architecture during rice domestication. Nat Commun 4: 2200. doi: 10.1038/ncomms3200

Zoschke, Reimo, Karsten Liere and Thomas Börner 2007. From seedling to mature plant: Arabidopsis plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development. The Plant Journal 50: 710-722.

Appendices

1 Appendix 1. *Oryza meridionalis*

1.1 Economic/Academic importance

Oryza meridionalis is an Australian wild rice in the AA genome group of close relatives of domesticated rice. The economic and academic interest in this species is associated with it being the most distant from domesticated rice of the species within the AA genome group making it an important resource for rice improvement and the study of rice evolution.

1.2 Brief botanical descriptions including distribution

Oryza meridionalis was described by Ng *et al*. in 1981. It is found across northern Australia from the Kimberley region in Western Australia to Queensland (Figure 20). *O. meridionalis* has also been reported from New Guinea. This is one of four *Oryza* species found in Australia (Henry et al., 2010).The description in the flora of Australia (Groves et al., 2009) includes the details provided in (Table 14) *O. meridionalis* is depicted in (Figure 21). It can be distinguished from other *Oryza* species found in northern Australia on the basis of the closed panicles and small anthers. *O. meridionalis* was originally described as an annual (Ng et al., 1981).

Table 14 Description of *Oryza meridionalis* (Groves et al., 2009)

Life cycle	Annual or perennial
Clums	0.3-2 m
Leaves	ligule 5-20(-30) mm, blade 6-47 cm long 4-14 mm wide
Panicles	9-30 cm long
Spikelets	6.5-10 mm long
Awn	(30-) 60-150 mm long
Anthers	1.3-2.5 (-3) mm long
Caryopsis	oblanceoloid or oboid-ellipsoidal laterally compressed (5-) 5.5-7.5 (-8.3) mm long

The presence of populations with similar appearance but apparent perennial habit led to some uncertainty about the identity of these perennial populations. *O. meridionalis* like plants were designated as Taxa B by (Brozynska et al., 2014b). Subsequent analysis (Moner et al 2017) has suggested that these are all part of one clade supporting the description of *O. meridionalis* as, an annual or perennial as in the Flora of Australia (Groves et al., 2009). (Julia et al. 2016) reported details of the morphology of some *ex situ* collections *of O. meridionalis*. Herbarium samples may be labelled *O. rufipogon* especially if collected before *O. meridionalis* was described.

1.3 Cytological details of genome including karyotype data

O. meridionalis is a diploid 2n=24.

1.4 Physiological studies

The grain physical traits (Kasem et al., 2010; Kasem et al., 2012; Tikapunya et al., 2016) and starch properties (Kasem et al., 2014; Tikapunya et al., 2017) have been investigated. Starch gene sequences were reported by Kasem et al. (2011*). O. meridionalis* has a high amylose content relative to domesticated rice.

1.5 Enumeration of sequences

The genome has been sequenced using Illumina and PacBio sequencing techniques (Brozynska et al., 2017) based upon 47.1 Gbp of shot gun Illumina sequence data and 15.0 Gbp of PacBio sequence data representing an estimated 127X and 41X coverage respectively of the estimated 370 Mbp genome.

1.6 Assembly

Brozynska et al. (2017) reported both hybrid (Illumina/PacBio) and Pac Bio only assemblies (Table 2). Hybrid assemblies covered 446 Mbp and PacBio alone, 355 Mbp.

1.7 Repetitive sequences

The most abundant group of transposable elements was found to be the Gypsy family representing almost 40% of all repeats with Copia elements accounting for 9.3% (Brozynska et al., 2017).

Table 15 Hybrid and PacBio assembly statistics calculated for scaffolds and contigs for hybrid assembly and for scaffolds only for PacBio assembly (Brozynska et al., 2017).

	Hybrid only	Pac-Bio
Assembler	Sparse Assembler + DBG2OLC	Celera Assembler
Number of scaffolds	4,718	3,242
Total size of scaffolds	446,369,637	354,906,376
Longest scaffold	2,079,733	3,232,522
Mean scaffold size	94,610	109,135
N50 scaffold length	163,003	159,640
Number of contigs	4808	
Total size of contigs	446,351,110	
Longest contig	1,449,836	
Median contig size	54,495	
N50 contig length	159,759	

1.8 Gene annotation

Bonskya et al. (2017) identified 21,169 protein encoding genes, and 5,624 non-coding RNA genes (including; 615 tRNA, 4,892 miRNA, 453 snoRNA, 87 sRNA and 129 rRNA).

1.9 Organelle genome

The complete chloroplast genome of *O. meridionalis* was reported by (Nock et al., 2014; Wambugu et al.,2015) used the whole chloroplast genome sequence to relate *O. meridionalis* to other taxa. Some of the variation in the chloroplast genome within the species has been explored (Waters et al., 2012; Brozynska et al., 2014a). The mitochondrial genome has not been reported.

1.10 Impact on plant breeding including pre-breeding work

Sanchez et al. (2013) produced hybrids between *O. sativa* and *O. meridionalis* that had heat and drought tolerance in extreme temperature conditions. Introgression from *O. meridionalis* into *japonica* cv. Taichung 65 lead to the identification of genes that control awn length on chromosomes 1, 4, 5. Awn length is controlled largely by a single dominant gene. However, other genes increase the expression and produce longer awns (Matsushita et al., 2003). Arbelaez et al. (2015) reported introgression lines with *O. sativa* cv Curinga as the recurrent parent.

1.11 Comparative genomics

Comparison of the genome with that of domesticated rice by mapping of sequence reads suggests that *O. meridionalis* has more diversity in regions of the genome that lack variation in the domesticated rice gene pool (Krishan et al., 2014).

1.12 Future prospects

O. meridionalis represents an important genetic resource for rice improvement providing a potential source of abiotic stress tolerance (Atwell et al., 2014) including heat tolerance (Scafaro et al., 2009; Scafaro et al., 2011; Scafaro et al., 2016). Photosynthesis traits may also be useful (Giuliani *et al*. 2013). Grain quality traits including starch properties may also add useful diversity to the rice gene pool. Further sequencing of this species will be of value (Henry, 2014) especially to explore diversity within the species. This resource will be important in developing rice for production in new or altered environments (Henry et al., 2016).

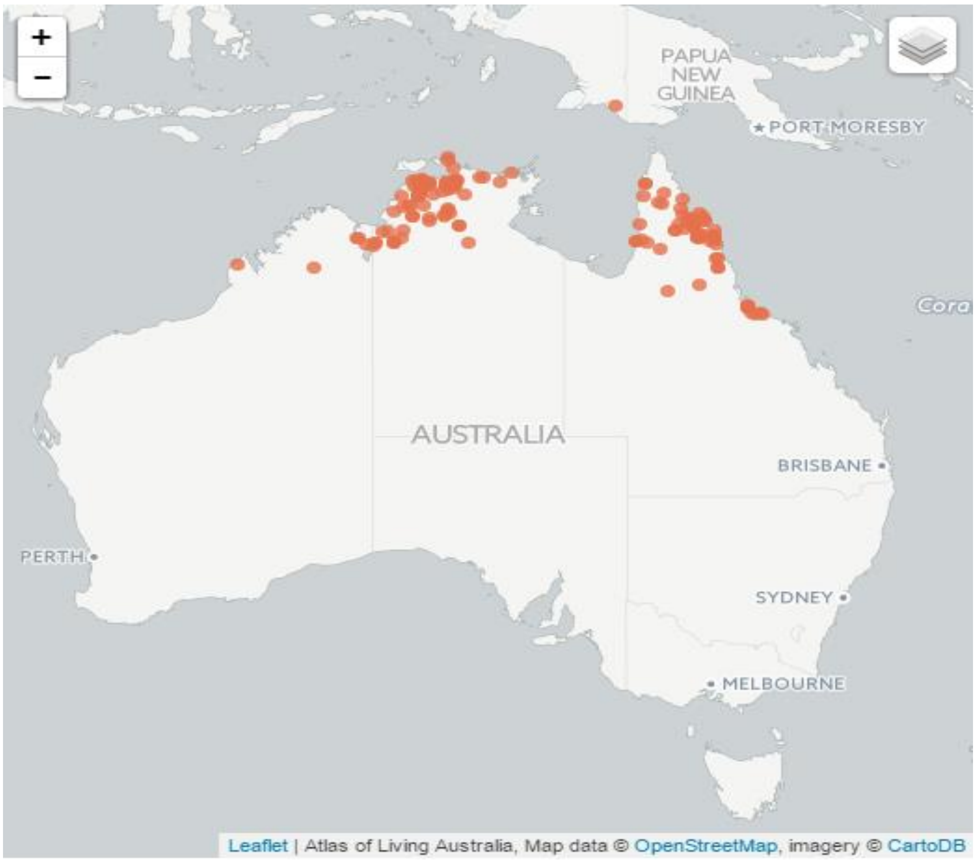

+
−
PAPUA NEW GUINEA
PORT MORESBY
Cora
AUSTRALIA
BRISBANE
PERTH
SYDNEY
MELBOURNE
Leaflet | Atlas of Living Australia, Map data © OpenStreetMap, imagery © CartoDB

Figure 20 Distrubution of O. meridionalis http://www.ala.org.au

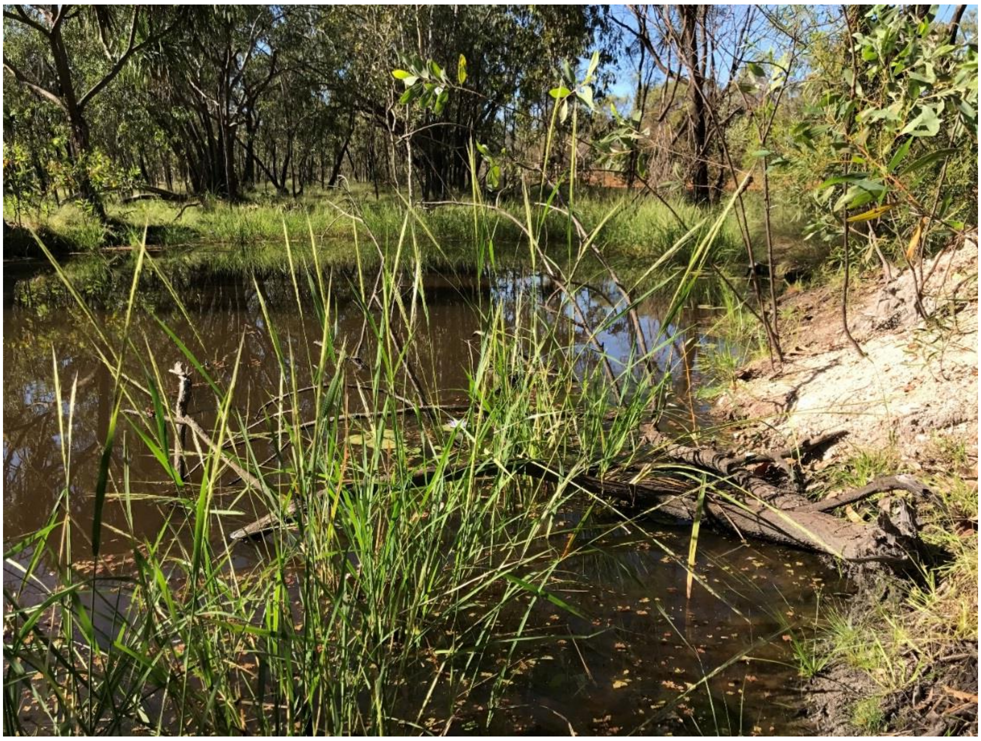


Figure 21 *Oryza* meridionalis in northern Australia.

2 Appendix 2.

Table 16 *Oryza* species the genome group, chromosome number and the geographical origin is provided for each species (Joseph & Thomas, 2008) and (Koh & Thomson, 2015).

	Oryza species	Genome group	Chr. number	Origin	Wild Domesticated
1	*O. officinalis* Wall ex. Watt	CC	24	Tropical Asia	Wild
2	*O. perennis*	AA	24		Wild
3	*O. punctata* Kotschy ex Steud.	BB, BBCC	24, 48	Philippines and Papua New Guinea	Wild
4	*O. rhizomatis* Vaughan	CC	24	Sri Lanka	Wild
5	*O. ridleyi* Hook	HHJJ	48	South Asia	Wild
6	*O. rufipogon* Griff.	AA	24	Tropical Asia	Wild
7	*O. sativa* ssp *japonica* and ssp *indica*	AA	24		Domesticated
8	*O. schlechteri* Pilger	HHKK	48	Papua New Guinea	Wild
9	*O. alta* Swallen	CCDD	48	South America	Wild
10	*O. australiensis* Domin.	EE	24	Tropical Australia	Wild
11	*O. barthii* Chev. et Roehr	AA	24	Africa	Wild
12	O. brachyantha Chev. et Roehr	FF	24	Africa	Wild
13	*O. coarctata* Roxb.	KKLL	48	India	Wild

14	*O. eichingeri* Peter	CC	24	South Asia and East Africa	Wild
15	*O. glaberrima*	AA	24	Africa	Domesticated
16	*O. glumaepatula* Steud. (*Oryza* glumaepatula)	AA	24	South and central America	Wild
17	*O. grandiglumis* Prod.	CCDD	48	South America	Wild
18	*O. granulata* Nees et Arn. ex. Watt	GG	24	Southeast Asia	Wild
19	*O. latifolia* Desv.	CCDD	48	South America	Wild
20	*O. longiglumis* Jansen	HHJJ	48	Indonesia	Wild
21	*Oryza* malampuzhaensis	BBCC	48	South India	Wild
22	*O. meridionalis* Ng	AA	24	Tropical Australia	Wild
23	*O. meyeriana* Baill	GG	24	Southeast Asia	Wild
24	*O. minuta* J.S. Presl. ex C.B. Presl.	BBCC	48	Philippines and PapuaNew Guinea	Wild
25	*O. nivara* Sharma et Shastry (*Oryza* sativa f. spontanea)	AA	24	Tropical Asia	Wild
26	*O. longistaminata* Chev. et Roehr (*Oryza* glumaepatula)	AA	24	Africa	Wild

Table 17 Chloroplasts sequences of *Oryza* spp. (http://www.ncbi.nlm.nih.gov/genome). Refseq, size, genes number and released date were demonstrated. Last update 15.1.2018

Organism Name	**BioProject**	**Size(Mb)**	**GC%**	**Replicons**	**tRNA**	**CDS**	**Genes**	**Release Date**	**Modify Date**
Oryza alta	PRJNA387897	0.13518	39	NC_034760.1/KF359913.1	37	87	132	24-May-17	24-May-17
Oryza australiensis	PRJNA256411	0.13522	38.95	NC_024608.1/KJ830774.1	38	83	129	29-Jul-14	29-Jul-14
Oryza barthii	PRJNA289787	0.13467	38.99	NC_027460.1/KM881634.1	33	82	123	14-Jul-15	14-Jul-15
Oryza brachyantha	PRJNA328726	0.1346	38.98	NC_030596.1/KT992850.1	38	83	129	12-Jul-16	19-Jul-17
Oryza eichingeri	PRJNA387861	0.13482	39	NC_034759.1/KF359912.1	37	87	132	24-May-17	24-May-17
Oryza glumipatula	PRJNA289804	0.13458	38.99	NC_027461.1/KM881640.1	33	83	124	14-Jul-15	14-Jul-15
Oryza grandiglumis	PRJNA387860	0.13515	38.99	NC_034761.1/KF359914.1	37	87	132	24-May-17	24-May-17
Oryza latifolia	PRJNA387768	0.13519	38.99	NC_034762.1/KF359915.1	37	87	132	24-May-17	24-May-17
Oryza longiglumis	PRJNA387852	0.13564	38.93	NC_034763.1/KF359918.1	37	87	132	24-May-17	24-May-17
Oryza longistaminata	PRJNA289799	0.13457	38.99	NC_027462.1/KM881641.1	33	83	124	14-Jul-15	14-Jul-15
Oryza meridionalis	PRJNA86637	0.13456	39.01	NC_016927.1/JN005831.1	41	75	124	28-Feb-12	28-Feb-12
Oryza meyeriana	PRJNA387854	0.13613	38.94	NC_034765.1/KF359921.1	37	86	131	24-May-17	24-May-17
Oryza minuta	PRJNA325260	0.13509	38.96	NC_030298.1/KU179220.1	39	89	138	10-Jun-16	10-Jun-16
Oryza nivara SL10	PRJNA12441	0.13449	39.01	NC_005973.1/AP006728.1	38	119	165	12-Jul-04	11-Mar-11
Oryza officinalis	PRJNA289798	0.13491	39	NC_027463.1/KM881643.1	33	83	124	14-Jul-15	14-Jul-15
Oryza punctata	PRJNA291899	0.1346	38.97	NC_027676.1/KM103375.1	41	100	149	4-Aug-15	4-Aug-15

Oryza rhizomatis	PRJNA387890	0.1348	39.01	NC_034758.1/KF359911.1	37	87	132	24-May-17	24-May-17
Oryza ridleyi	PRJNA387853	0.13573	38.92	NC_034764.1/KF359919.1	37	87	132	24-May-17	24-May-17
Oryza rufipogon	PRJNA162601	0.13454	39	NC_017835.1/JN005832.1	37	77	122	9-May-12	9-May-12
Oryza sativa	PRJNA291900	0.1345	39	NC_031333.1/KM103369.1	40	100	148	5-Oct-16	26-Jan-17
Oryza sativa indica Group	PRJNA17293	0.1345	39	NC_008155.1/AY522329.1	0	64	65	16-Jun-06	15-Apr-09
Oryza sativa indica Group	PRJNA368975	0.13455	39	NC_027678.1/KM103382.1	41	94	143	4-Aug-15	26-Jan-17
Oryza sativa indica Group	PRJNA318714	0.13455	39	Pltd: CP018170.1	0		0	4-May-17	4-May-17

Table 18 Comparison of chloroplast sequence generated by mapping and de novo procedures. Two different reference genomes were used. The degree of manual correction required for assembly and the final chloroplast size is given.

Accessions	Improved mapped sequence		Improved De novo sequence		Differences between Mapping and De novo seq. manual correction required	Final chloroplast sequence size bp
	Differences Vs *O. rufipogon*	Differences Vs O. *sativa* Nipponbare	Differences Vs *O. rufipogon*	Differences Vs O. *sativa* Nipponbare		
Z1W1743					Gaps	
Z1W1998	73	49	77	52	4	134,595
Z1W1782	80	54	81	55	1	134,595
Z1W1777	74	19	74	19	0	134,536
Z1W1683	74	19	74	19	0	134,536
Z1W2066	76	84	79	87	3	134,542
Z1W1804	70	42	72	44	2	134,582
Z2W0634	70	78	70	78	0	134,511
Z2W0628	68	41	70	43	2	134,583
Z2W1083	74	20	74	20	0	134,537

Z2W0153	76	82	77	83	1	134,484
Z2W1126	56	72	58	74	2	134,494
Z2W1096	72	17	74	19	2	134,536
Z3W3085	75	86	80	91	5	134,517
Z3W3091	78	87	84	94	Gaps	
Z3W3002	58	71	65	78	7	134,501
Z3W3052	72	83	78	89	6	134,516
Z3W3065	75	89	80	94	5	134,539
Z3W2331					Gaps	
Z4W0626	71	85	71	85	0	134,456
Z4W2308	80	22	80	22	0	134,553
Z4W1939	55	71	57	73	2	134,494
Z4W1554	59	72	60	74	2	134,495
Z4W1870	73	84	75	86	2	134,516
Z4W1854	81	24	86	29	5	134,116
Z4W2316	81	14	82	15	1	134,556

Z5W1236					Gaps	
Z5W1230	75	81	76	82	1	134,521
Z5W2078	126	127	126	127	0	134,553
Z5W2108	127	128	127	128	0	134,542
Z5W1975	57	71	59	73	2	134,495
Z5W1977	74	95	75	96	1	134,508
Z5W2024	60	73	62	76	2, 3	134,520
Z5W0576	61	77	62	78	1	134,502
Z5W1214	127	132	127	132	0	134,549
HP483_*indica*	61	77	62	78	1	134,502
HP179_*indica*	57	72	59	73	1,2	134,496
HP49_temperate_japonica	79	1	79	1	0	134,551
HP46_temperate_japonica	82	5	85	8	3	134,553
GP715_aus	78	23	78	23	0	134,534
GP706_tropical_japonica	80	13	85	15	2,5	134,556
GP294_aromatic	77	22	77	22	0	134,532

GP285_aus	61	74	62	75	1	134,540
GP284_aromatic	77	22	78	23	1	134,532
GP629_tropical_japonica					Gaps	

Table 19 Phylogenetic software tools applied to chloroplast genome analysis, analysis model and bootstrap number used in this study.

	Program	Analysing method	Substitution model	Rate variation	Bootstrapping	Out group	Options Chosen
1	Fast tree	Maximum likelihood	GTR	Gamma	-	-	Gamma20 likelihood
2	Garli	Maximum likelihood	-	-	-	-	Default setting
3	PHYLM	Maximum likelihood	GTR	-	1000	-	
4	MrBayes	Bayesian	GTR	Gamma	2000	*O. australiensis*	
5	RAxML	Maximum likelihood	GTR	Gamma	2000	-	rapid bootstrapping and search for the best-scoring ML tree

Table 20 SNP frequencies in each clade as described below

Clade group	SNPs	Clade group	SNPs	Clade group	SNPs	Clade group	SNPs
A	35	E	12	A,B,F,G	2	C2	1
B	28	F	35	A,F,G	6	E1	2
C	2	G	102	C,D	10	E2	2
D	2	A,B,F	1	C1	9	E3	1
C2	1	E4	1	C2	1	E4	1
E1	2	F,G	4	E1	2	F,G	4
E2	2	officinalis	8	E2	2	officinalis	8
E3	1	australiensis	2	E3	1	australiensis	2
total	265						

Clade A: W1214 Z5 Philippine, W2078 Z5 Australia, W2108 Z5 Australia, Australian taxa A and Australian taxa B

Clade B: *O. barthii1, O. barthii2, O. barthii3, O. barthii4* and *O. glaberrima*

Clade C: *O. nivara,* W0153 Z2 India, W0626 Z4 Burma, W0634 Z2 Burma, W1230 Z5 Papua New Guinea, W1554 Z4 Thailand, W1870 Z4 Thailand, W2024 Z5 Indonesia, W2066 Z1 Nepal, W3052 Z3 China, W3065 Z3 China and W3085 Z3 China

Clade D *O. sativa indica* JN861109.1, *O. sativa indica* NC_008155.1, W0576 Z5 Malaysia, W1126 Z2 India, W1939 Z4 Thailand, W1975 Z5 Indonesia and W3002 Z3 China

Clade E: *O. rufipogon* Asian1, *O. sativa japonica* NC_001320.1, *O. sativa subsp. japonica* Nipponbare GU592207.1, W0626 Z4 Burma, W1083 Z2 India, W1096 Z2 India, W1683 Z1 India, W1777 Z1 India, W1782 Z1 India, W1804 Z1 Sri Lanka, W1854 Z4 Thailand, W1998 Z1 India, W2308 Z4 Laos and W2316 Z4 Vietnam

Clade F: *O. glumipatula, O. longistaminata1 and O. longistaminata2*

Clade G*: O. officinalis* and *O. australiensis*

Table 21 SNPs / InDels markers distinguishing the clades defined in the chloroplast phylogeny.

	Sequence	SNP clade
1	CGCGACCTTGGCTATCAACTACAGATTGGTTGAAATTGAATCCGTTTAGG/ATTGAAAGCCATAGTACTAATACCTAAAGCAGTGAACCAAATCCCTACTAC	G in clade G
2	GGAAGATTAATCGGCCAAAATAACCATGAGCGGCCACAATATTATAAGTT/CTCTTCCTCTTGACCAAATCTGTAACCCTCATTAGCAGATTCGTTTTCAGT	T in clade E
3	CTTCCTCTTGACCAAATCTGTAACCCTCATTAGCAGATTCATTTTCAGTA/GGTTTCCCTGATCAAACTAGAGGTTACCAAGGAACCATGCATAGCACTGAA	A in O. australiensis
4	TACCATCAGAGAAACTTCCTTGACCAATAGGGTAAATCAAGAAAACAGCG/AGTAGCAGCTGCAACAGGAGCTGAATATGCAACAGCAATCCAAGGACGCAT	G in clade F
5	TTTCATTGCACACGACTTTCCCTATGTAGAAATAGGCTATTTCTATTCCA/GAAGAGGAAGTCTACTAATTTTTTTAGTAGTAAGTTGATTCACTTACTATT	A in clade A
6	ATCGTGCTTGCATTTTTTCATTGCACACGACTTTCCCTATGTAGAAATAGC/GCTATTTCTATTCCGAAGAGGAAGTCTACTAATTTTTTTAGTAGTAAGTTG	C in clade G
7	TTACCTTGATCATTTATCAATCATTTCTAGTTTATTAGTTTTGTTTAATA/GATTAATTAAGAGGATTCACCAGATCATTGATACGGAGAATATCCAAATAC	A in clade E1
8	ATTTATTGGTACACTTGAAAAGTACCCCAGAAAATCGAAGCAAGAGTTTG/TCTAATTGGTTTAGATGGATCCTTTGCGGTTGAGTCCAAAAAGAGAAAGAA	G in clades A, F and G
9	GAAACAACAAGAAAAATTCATATTCTGATACATAAGAGTTATATAGGAAT/CCGAAATAGTCTTTTATTTTCTTTTTTCAAAATAAAAATGGATTTCATTGA	T in clade G

	Sequence	SNP clade
10	GGACAAGACTGTTCTCGTAGCGAGAATGGGATTTCTACAACGATCGCAAC/ACCCCTCAGATAGAATCTGAGAATAAAACTCAGAATAAAAAAAATTGTTGT	C in clade B
11	ATTAACCGTTTCACAAGTAGTGAACTAAATTTCTTGTTATTAGAACCAAG/TAATTTCGACAAGTTCGGAACCATTTAATCCATAATCATGGGCAAACACAT	G in clade A
12	AGAAAAAATCAAAGGTCTACTCATAGGAAAACCT/AGCTTTTCCCTACATCAGGCACTAATCTATTTTTAACGTCTAATTAGATCAGGGAGTTCTTCCAATT	T in clade A
13	CTTCCAATTAAGAAGTTAAGCTCGTTGCTTTTTA/GTTTTACCAGAATTGGAGCCAGGCTCTATCCATTTATTCATTAGACCCAGAAAATCG/AGAATTTTTTTATT	A,G in clade G
14	TTCTTTCTTTTTCTTTAAAGAATTCCGCCTTCCTTAAAATATCAGAAACA/TGTTCTTGTAGGTTGAGCACCTTTTTCAAGGAAATAGAGAATAGCTGGAAC	A in clade A
15	TTCTTTCTTTTTCTTTAAAGAATTCCGCCTTCCTTAAAATATCAGAAACG/TGTTCTTGTAGGTTGAGCACCTTTTTCAAGGAAATAGAGAATAGCTGGAAC	G in clade E
16	TCATCTCGAACAAATTCACTTTTATTCCTTATTCCGGTCCAATTCTATTGTTGAGGTTGAGACAGTTGAAAATCGTGTTTACTTGTTCGGGA	Ins in clade C1
17	CTAATTTATTAGTTTTCACTAACCCTAGATTCTTTCCCTTGATAAAAAAG/TAAATTCTGTCCTCTCGAGCTCCATCGTGTACTATTTACTTAGCTTACTTA	G in clade F
18	CTTCAAGTCGCACGTTGCTTTCTACCACATCGTTTTAAACGAAGTTTTAC/ACATAACATTCCTCTAATTTCATTGCAAAGTGTTATAGGGAATTGATCCAA	C in clade G
19	TATAGGGAATTGATCCAATATGGATGGAATCATGAATAGTCATTAGTTTA/CGTTTTTTGTATACTAATTCAAACTTGCTTTGCTATCTATGGAGAAATATG	A in clade G

	Sequence	SNP clade
20	TTCTCGTATTTCTTCGACTCGAATACCAAAAGAAAGAAAAAAAATGAAGTAAAAAAAAAACGCATTTCCTGTAAAGTAAAATTAAGGTCTTTGCTTTTACTT	Ins in clade A
21	TAGTTAAACTATTGCAATGAAAAGAAAGTTTTTTGGTAGTTATAGAATTA/CTCGTATTTCTTCGACTCGAATACCAAAAGAAAGAAAAAAAATGAAGTAAAAA	A in clade B
22	TTCTATCTAACGAGCAGTTCTTATCTTATCTTTACCGGGATGGATCATTT/CTGGATATTTAAAAAATCGCGGATCGAGATCGTTTTCGCTTAACCAAAGAA	T in clade A
23	AAAAATTTATCTCTATCATAAATCTATCTCTACCATAAAGGGG/AATAGGTCTCGTTTTTTATACAATGTTCTAT/CGTCAAGTTTAAAA	GG,T in O. australiensis
24	TAGGTCTCGTTTTTTATACAATGTTCTACGTCAAGTTTAAAAATTTTTCATGAAAATGAAAAAAAGATTTTCAATTTGACTGGACTTGACACTGGATTATGTTT	Ins in O. officinalis
25	ACTTGACACTGGATTATGTTTTCTGAGACAGAAAATGAACGCATTAGGAA/CTGCATCGAATCTAAGAGTTTATAAGAGAAAAAAATTCTCTTTAATAAACTT	A in clade F
26	CTTTATGTCTCGTGCAGAATACAATACGATTTCATCTTTCGTTTCATCAT/GAAAAAATCTGGGACGGAAGGATTCGAACCTCCGAGTAACGGGACCAAAAC	T in clade C
27	GGAAGGATTCGAACCTCCGAGTAACGGGACCAAAACCCGCTGCCTTACCG/ACTTGGCCACGCCCCATTTCGGGTTTTATGCGACACTAATAAACAGTATTA	G in clade A
28	CATTACATGGAATTCTATTAAGATATTATATGAAAGTCGAATTTCTTCCT/ACTCTCATTTGAGAGTGCGAATACAAGGAGGTATTTTGTGTTTGGGAA	T in clade E1
29	TTATTTATCCGACTAGTTTTTTCTTCGCCAAATTGCCCGAAGCTTATGCG/CATTTTCAACCCAATCGTGGATTTTATGCCTGTCATACCTGTACTCTTTTT	G in clade B

	Sequence	SNP clade
30	CGACTAGTTTTTTCTTCGCCAAATTGCCCGAAGCTTATGCCATTTTCAAT/CCCAATCGTGGAT TTTATGCCTGTCATACCTGTACTCTTTTTTCTATTAGCC	T in clade E
31	TACGAGAAAATCCGGGGGTCAGAATTCCTTCCAATTCGAAAGTCCCAAAT/CGATCCGAGGG GGCGGAAAGAGAGGGATTCGAACCCTCGGTACAAAAAAATT	T in clade G
32	TTCTTTTTTCTTTCTAATTCTAAAATTGGATATTGGCTAAAAGACAATCG/AGATAGATTTTCT CTTCAGCAGGCATTTCCATATAGGACTTGTTATAATAA	G in clade D
33	ATTCTAAAATTGGATATTGGCTAAAAGACAATCAGATAGATTTTCTCTTT/CAGCAGGCATTT CCATATAGGACTTGTTATAATAAAACAAGCAGGTT	T in clade A
34	GCAGGCATTTCCATATAGGACTTGTTATAATAAAACAAGCAGGTTATAGAAAGAAAAAAACT CTTTTTTTTATTATTTATCAACAAAGCAAAAAGGGGTCTTATC	Ins in clade A
35	TGTATAAGTGGATTTTTTTGTATTTCCTTAGACTTAGACCG/ACGCAAGGCAAGAATTTCTCGC TATTTACG/TATTTCATATTCTTGTTACTAGATGTT	G,G in clade G
36	GATTTCGAAAGTCAATTTTTCTTTTCAATATCTTTATCTTTCTTTTTTTTCAGAATCCTATTTTT GTTCTTATACCCATGCAATAGAGAGCGAGTGGG	Ins in O. officinalis
37	GCAAATACCTTCCGCGCTTTTAACCCAACTCAAGCTGAAGAAACTTATTCC/AATGGTCACCG CTAATCGCTTTTGGTCCCAAATCTTTGGTGTTGCTTTTTC	C in clade C2
38	TTCCCTGAGGAGGTTCTACCACGTGGAAACGCTCTTTAATGGAACTTTC/TGTTTTAGCTGGTC GTGACCAAGAAACCACCGGTTTTGCTTGGTGGGCCGGGAATGC	C in clade G
39	GCCGTGCATTTGTATGGTCTGGAGAAGCTTACTTGTCTTATAGTTTAGGT/CGCTTTATCTGTC TTTGGTTTTATCGCTTGTTGTTTTGTCTGGTTCAATAA	T in clade F

	Sequence	SNP clade
40	AAAAAAAACAAATAAAGAAACAAACGTATTCAATACGCAAAAGAAAAGAGAGAGGAAAGCAAAAGGAGAGAGAGAGGAAAGCAAAAGGAGAGAGAGGGATT	Ins in O. officinalis
41	AATAAAGAAACAAACGTATTCAATACGCAAAAGAAAAGAGAGAGGAAAGT/CAAAAGGAGAGAGAGGGATTCGAACCCTCGATAGTTCCTAGAACTATACCG	T in clade B
42	GAACATAGCCATACGAAATGACC/TCACTAACCTCTAGAAACATCTCAAATACAAATCCCTTTTCGATATATTTCTGTATACTGTATA/CCATGG/TATACAGGATCCG	C,A,G in clade G
43	CGATATATTTCTGTATACTGTATCCATGTATACAGGATCCGCTATATCT/CGCTTGTGAAATAAAGCATAAAAT/CCCCCTCAACCCCATATCCAAATAAAAAAAGTGG	T,T in clade A
44	GATTGGACTGGTCTTTCTGGTAGCTATTCTAAATTCTCTCATTTCTTAAG/ATGTGTTTAGTATTTAGTAGCCCGATACAAAATAAAAAAGGGCCGTTTATTCG	G in clade F
45	AATAGAAAATGAAACGGTCGACCCAGACATAGACGGTCGACCCAGGCGGATATAATATACCCTATAAAATATAGGACGTAGCGAGCGTAGTTCAATGGTAA	Ins in clade C1
46	ATAGACTGTGCCTTTCTTTCATTTATTTTTTCTTTTCTGCAAGGTAGGGAGGGGGCCTTGAGAGTTCCTCTTGTGGTAGCAAGTTACTTCGCAACCTGCT	Ins in clade E
47	ATAAAAAGGGTTGGATACCGCCCAACCACCCAGCCCTCTACCATG/ATCTAGACAAATAGAATAGTTA/CCTTTTATACAGACTGCTAAGTGCGGAGACGGGAATCGAACCC	G,TA in clade E2
48	CTGCTCTACTCCGCTCTGGAGCGCTGGAAACCGGTGGACGAAAAAGGTTGAATACAATACAGGCCTCTACCATGTCTAGACAAATAGAATAGTTATTTTATAC	Ins in clade B
49	CGACTCTGTACTCATAATCCAAATCCA/TATTTGTTTTTTGGATGCAATTTCAATTAGTCTTTGGA/GTACAAATCGCGAAAATGCATATTCTTCCTCAATATGCTATTGAGAG	A,A in clade G

	Sequence	SNP clade
50	AATGAAACAGAGAAGGTTCCTCACAGTTAGCA/GGTTGGTACTTCGATCGCGGGCCTTTCCTTTACTTTCTTTTTTGTTCAGAATTGAACAAAGAATTTGGGGAAGAAAACATCTT	A in clade B
51	TTCGATCGCGGGCCTTTCCTTTCCTTTCTTTCCTTTCTTTTTTGTTCAGAATTGAACAAAGAATTTGGGGAAGATGTTT/AAACATCTTCCCCCACTTATCATGAAATCTGGGCCATAGA	Ins, TGTTT in O. officinalis
52	TGTTCAGAATTGAACAAAGAATTTGGGGAAGAAAACATCTTCCCCCACC/TTATCATGAAATCTGGGCCATAGAGAAAGAGTGAGATGTTTTTTTTT	C in clade B
53	AAGAAAACATCTTCCCCCACTTATCATGAAATCTGGGCCATAGAGAAAGC/AGTGAGATGTTTTTTTTTATTTATCATAGACTTTCCCTATGGCTTGAGAGAAACA	C in clade A
54	TTGAGAGAAACAATAAATAACTTAAAGAAAAAGGCG/ACATAGGAGCCGAAGGATTTACTTGATGTAAAGAAGATTCTGAATGTCTCC/TGCTTAGTCGATTCTCTCCGTTTAACTA/TTTTCTTCTCTTCTTTTCCACTCAATTCTAGTTTATTAGA	G,C,A in clade G
55	TAAAAGAATCAAAGAAGATGAATAGAACTAAGAACACAT/CAAAAAAGAGCATATAGGCCCGAGACCATTACCAAAAGTTCTTCCCAATAATCATATTGGGTAT	T in clade G
56	GTATATCACTGAAAATTAATACCCAGCCATATGGGTATATGAAGGGCGCA/GAATTCGTTTATACCCCACCCAATTAGAGGAAATAAAACATAAATGGAGAAAGTTT	A in clade B
57	AGCCAATAGAAGAAAAAAGTCCCTAATTTTTCAGACCGTTCTGAGCATGC/TGAAAAGTCAATAGCCTAAAGATAAAAAACCCTATACTTTGTGCAAGTGAT	C in clade F
58	AGACTTATATATCTCGATATATACAGATATAATGTACATTATGGAGTAGACT/CTATAATGGGAAATGAAAGTGGCTAATTTTGGAATTGAATAAGAAGCCCTTTT	T in clade E3
59	GTTTAAACACTAAGCGAAGCAGGGGGGTGTAAATTCCAAAAAAGAAATTGT/GACTCTTTTTCCTATTAGATCAATCAAATCACTACCCGTACTGAACTAATATAGAATCCC	T in clade C1

	Sequence	SNP clade
60	TTTTATTAATCTATTCTTATTCCATATCCTTTATAAACGAATTC/TCCCTAAAAAGTAGGGGATGATCCGTGAATTAACCTAACCATCAACTAAAA	C in clade F
61	AAAAACTGCTCATACTATCATTATAGTATAATGAGGAGCGGTTGTATACG/CGCCCTATCGTCTAGTGATGCCCCTATCGTCTAGTGGTTCAGGACATCTCTCTTTCAAGGAG	G in clade A,F,G
62	TTCCTGGGTCGATGCCCGAGCGGTTAATGGGGACGGACTGTAAATTCGTTGACG/AATATGTCTACGCTGGTTCAAATCCAGCTCGGCCCAAAAATCTAGGGCTTCGTGA	G in clade G
63	GCTCGGCCCAAAAATCTAGGGCTTCGTGAATATGAGTTAAATCCATTTTTA/TTTCTTCCATAAAAAAGAATATTTGATCCATAGAAATAAAAGAAATAAAGGAT	A in clade D +W2024 Z5 Indonesia, W1554 Z4 Thailand
64	TTCCTCTCTTACAAACAAAAGACCTTTTCTTATTGGTTATTGAAAGGTGGATTC/ATTATCTATTTTTAGCGATAATAAATCGCGACATACTAGTTATGTCATTCTCACTATA	C in clade A
65	CACCGCCCTGTCAAGGCGGAAGCTGCGGGTTCGAGCCCCGTCAGTCCCGAACTAGGGTC/TCAATGAATGGAGAAATTCATCTTTCCTTTTTCCATGAAAAAAGGGGGGCAGGAAGCAAGATCAAATA	C in clade A
66	TTTTTTAGTTCGCGTTTCTCAGTAAAGAGG/AGA/GAGAGTATAGGAATTTTTTTATCACTACTTCTGGTTGATAGCGAAAGACATACATATCATACGT	G,A in clade G
67	CTACTTCTGGTTGATAGCGAAAGACATACATATCATACGTGGAAGGGATCT/CTCCTATGTTATACTATTCCACTCTCAACCATGAATTGATTTGATAGATCCGATATTCATAATATTGAAT	T in clade G
68	GAAGTTCAATTAATCATTGAAGAAATGAAAAAGGATTAAATAAAAA/TAAAAATCCAAGTCTTAAATGAAAGGATCCGGTTGGAATCATAAAGTGTGGTAGAAAAA	A in clade A

	Sequence	SNP clade
69	CAATTTCTTTTTTCACTGCATCCACTTAATTTCAATCAAGTCAAAATA/GAAAAAATCCATGGAGGGAGAGAAAAATAATATGAGAATAGACTATAGTAAAAG	A in clade G
70	TGTTTCAAAAGAGCATAAAATTTATTTTAAGAACTAAGAATAAGAAAAGAA/GTATAAAACAAATGGAAAATGTGCGATATGTTGGGAATAGCTCCGCGGAAGAAA	A in clade E
71	AAATGGAAAATGTGCGATATGTTGGGAATAGCTCCGCGGAAGAAAATCTAAAA/GTTCTTATGTATAGAACTTTTTTAACCATGGGTCGCTTCTAGTAGCGATTATGA	A in clade A
72	CTAAAGTTCTTATGTATAGAACTTTTTTAACCATGGGTCGCTTCTAGTC/AGCGATTATGAATTGCTCTCACCGCTCTTTCTATTTCTATTTTCTATTC	C in clade F
73	ATGAAGGAATTATTCTACTATTGATGAATAATCATAGTAGAATCAAGGGTACAGAGTCAAAAAGGGGTTCTGACCTAAAA/GGCTATGGATGAATCAGTTCAAAGAATTTACTC	In clades C,D
74	GAAACGCTATCTCATCCCTATTGGTAG/TCGGTTTGGGCCACTACTGCTAAAACAAACCCCAGTTTGAGGAAAGAACGGTGGGTTCTCAAAATCCAGTATCGCCGAGCCT	G in clade G
75	GAAACGCTATCTCATCCCTATTGGTATCGGTTTGGGCCACTG/ACTGCTAAAACAAACCCCAGTTTGAGGAAAGAACGGTGGGTTCTCAAAATCCAGTATCGCCGAGCCTTGTTATTCTC	G in clade F
76	GGTTCTCAAAATCCAGTATCGCCGAGCCTTGTTATTCTCTTGCCCCAACTTATGCGGGGTGCACAAATTTGTCGATTTGGATCAGTACTATAAGCCTAAGTATTTTATTGATCAGGCGGCAC	Ins in clade F
77	CCATGCCGCCAAAAAAATACGATCTAAAATCGAGAAAAGAGCAAGTATTCATG/CCACGTTTCTTACTAAAACTAACTTTCTTTTATCTTAAATCTAATTCTACTTA	G in clade A
78	ACGTTTCTTACTAAAACA/TAACTTTCTTTTATCTTG/AAATCTAATTCTACTTACTTTTTTCCAATCTTTTTCAAAAAATCTATTCATGCTTTTTTTGGATCCAGTTTCGATTATTCTCCTCG/AAAGGATTCTATCTTAAAACACACATTGCTAACACTAGAAAACTTC	G,G in clade G

	Sequence	SNP clade
79	TCGGTATTCTCTCCCGCCTGCCATTTAATGT/GCATAATAAAAGACAATGGATTTATGCCTAATCCGTATATAGGTAAACTCCAGGTCCGAACA	T in clade E
80	TGCCATTTAATGGCATAATAAAAGACAATGGATTTATGCCTAATCCGTATATAA/GGTAAACTCCAGGTCCGAACAGCATTATTATCTATGGATCCCCCTTATGTACATATCTC	A in clade C
81	TATTCTCAATCGAACTAAAGTCAAACTTTCTAGTGCTTATAAATTATTATC/ATTTTGGTTTTATCCCATTCATAGAAAGGAGAAAAAATGAGAAATCTTTGCCGTC	C in clade A
82	GTCAAAAGAAAAAGCTATTTTGGAGTTTTATCAACAATTTGCTTGTGTAGGC/TGGGGACCTGGTATTTTCGGAATCCTTATGTGAGGAATTACAAAAGAAATT	C in clade G
83	TTACGAGACCTTCTTTGGTACATATCCCTTATCTCAAGTTTTTGATCAAACCAATCCATTGACACAAACT/GGTTCATGGGCGAAAAGTGAGTTGTTTGGGTCCTGGAGGATTGAC	T in clade G
84	AGGATATTTATACTTCTTTTCACATCCGAAAATATGAAATTCAGACGGATACG/AACAAGCCAAGGCTCCGCTGAAAAAATCACTAAAGAAATACCACATCTAGAAGAACATTTA	G in clade G
85	AATCACTAAAGAAATACCACATCTAGAAGAACATTTACTCCGCAATTTGGAT/CAGAAATGGAGTTGTGAAGTTGGGGTCCTGGGTAGAAACAGGCGATATTTTA	T in clade B
86	TCTCAGAAGAACTTCCAGGTTAATAGGGAAGAAGTTTGATCGGAATAAATC/ATAAATTCTTTTCTTATTTCTATTTTATGATTGACCAATATAAACATCAACA	C in clade A
87	ATTTCTATTTTATGATTGACCAATATAAACATCAACAACTTCAAATTGGC/ACTCGTTTCCCCTCAACAAATAAAGGCTTGGGCTAACAAAA	C in clade G
88	TTGATTCTCGGATACGAAGATATCAAATGGGATACATCAAACTCGCATGTCCCGTGACTCATGTGTGGTATTTG/AAAAGGTCTTCCTAGTTATATCGCGAAT	G in clade G

	Sequence	SNP clade
89	ATAGCAATAAAGCTTTTTCAGCTATTTGTAATTCGCGATTTAATCACGAAACGC/TGCTACTTCTAATGTTAGGATTGCTAAAAGGAAAATTTGGGAAAAGGAACC	C in clade B
90	GCTACTTCTAATGTTAGGATTGCTAAAAGGAAAATTTGGGAAAAGGAACCT/CATTGTATGGGAAATACTTCAAGAAGTTATGAGGGGACATCCTGTACTGTTGAATAGAGCACCT	T in clade F
91	ACTATTTGTTTACACCCATTAGTGTGTAAAGGTTTCAATGCG/AGACTTTGATGGGGATCAAATGGCTGTTCATCTACCTTTATCCTTGGAAGCTCAGGCGGAAGCTCGTTTACTTATGTT	G in clade G
92	GCGGAAGCTCGTTTACTTATGTTTTCTCATATGAATCTCCTATCTCCCGCTATTGGA/GGATCCTATTTGCGTACCAACCCAAGACATGCTTATCGGACTTTATGTATTAACGATT GGAAAC	A in clade G
93	CGAAAAAGGGGGTACTTATTTATGGCGGAACGGGCCAATCTGGTCTTTCAT/GAATAAAGAGATAGATGGAACTGCTATGAAACGACTTATTAGCAGATTAATAGATCATTTCG	T in clade A,F,G
94	GTTTTCTTTTGGAAAAACACTATTATTATGGGGCTGTACACGCGGTAGAAAAG/ATTACGCCAATCCGTTGAAATCTGGTATGCTACAAGTGAATATTTGAAACACG	G in clade A
95	TTACGCCAATCCGTTGAAATCTGGTATGCTACAAGTGAATATTTGAAACAA/CGAAATGAATTCGAATTTTCGGATAACAGATCCTTCTAATCCAGTCTATCTAATGTCTTTTTCAGGAGCTAGAGGAA ATGCATCT/GCAGGTACACCAATTAGTAGGTATGCGAGGATTAATGGCGGATCCTCAAGGA	A,T in clade G
96	GATATTCTACATAGTGTGACTATTCCC/TTCAAAAAGCTTGATTCTAGTGCAAAATGATCAATATGTAGAATCCGAACAAGTAATTGCGGAGATTCGTGCCGGAACGTCCGCTTTGCATTTTAAAGAAAG/AGGTACAAAAACATATTTATTCCGAATCAGAC/TGGGGAAATGCACTGGAGTACCGATGTTTATCATGCGCCCGAATATCAATATGGTAATCTTCGTCGATTACCAAA	C,G,C in clade G

	Sequence	SNP clade
97	GATTACCAAAAACAAGCCATTTATGGATATTGTCAGTAAGTATGTGCAGAG/TCTAGTATAGCTTCTTTTTCGCTCCACAAGGATCAAGATCAAATGAATACTTATTC	G in clade B
98	GTATAGCTTCTTTTTCGCTCCACAAGGATCAAGATCAAATGAATACTTATTCT/CTTTTCTGTTGACGGAAGGTATATCTTTGGCCTCTCGATGGCTGATGATGAGGTAAGACATAGAC	T in clade G
99	TTGACGACCCACGATACAAAAAAGATAAAAAGGGTTCG/AGGAATTGTTAAATTTAGATATAGGACCCTAGAGGACGAATATAGGACTCGAGAGAAAGACT	G in clade G
100	CCCGAGAGGAAGAATGTAAAACCCTAGAAGACGAATATAGGACTCT/GAGAGGAC/GGAGTATGAAACCCTAGAAGATGAATATGGGATCCCAGAGG/AACGAATATGAAACCCTAGAAGATGAATATGGAATCCTAGAGGACGAATAT	T,C,G in clade G
101	ACCACTAGAAAGAGAAAAAAAAGATTCGAAGGAATCAAAAAAAAGGA/GAAAATTGGGTCTATGTTCAA/GTGGAAAAAAATTCTCAAGAGCAAGGAAAAGTATTTTGTTTTGGTTCGACCTGCAGTC	A,A in clade G
102	AAAAAGAGGAGGCTCGTGCTTCCCTTGTTGAGATAAGAGCAAATGA/GTCTGATTCGCGATTTCCTAAGAATTGGGTTAATCAAATCCACTATTTCGTATACACGAAAAAGGTATGA TAGCAC/GAAGTGCAGGACTGATTCTCCATAATAGGTTAGATCGCACCAATACCAATTCCTTTTA	A,C in clade G
103	GGTTTTGTCGGCATCCAACTGTTCTCGAATTGGTTTTTTTAAGAATTCC/AAAAAATCCCAATGGGGTAAAAGAATCGAATCCTAGAATTCCTATTCCAAAATTTT	C in clade A
104	CAGTTAAATTGGCACTTTCTCCCTCATGATTCTTGGGAAGAGACATCAGCT/AAAAATTCACCTTGGACAATTTATTTGCGAAAATGTATGTCTATTTAAATC	T in clade F
105	GGATTGGAATGAGCGTATACCAAGAATTCTTGGGGGTCCTTGGGGATTCTTGATTGGAGCTGAGC/TTAACCATAGCCCAAAGTCGTATCTCTTTGGTTAATA AGATCCAAAAGGTTTATCGA	C in clade G

	Sequence	SNP clade
106	TTATACCTGTTGGTACCGGATTCCAAAAATTTGTGCACCGTTACCCACAAA/GACAAGAACCTTTATTTCGAAATTCAAAAAAAAAAACTATTTGCGTCGGAAATGAGAGATA	A in clade E2
107	TCGGCTCTTTCTTAATCTTCGAAAAGAAAGAAATTTCGTAATGGAAT/AGGTAGGATGAAAAAAAAGAAAAAAATCAAAAGGAAGTGTGGAAAAAATGACAAGAA GATATTGGAACAT	T in clade G
108	TATGTTAACGAATTGGTCGATTACTAAAACTAGACTTTCTCAATTTAGAGAT/CTTAAGAGCAGAAGAAAAGATGGAAAAATTCCACCATCTCCCAAAAAGAGATGTGGCAATCTTGAAGAGAAAATTATCTACCTTGC	T in clade A
109	TCGGGCGTAGAAGTAGGCCAACACTTCTATTGGCAAATAGGAGGTTTCCAAATTCATGCCCAAGTACTC/TATCACTTCTTGGGTCGTAATTACTATCTTGCTAGGTTCAGTTATC	C in clade G
110	ACTATCTTGCTAGGTTCAGTTATCATAGCTGTTCGCAATCCACAAACCATT/CCCAACCGATGGTCAGAATTTCTTCGAATATGTCCTTGAGTTTATTCGAGACTT	T in clade B
111	GAAGAGGAAAGAAAGAAGGATGGAATGAAAGATCAGTTGGTTGGAAAGAAAGAGAAATAGAATAG/ATGAGTACACAAACCTCTAATGATTAGAAACTAAAAAGGAGATCTCGAAGCAGTTCGGAGAATT	G in clade A
112	CTTAGTCTAGCTTTTATGGAAGCTTTAACAATTTATGGACTAGTTGTGGCACTA/GGCGCTTTTATTTGCGAACCCTTTTGTTTAATCCTAAAAAAGAAAACGAGTCCTTTAGATT	A in clade G
113	GATTTGAGGATGATCAATTTAGAGGATATGTTCGCCGTCTTGCTTCCCGT/CCCTTTGTTTAGGGCAGTGGAAAGTATTTTTCCTTTTATTTTAGGAATTTTGGGAACATT	T in clad G
114	AAATTTTAACTAAAGGGCAAATACAAATAAAAAAACAACTTTGCTGCCCAT/CGATAGATTTTTATCTAGGCGGAAGAGTCCTCTTAATATTTATCTAGTCTTA TATGGGTTTCGGTATATTGAA	T in clade G

	Sequence	SNP clade
115	AATCGAAAACAGAGGATCTTGAGTACTATTCGAAATTCGGAAGAATTGCGTAGAGGG/AACCATTGAGCAGCTCGAAAAAGCTCGAATTCGATTACAGAAAGTCGAACTAGAA	G in clade G
116	CCATGAAACAAGTAGCTGGCAAATCAAAATTGGAATTAGCTCAATTCGCG/AGAGTTACAAGCCTTTGCACAATTCGCCTCTGCTCTCGATAAAACAAGTCAGAATCAATTG	G in clade E
117	AAAAACAAATGCATATACAAATGTATGATGCATATATCATAAAGAAGGAATATATATGGAGCGGGTAGTGGGAATCGAACCCGCAACCCCACGGTTATGAGCCTTGTCAG	Ins in clade B
118	AAACTGCCAAATAAAACGCGTCCCAAGCAGAAATATCACAAGTACCGCCGCGACCT/AGGGCCGTCGCAAGGAAAACTATATCCAAAATCTTTTTTATCCGGCATTAATT	T in clade G
119	GTCTAACATTCTTGCCAATACATTATCCTCATTCTGTTCCGGATTGTAATCC/TCTAATGAAAAAAATAGCTCCATGAGCAAAAGCCCCTGTCATGATGAACCCTGCAATGT	C in clade G
120	CTTGGTTTCCATTTGGGTTGTAGATGTAACCAACCCCCTATTAAGGATAGC/GGTAGAAAGAAATAATAGAAAAAGAGCTCCTGTATAAAGATCTTCATTGGTCCGTAA	C in clade G
121	CCTTAGGATCAACCCCAGCGTCAAGAAATTGGTTAATCGGTAAAGATACG/ATGGATTTGGTGCCCCGCCCAAGAAAGAGACCCAAGTCCTAATAACCCTGCTAAGTGAT	G in clade B
122	GAGGCAAGTGTTCGGATCTATTATGACATAAGGATTGGGTGCCTAACGGACTTTTTTTTATCTTGGATTTCTCCACGTAACAAAAAAACCTTTTTTTAATTTAAA	Ins T in clade A
123	TTCCGACCTAATTTATTTGATTAATGGATCAACAACCAAACCCCCATTTTA/CTGAAAAAGGAGAGTGGTCTTATTCAAATTCAAAGCGCTTCGTAATCTTCAACCAGTTCTG	A in clade A,F,G
124	GCAGAAAAATGAAGCATAGATAGACCTATATCCTTCGTCCA/GAATTTTCTGAAAGGTAACTATCTCGGTTTCATATATGAAATTTCTATAGAATCC	A in clade F

	Sequence	SNP clade
125	ATGGGATAAGTAAGCAGTTTTTTTTAGTTGTATCGACCCAGTCGC/GTCACTAATTGATCTTTACGGTGCTTTCTCTATCAATTTGAGAACTCTATCCATAGAGTAGTATAGGCCATACTT	C in clade G
126	TATAACTTCGATCATAGGGATCAATTTCTAGTCGCGTAGCTTCATAATAATTC/TTGCAAAGCTTCCGCATAATTTCCTTCGGATTGAGCCAACATCCGTTACGGTCGT	C in clade G
127	TCCCACAAGAGGTTTTTCTTAACACCAATGAATTCTATTAATGCTAGAGA/GAAAACGATAGCTCCAAGAATTTCTTTGTTCTCAACGCCTCCTATTTAGAGGAAT	A in clade F
128	TGTTATCCCAACCATTCTTCCCAGCCCTGATACCAATCAGGAAAGGGC/TTAATTTCTAACAAAGTTTTTCTCTTGTTGATTCCTATTTCTAGGTGTAGTGCTTTTA	C in clade G
129	CCCCTATGCTACCTATTAGTACTAGTAGAGTAGGATTAGCCTGTAATACAA/GAACCTATCCTGTAGGTGTAACCTTTCGCTCAATACTAAAATCTACAATTGAAGCAT	A in clade F
130	AAAGAGTCAAATCGCACCATCTCTATAATAAGTAAATGCCCTTTTTTCCCCT/GGAGGTTGTCGGAATTATTCGCAATAAAATATTGGCTACAATTGAGAAGGTCTTA	T in clade G
131	TTGAGAAGGTCTTATCAATGAAATTTCCATTTATACGGGATCTAGGCATAATTCCCAAT/CCCATTCTATCATTCTATATAGAATTCTTTTCATTCCTTCACAAAATAACAT	T in clade F
132	TCCATTCAATTCTTAT-AAATCGATCCCTATGCTCCAAATGGATAAGG/AGAGGTATTTCTGCTCAGCCCAAATTCTCTCTTTTTCCTTCTGTTTGAACAAGAAGAGAT	G in clade F
133	AGGAATAGGAAAACTCGCTATTCACTCAGTTTTTTTTCCATAATAAGAG/TTATGGAGGAGAGATGGCCGAGCGGTTCAAGGCGTAGCATTGGAACTGCTATGTAG	G in clade C1
134	TGGTTGTACCTGTACTGCAGGAATAGGAAAACTCGCTATTCACTCAGTTTA/TTTTTCCATAATAAGATTATGGAGGAGAGATGGCCGAGCGGTTCAAGGCGTAGCA	A in clade G

	Sequence	SNP clade
135	GATATTTTTAATTTGATATGGCTCGGACGAATAATCTAATACATGGATAAAGAATAAATAATATATATACGAAAACATAATAAAGAGAACATGCGAATTTCTTGTATT	Ins T in clade B
136	TCTAGTATTTATCCTGTTTTTTTTATTAATAGGTTTAAGATTCATTAGCTTTATCATTCTGCTCTTTCACAAAGGAGTGCCAAGAGAACTCAATGGATCTTATGTTATTCATTGAATACATTTCTTTTTTATTATAGTATCGGCAAGGAATGTCGATT	Ins in O. officinalis
137	TTGTACAATGCATAGGACTGCCCCCTCCCCATTTCCAAATTTTGGATTTGGAATACTTTATTGATTTTTTAGCCCCTTTAATTGACATAGATACAAATACTCTACTAGGATGATGCACAAGA	Ins in O. officinalis
138	TTCAGAAGATATGTCTAAAGTAGATGGTGATTGATAGAGCAATTCTTGCTCG/ATAAGTTCCAGTATTAGTACTGCGCCGAACATAAAGCTTGTGGCTGGTAGTAA	G in clade G
139	AAAACTAGCGCAAACATGTAATAGCGTATTCGGAATTGTAACCAAGCCCCT/CCCCATGGGTTCTATACCCGATTCATAACTAGAAAGCTTCTCTGGTCCTTCACGAAC	T in clade G
140	TGTTTCCTCTTTGCCACGTCTTCTTTAAAGATTCATCCAATGGAATCCCGACC/TCCCTTTCTTTTTGATTTCCTTTCTATTTAGGTATGGTGGAGACATAATTCTTATAGAA	C in clade E
141	TTCTTTAAAGATTCATCCAATGGAATCCCGACTCCCTTTCTTTTTGATTTA/CCTTTCTATTTAGGTATGGTGGAGACATAATTCTTATAGAAACAAAACTCTC	A in clade C1
142	ATGAGGAGTAATTCTATAAAAATAAAGAACTCTATTTCAGAACGTAGATC/TGATTTAGATTTAGGTAATCTATAGATATAGATAAGCAAAGTAATATACTTCAAACAAAGTAGGAATT C/TGCAAGATGGAGAACATCTTGCAGTTGATTTGATAGAAATTCATTTTTCTTTT	C,C in clades A,B,F,G
143	TACTTCAAACAAAGTAGGAATTCGCAAGATGGAGAACATCTTGCAGTTATTATAGGGAAGTCTAGGGACTTAGAGCATATCCTATTTGAAGGAGGGTGGAATTCAAATCTGGTAAAGGATCTTTGCTTCTATTGATTTGATAGAAATTCGTTTTTCTTTTCCTGTCTCTATAATTTTC	Ins in O. officinalis

	Sequence	SNP clade
144	TCTATAATTTTCGATGAATGAGCCTCTGGTAATCCTTTTC/ATCTCTATTTTATGGCGCAGGCG CCTGTCCAGTCTATAAACAAGTACTAATAGGGAAATGAAAACTATA	C in clades C,D
145	GATGAATGAGCCTCTGGTAATCCTTTTATCTCTATTTTATGGCGCAGGCC/GCCTGTCCAGTCT ATAAACAAGTACTAATAGGGAAATGAAA	C in clade F
146	GACATTGATTTTGCAAGAAGATCCACTATGTTCATTGCATAATAAGCTCCT/CTTGAAAAGCA TTGGCGCACGTGTAAACGAGTTGCTCTACCGAACTGA	T in clade A
147	CAATAGTAGGTAGGTAGGTAGAAAAATTACTAGATAGCATTGGA/CCCTACTTCGCTTCGCTA TCTAATAAC/TTTTTTCTACCCCTCTTCCCTTTTTCTTTGTATCA ACTAAACCGTTGGGTTGTCTTCAATTAGATG	A, C in clad G
148	TGGGGGAATCCAATTAACAGCCTCGACTCGTATCCTAGCTCGTCTGAGAGCTAG/CCTTCGCT TCAACCAATTCTTTCGTACCCTCAGCTCTACTCACGTTAGCTTCG GCTA	G in clade G
149	AATTTGCTTACCGTCAGTGTCTCGACTCTTGACTACCAAAGCG/ATTATAAATATAAGGTAAC TTGCCCGGGGGAAAAGTGACATCCAGCACGGGTCCAATAATTTGATC	G in clade G
150	TTGATTTCGTTGCCCAAACGAATCCCATTCAATCGTTTACTCATGGAATGAGC/TCCGTCGGA AAGTTCAATCAATCTTTTTTTCATATACATTTTGCCTTTTGTAAACGATT	C in clade E
151	GTCCGTCGGAAAGTTCAATCAATCTTTTTTTCATATACATTTTGCCTTTTGTG/AAACGATTTG TGCCTACTCTACTTTCTTATCTAGGACTTCGATATACAAAATATATAC	G in clade G
152	CTTGATCGTTACAAAGGCCGATGCTATCACATCGAGCCCGTTGTTGGGGAGGAA/TAATCAAT ATATCGCTTATGTAGCTTATCCATTAGACCTATTTGAAGAGGGTTCTGTTACTAA	A in clade G

	Sequence	SNP clade
153	CGTCCTTTATTGGGATGTACTATTAAACCAAAATTGGGATTATCC/TGCAAAAAATTATGGTAGAGCATGTTATGAGTGTCTACGT/CGGTGGACTTGATTTTACCAAAGATGATGAAAACGTAAACTCACAA	C, T in clade G
154	GATTCTGTATTGCAATTTGGTGGAGGAACTTTAGGACATCCTTGGGGTAATGCG/ACCTGGTGCAGCAGCTAATCGGGTGGCTTTAGAAGCCTGTGTACAAGCTCGTAA	G in clade A
155	AAACTAAAGGAGAATGAATGAAAAAAGACAGAGTTTGGAAGTTAGACCCCTTTA/CTAAGACTCTCTTTCAAAAAAGAGGACATTTTGAAACTTTTAACAGGCACAATCGTGAGTCAACAAGTGACTCGAAATGCC/TCGTAAGAAAAGAGAATTGATTTTCAAAATGGTAGAACTAGATGA	A,C in clade G
156	TAAAAGAATTTGTCCCTTGATTGAGTATGCTATTTTT/CCCTCCTTTACCGCGCATTATTGTATAC/TGCTTCTAGAAGAGCACGTATGCAGAGAGGAAATTACAGTTTAATAAAAAA	T,C in clade G
157	ATGTGGAAGGAAGTAGACGAAAAAGATTTTGGATTCGAAATAGGCA/GCATTCGACTAAGTCA/GTACTTTGAATCCAATTTCAAGTTCA/GATTAGAAGGATAGG/AAAGGCCGCGAGGATCGGAAAAGAAAAATCAAATCTTTTTAATTGCTTCT	A,A,A,G in clade F
158	TGCATGTGGAAGGAAGTAGACGAAAAAGATTTTGGATTCGAAATAGGT/CGCATTCGACTAAGTCGTACTTTGAATCCAATTTCAAGTTCGATTAGAAGGATAGA	T in clades C,D
159	AGATATACTTAATTATATCATAAGAATCTTAAGATATTTTTC/TGAATAGATAG/CAAATCGAATAGATAGAAATAGTAAATTTGAATGGAGACACCTATTCTATGATG	C,AG in clade G
160	ACAGGATCATAATACGGATCTTTTGTAGTGTAAGTAATATAATATGGTAC/TGTTATGTGGCTCTTTCTACACACAAATGCAAACCCGCTATGGATGC/GGGATTATGGATGCGGATATAGGCTACGAGCATAAATGCATGCATATGCGGAACCGGGTAT	C, C in clad G

	Sequence	SNP clade
161	TTTTACAGGAGTATCTAGTTGGCGAAGGCGATTTCAGAATCAAAAAAAGTAAAGTAAAGTCAAAATCATTTAGCTTATTCTCTCAATTTCAATCGACCGCTG	Ins in clades C,D
162	TGAATAGAAAGTCAATGTATCTAACCAATTATTTTACAGGAGTATCTAGTTGC/GCGAAGGCGATTTCAGAATCAAAAAAAGTAAAGTCAAAATCATTTAGCTTATTCTCTCAATTTCA	C in clade G
163	TTTCTCAACACGAGGGAAAAGGTCCCTTCGAAATTGCATTATTGTAAGGGGATTTTGAGTATTTATCTAAAGGAAGGAACAAATGAGGATAAGAGAAAATTGCTTC	Del. In clade F
164	AAGACCTTTTTATCTTGGACGAAATGATAAAAGAGAAACCGAATACACATGTACAAAAA/CCCCCCTATAGGAATACGCAAGGAAATAATACAATTGGCCAAAATAGATAATGAGG GTCATCT	A in clade G
165	CCCACGAGAAGCAACTGGACGAATTGTATGTGCCAATTGCCATTTAGCT/GAATAAGCCTGTGGATATTGAAGTTCCCCAAGCAGTGCTTCCCGATACTGTATTTGAAGCAG	T in clade G
166	TTGGGCACAAGAAAAAGGCTTTTTTGCCTTTTTCTTGTGTCGATTCTTCTTGTATTGTATCGAAATATGAATCTTTTTTCTTCCTATTCGGCAAAGATTACTATTTC	Ins in clades C,D
167	ATTAGTTTATTACTCTAAATTAAATCAATGATTTACAAGAGACTTCCTCCGGGT/GAATAAAATATTGGATCCTCGATTGATCCTTTCTTTCTCCTCGCTTCATAAAAG	T in clade A
168	TACAAGAGACTTCCTCCGGGGAATAAAATATTGGATCCTCGATTGATCCC/TTTCTTTCTCCTCGCTTCATAAAAGTGAATC/TAATTTCATTGGCGAGGGGGTTATAAATCAACTGATGGATTACTTCACTAACATTATT	C,C in clade G
169	AACAAACAAAATTAACAAACAAAACGAATAAATAGAGGGATTCTGACCATCAGAT/GCAAAGGCTTTCTCTTTGTTATTTTTACAAATCAAAATAGGAAACCCGTTTGTAGGTTATGGAATA	T in clade A
170	GGGGGTAAGGACCCGCTAAGTTCCTATTTTTTCATGTTTACAAT/CCTGGTCCCTCCAATTACTATAGAGATGAACCCAATCCAGAATATGAACCGTAAAAGAAAACACCTATTAAAC	T in clade F, G

	Sequence	SNP clade
171	CATCCTTGTGAGATTGTCAATTTTGTACCAAAGGTGTATTTTGAGTATACCG/AAATTAGTATAGCTATCCTTCCTATGGCACAGCAATCCTGTTTCG	G in clade G
172	ACCAAATTAGTATAGCTATCCTTCCTATGGCACAGCAATCCTGTTTCGAGACCAAG/CTTGGTCTCGAAACAGAATTCTTTTTTTCTCTTCTTTGTTCCTTGTCTATAGGGTAAGCTA	AGACCAAG in clade C,D
173	CAATAGAAAACCTCAATTTTGAGGGTCCTACTTAATTTTCACCGGCTTCGGATCGGAATAGTAGAATAATTCGGAATAGGGCTCAAGATCTTGGGAAAATCTA	Ins in clade B
174	AAGAGAAGTAGATGCGAAAGCTATCCCTTCGAATCCAACCTTTCCCT/CTTAAAGAATTTAATTGGTTAGCATAATATAATATCTAATAAATAGAAAATCAAATAGTAGATAATCTGTTATGAAAGAGAGAAAACATTCTTTGAAGAATCAAGATTCGTAATCAAT/CCCTTGCCTTGTTTACTAACTTTCTT	T,T in clade G
175	CTTATTCCATATGGAATACAATCAATTAAAATAAGAAGGAATAGGGGAATATTT/CGACTGTTCGCTCCAAAAAGAAGGTTAAATCATCCTATTGAAAAAGACCAAAATAGAAAGAACTTTTTCA	T in clade B
176 177	TTTTTCACTGGGGTTAGC/ATGATCTAGTTCTTAATATTA/TTTACTTTACTCAATTGACAGATTACACAGCAAATCTCTTGATTCGGAATTA	A in clade A,B,F,G C/A in just in G
	TTTCTTGGTCATTGAGATTCGTGGATAATTTAGACTACTATTTAGGGATAG/AATCGTACCTCTTTTTTTTATCT/CCCTCGAACAAATCGAAATGATTGAAGTTTTTCTATTTGGAATCGTCTTAGGCCTA	G,T in clade F
178	GTCATTGTACACAATTCCTATCTTGTTTTCCACATCCTAATTTTCTTC/GTCTTTTTCTATCTATAGAGAATCT/CTCGTGTCATTTCTTCTTTTTGGTCTCATATAAT CAAGGAATGGTATATAT	C,T in clade G

	Sequence	SNP clade
179	TTGGACCTTAGAGTCATGAAAAATTTGGTAAATCTCATTTTTGAAAAAAT/GAAATTCAATTAAAAGCAGTATCCAAGCTAAGTCAGGCCTCAGAAATCAGAGC	T in clade F
180	AAAATTTTTTATTCTAATGGATTTTCTTCTTCCTCTTCGGTTTCAAAATAGAGGAATAAAAGAACAAAATAGAAGAATAAAAGAATAAGTAGAAGAATTAAGTTAAGTCAATCCAAAAAGGAAAGG	Ins in O. officinalis
181	GATGTTAGAATCAGAGTTATTTTGCAATGTGTGAGTTGTGTTCGAAAAGGC/GGCCAATGAGGAGTCGGCAGGGATTTCTAGATATAGTACTCAAAAGAATCGCCAC	C in clade G
182	TGTGTTCGATCTTTCCAAAGATCAAAAAGAATAAGAACTTCCTATTTAATATTCCTATTTAATATATAGAGT/CATAGATAGAATACAAAATACAAATCAACTTGTCTGATTTCCATTAGATAT	Ins ,T in clade F
183	TCTTGTGTTCGATCTTTCCAAAGATCAAAAAGAATAAGAACTTCCTATTTAATATTCCTATTTAATATATAGAGCATAGATAGAATACAAAATACAAATCAACTTGTCTGATTTCCATTAGATA	Ins in clade B
184	CAAAGATCAAAAAGAATAAGAACTTCCTATTTAATATATAGAA/GCATAGATAGAATACAAAATACAAATCAACTC/TGTCTGATTTCCG/ATTAGATATTATTTCATATGTAT	A,C,G in clade G
185	GAGGGTATTCATCTAATATATGGACCAAAGAGAGACTAC/TTTCTTCTGGATCCAAAATTAATAAAATAAACAAATCAATTTTTT	C in clade A,F,G
186	AATAAGGAATAAATCATGTATACATCTAAACAACCCTTTCATAAATCCAAG/ACAAACTTTTCATAAATCCAAGCAAACTTTTCGTAAATCCAAGCAAACTTTTCGTAA	G in clade G
187	AGTCAATTTCAATAATTACAGGTCCTAGACCCAGAAAAAATAGACATATTCCTCA/CATTAACACAAAAGTTCAATTCCAATCGAAACTTAAGAAACTCCAACCAGACTTTAAGAAA	A in clade F,G

	Sequence	SNP clade
188	TCGAAAGGGCCAGACTATATATAAAGAAAGTAATCCAATTTAGATTCTTGT/GGTTTGTTATAAGAAAGAACAATGGGGAAGAAG/AAAATAGTTTTTTTATTTATTGCAACATGCTCGTTGATTC CTACCACTTAATC	T,G in clade G
189	TACAGCTACTTGTGCAAGGATTTTACGATTAAGAATCAATTCTTTCTTGTAC/AAGATTGTGTATTAATTTACTATAATTATCGAATACTTTATGTATCCGCGTTGCTGCGTTTATCCG	C in clade G
190	TTTGCATCGTATCAAAAATCGCCATTCCTGAGATTAACCACCCGCCT/GGGGGAGTTTATAAACAAAAAAATATCGCTAATTCCATCTTCTATACT	T in clade G
191	GTGGTTGGAGTATTTCAGGAGGAACTGTAACGAATCCGGGTATTTGGAGTTATGAAGGT/CGTGGCAGGGGCGCATATTGTGTTTTCTGGCTTGTGTTT CTTGGCAGC TATCTGGCATTGGGTA	T in clade G
192	TCTTTAGAGATAAAGAAGGGCGCGAACTTTTTGTACGCCGTATGCCTACC/TTTTTTTGAAACATTTCCGGTTGTTTTGGTAGATGAAGAGGGAATTGTGAGAGCGGACG	C in clade C1
193	ACATGGGAAACATCTCCCATCCCTTCTTTGACTCTTTTTCCTTTTTTATAT/CGGGAAATGATCCCAAATGACAAATGAATAGGTGTGGAAGTTATAATTGTAAATAA	T In clade G
194	CTAAGGTTCCGACTAAAAAAGTGAAATAATTTAATTGAAGTAAGAAGTCTCCCAGATG/CATCTGGGAGACTTCTTACTTCAATTAGTCCCCGTGTTCTTCGA ATGGATCTCTTAATTGTTGAGA	AGATG in clade B
195	TTATGGCTACACAAACCGTTGAAGATAGTTCTAGACCTGGACCAAGACG/AAACTCGCGTAGGTAATTTATTGAAACCCTTGAATTCGGAATATGGGAAAGTAGCTCCGGGTT	G in clade G
196	AAACCCTTGAATTCGGAATATGGGAAAGTAGCTCCGGGTTGGGGGACTACC/TCCTTTTATGGGGGTCGCAATGGCTTTATTCGCGGTATTCCTATCTATTATTTTAGAAATT	C in clade F
197	GAGTGTGTGACTTGTTAGAATTTGCTCCTATTGATAATACATAGAAAGGG/CACCTGTTATCTCTATCAAGATGATTCTAATTCGTCGGATATTATTTATTCTAGTATCTGGAAC	GG in clade G

	Sequence	SNP clade
198	ATAGACGAGCCAACTTGAGATTTTTTGGCATTATCATCACAAAGAAGAAATTA/CTGGATTTTTCTTATTTCATATCTTCAAGGCAAATCGACCCAACCCAGTGGCTGATGA	A in clade C1
199	AAGTTGCTACCGGTTTTGCTATGACTTTTTACTATCGCCCAACCGTTACAGAA/GGCTTTTTCCTCGGTTCAATACATAATGACCGAGGCCAACTTTGGTTGGTTAATCCGATCAGT	A in clade F
200	CCCGCGAATTAACTTGGGTCACTGGTGTGGTTTTAGCTGTATTG/AACTGCATCGTTTGGTGTAACTGGTTATTCTTTACCTTGGGATCAAATTGGTTATTGGGCAGTCAAAATTGTGACAGGTGTA/G CCTGACGCGATTCCGGTAATAGGATCACCTTTAGTGGAGTTATTA	G,A in clade G
201	TCTAATGATACGTAAGCAAGGTATTTCGGGCCCTTTATAAGGAAGGCATA/CTCATAGAGAGTTCTAATTCTCATATATCATATCGGGTAGGTTGTGGTATTTCATTGCTACAAACATGG	A in clade G
202	CTTGGATATTGAGCATTTACCCATAAGAGTAGGATTCTTTTCAATGAG/ATAGTTGTAGGTGCAACTTCGGAAAATAGAATCTGATAAAGCTTTTCTTACTTAGAG	G in clade G
203	TTTTTGTTTTTCTTTAGATTAGTTAATCTTTTTTGAAAGCTTAAAAGGGGT/GGAAGTAAACCTGTTTTTATTTTCTTGGAAACGAGTACCCTCTTCCTCCGTGTGAAGAA	T in clade G
204	GTCCATATTTCTAGAAAAAGTATCTCATATTTTGCATTTCCATTCCCACAAGA/CAAAAATACTATAATTCACATTTCGAACAGGCATGGATACAGCATCTATAGGATAAC	A in clade G
205	TAATATCTTGGGCAGTTATGTATCTAGGACCTTTGACGCAAATTGATGCGT/GTTCTAACTCCATAGAGATTACTTCTCAATACAATTTCTTTCAAATTTAGT	T in clade G
206	GTAGGGCTTCCATAACTAAACCCTCGAAAGTAATTTTTGCTTCTCTCGGGG/TTTTTTTTTTCTCTCCTATTTTTCTTTTCTGTCATA-TTTTTTTTCTCCTATTTTT	G in clade B
207	ATAGCAATTCCCATTCCGCCCAAAACCTTAGGAATTCCTTGATAGTTGGT/CATAAATTCGTAAGCCAGGTCGGCTGATACGCTTTAAAAAGGTTCTAGTTCTATATATT	T in clade F

	Sequence	SNP clade
208	CGAAGAAATTGACTTCGTATGGGCATTTTGCTGGCAGCTATGGAAATAGCTA/GCTCTAGCTACAGTTTCGGATACTCCGCCCATTTCATAAAGTATTCGACCTGGTTT	A in clade C,D
209	TTTAACAACGGCTACCCAATATTCGGGGGATCCCTTTCCCGAACCCATACGTGTTTCC/GGTG/CGGTCTTATTGTAACCGGTTTGTCGGGAAATATACGTACCCAGATTTTTCCA	C,G in clade G
210	CGAATATTTACTCTTTCCTGTCTTATTTGTTAATTCATAACCTTATCAAATAAGG/ACAATTTTTTTGGTTTGTTCCGCCATCCCACCCAATGAAGTATTGGGATTCTTT	G in clade G
211	TTTTCCCGCGAGACGGCCTGCAATTTTTACTTTTACTCCCTTTATATCC/TGTTTTTTTAGTTAATTCAATGGCTTTTTTCATTGCCTTTCGGAATGAAACTCTAT	C in clade G
212	CTACGTCCTCGAGCCCGAGGTCTGAATTTATTCATAATAGTACTCCTACTGACTTCC/GGCTTTAGTGATGAATAAATTTGCTTTGTCGAAATCCCTATAATGAGTAGCATTTGCT	C in clade B
213	ACCTCTCTGGATCCTCGAATTGAAAGAGAGATTGAGAGGGATCA/CAGAATCCTAATTCTCGCTATTTGGAATGGATCCAATTCTATTGAGTCTGACTCATAGTGATCATTTCTC	A in clade G
214	GATATGTCAAAAGCAGGTCTGATTACACCTATTCCTAATCCTAAATAGAATGTAAGGAT/CGTGGGGATTTCTATGTAAACAGAGTATCCTATTTCCATAGGCTCGAATGACCCCTTCTCATAATAAGAA	T in clade F
215	GGTATGGAATGAACTTATAATCTGATGATCGAGTCGATTCCATGATTATAAGTTCATA/TACCCTAGCGCCCATTCCCATTTTGGGCGGAACAGATCTACTAATTCTTTTATT	A in clade F,G
216	TTGTAGGGTGGATCTCGAAAGATAGGAAAGATCTCCCTCCAAGCCGTACATACG/AACTTTCATCGAATACGGCTTTCCACAGAATTCTATAGGGATCTATGAGATCGAG	G in clade G
217	CATTTCATGTTTCGAGGTCTCAAAAAAGGGCGTGGAAACAGATAGAAACTCTG/TGAATGGAAATTGAAAAGAAATGTAGCCCCAGTTCCTTCGGAAATGGTAAGATCTTTGGCG	G in clade A

	Sequence	SNP clade
218	GATGTCAAAAGGAAAGGGATGGAGTTTTTCTCGCTTTTGGCGTAGCAGGCCTCCCTTT/AAAGGGAGGCCCGCGCGACGGGCTATTAGCTCAGTGGTAGAGCGCGCCCCTGATAATTGCGTCGTTG	TTT in clade C,D
219	CTAGCCATAAGAGGAATGCTTGGTATAAATAAGCCACTTCTTGGTCTTCGACC/TCCCTAAGTCACTACGAGCGCCCC/TCGATCAGTGCAATGGGATGTGGCTATTTATCTATCTCTTGACTCGAAATGGGAGCAG	C,C in clade G
220	TGATCTTCATATCGATCTATTATCCACCTCTGCATCTATTCTTTCTTAGCTCTAAACGGGTGGAAGATCCATCCAATTTGGTTATATCATGGACTCAAAAAACGGAT	Ins in clade F
221	AGAACACAGATACATAACATAAAAAAAAGAATAAATAAGACGAAATTCGC/ACCTCCCCCTACATATTTAATTTCTTCTCCTATACAAAAACTAGCAAGACCTACTCCATT	C in clade B
222	GTTTTTAGTCCCCAATGAAGTACTAAAGGACCCTATCCTATTTCCTGTATTAC/ACATGAATTTTGGATAGATTTTGTGAAAAAAAAGAAACTCCACTCTTCGCTGTTG	C in clade B
223	TGAATATCCAACAAGAGGTTCCATTGAATGAATAACAGATCCGGATCCCAAGAAC/TAATAAAGCTTTCGAATAAGCATGAGTGATCAAATGGAATAAAGCAGCTTG	C in clade G
224	CTTATATAAACAAAAAATCTCAAATATCCCTCATCGTGAGACATATAATCG/ATCACTATAAATAAGAACCAGGATTCCTACAGTAGTAATTAGTATTAACATAATAGAAGT	G in clade G
225	AAAGTAAAACACTAGGAAAAGCCCATATGCGACGAAGATTTTTTGTTGCTGTC/TGGAAT/CAAGAAAAAGTCCAAACCCCATTGACATAATAACTGGAAGTGGGAGAAGAGGGA	C,T in clade G
226	TAATTTTTCAAAAATTTTCTCATTGAAACAATCAAAAAATAAGAATAGGTTTTGTTTTGTTGGTTAAAGTCAAAAAGTTAATGAAATAACTTCGTTACCTAGTTATTACCT	In in clade F

	Sequence	SNP clade
227	GAAAACCTTTGTATATATTCTATATTATTAAAACAAAGTCTAAAAAAAATATAG/TAATATGTTAAAAAACTCTTGTCTTATCCGCATTAGACAAAATGAAGTAAAAAAGAAT	G in clade A
228	TTTCAATATCTTTTAGTATCTAAGTATAAATACTAAGAAAAAGAAGAAAA/GATGGATTGATTTGCGGCAATAGATGTCTTTCACATACAACTAGAAAAAAGTA	A in clad C,D
229	ACAAACAAATAATAGGGTTTTGGGATAATATGAATTGACCTATCCCCC/AAAAAATTCCAATTATTTAATATGAATAATTAGGAATAATTAGGATTAATTAATGA	C in clade E
230	ACCAAACGAAGTCTATTTTAATGAAGATTCTAATGTCCTAAATTCTATGGAC/ATCTTCCAATCTCGACGATTCGCGAGAAAATAACTTAATATTCTTTTAATAA	C in clade G
231	GAATCTTCCAATCTCGACGATTCGCGAGAAAATAACTTAATATTCTTTTAATAAACCTA/GTTATTTCAACTTAGCCGCCATGGTGAAATTGGTAGACACGCTGCTCTTAGGAAGCAGTGCTC	A in clade A
232	TTCGAGTCCGAGTGGCGGCAGTCTCGAAAAAGAATACAATAGATTATAAAATAAAATGGATTCAATTCAATTCGAAATTTCCAATTTTGTAATGGGACCTTCTC CTTATGCTATTTGCAACTTTA	Ins in clade F
233	TGTTACTAAGCTATGCGACTCTTTTGTGCGGATCCTTATTATCCGCCGCTCTTCTAATC/GATTAGATTTCGAAAGAATTTAGATTTCTTTTCGAAAAAGAAGAAAAATGTTTT	C in clade G
234	TTCCAAATTATTACAAATATCAATTAATTGAGCGTTTGGATTCTTGGAGTTC/ATCGTGTCATTAGTCTAGGGTTTACCCTTTTAACCATAGGTATTCTTTGTGGAGCAGTATGGGCT	C in clade A
235	TCAAAAATTCGAGATAGATCTAATTAGACTCTTTTACTTTTTTCTGAATTTTTG/TAGTATTTCCACTATGGAATATAGAGCGGACTAGTAGAAGAAAAAAAATCCTATTTAGGA	G in clade E
236	GATACAGATTAAAAGAAAGAGTTCTCGCGGGCCGGAATCCTCAAAATTTG/TCGTTTGGAACATGAAATAGCTTGTATCCATAGAACATCTGTCGTAACATAGAT	G in clade A

	Sequence	SNP clade
237	TTACAAAAGTAATTAGCATTTTTGGCATTAACAGAAATTTTGGACTAGTAATG/TAGTCCAAAAAATACTACTAATTCCGCAACAAAACCACTCATTCCTGGTAAGGCAA	G in clade A
238	TAATGAAACCCATGTGAGAGACGGAGGAGTAGGCTATTCTTTTTTTGAAATTT/GCGTTGA/GCCAAGAGAAGTTGAAGCTGCATAGATTATTTGCATCGCTCCTATTATTACTAACC	T,A in clade G
239	CAATAAGATAGAGCCATGCTGCGGGTTGTCTCAGGTCCTAAATAAACGCGGACG/CTTAAAAAATCTGTTGGGCAGGCGGATTCGCATCTCTTACAACCCACACAATCTT	G in clade B
240	ACATTGAGTGCATCCTATACATGTATCATAAATTTTTACGGAATGTGACATTGGG/TCTATAAATTTTCCTTTTCAACATAAAAATTTTCGATCTGGTCAAAATGAA	G in clade E
241	AAAAATTTTCGATCTGGTCAAAATGAAATTAGTACTATATCAATCAAATGTATTA/GTAGACACCAGACGAAGCAATGGTTTATCCAAACTTCAACAAATAATGCAATATATTTCTTA	A in clade F
242	TATATTTCTTAATCCGTTTGTGAGAAAGCATGAAAAGAGCCAAGAGACTTG/TAATTTTGGGCTTCAACAATCATAATTATACGAATTGTATATACGAATTCG	G in clade A,B,F
243	AGAATACTATGGAATAACCTACTCAAAAAATAGATATTCTCAAATAATAAATAGTATTCATGTTAATATTTCATCAAATAATAAATAGTATTCATGTTAATATTTCATATTATTATTATATGTGTCCCTTTG	Ins in clade B
244	GCAGCCGCAAGGGCTATAACAAAAATTGCGAAAATGTCTCCTTTTAATTGGCGA/GCTATCAAATAGATCAGAAAATGTTACGAGATTTAGATTAATTGAATTCAGTATAAGTTC	A in clade G
245	TATTCGAAATATCTATGAAAAAGGTATGTTTCTTTCTCTTGTTTGAGAGA/GACTTTTGTGTTGAAAATATTCTTACTGTTATTGTAT	A in clade A,F,G
246	AATAAATAAGCTTTAGTTAATGTAATAAAGATACTCATTGTCATTTCTAG/AAATTCCAACCATTTTATTCATTTGGAAAAATCCAAAAAAGGATATATAGGG	G in clade A

	Sequence	SNP clade
247	GGTAATCTTTCACATTCCGCCAAAGAAGAAATTAGAAAAACCAGAAAACCTATG/AGGCTGACGCCAAAGATTCCATCCAAAAAAACCATATTTTGACTGTGCTTCAACTATA	G in clade B
248	CTGTCTGCTAGAATAATAAAAAACGCTTCGGAATTCATCTCATCCTTTATAATATAAA/TGGTACTTTTTCTTTGTTCAGCAATAACTTAATCTTGGAATAAAACACTCGTTAT	A in clade B
249	AAATTAGACCAAAGGAATTCTGTCTGCTAGAATAATAAAAAACGCTTCC/GGAATTCATCTCATCCTTTATAATATAATGGTACTTTTTCTTTGTTCAGCAATAA	C in clade G
250	TACTAATCCTTTATGTACTTTAGTGTTTCTAATCCCTCACTAACTTTTGAC/TGGATTCCCTTATGATTACAACTTTCTGTATCGGGAATCCCTTATTATTGCCCGCTTCAA	C in clade C1
251	TTTGTTTCACTCATATAGCTATCTAGTTTAACTTACTAACCTGAATATAGAATAAGAAAAGGAA/GGATAAATATTCAATGAATTTCAGAGGAAAAAGATCCTATTTTAACGAATCGCAC	G in clade G
252	ATAGCTTGAAGCAGTCCCAGGGGGCCAGCATATTCAGGACCAATACGTTGTTGTATTT/CGATGCGGATATTTCTCTTTCTAACCACACAATTACGAGTACTTCTATTGTGATTCCCAGT	T in clade F
253	GTCATGATATCAGCCAATTTCATTTTTTTGACTAGCTGAGGAAGAATTTGCAAATTAATAAAC/ACCGGGTGGACGAATTTTCCATCTCCAGGGGAAAAGACTATCATCTCCTACCAGAT	C in clade C1
254	CCTACCAGATAAATTCCTAATTCACCTTTTGGGGCTTCCACTCTTGCATAAAGCTCTTGC/TTTTGACAATTCAAAATTGGGTGAAGGTTTTTTACCAAGAAATCGATATTCAAAA	C in clade F
255	TTCGGAATTCTTTGCTTTCTTAAAGCGTCGGACTTCTAAATTCTCATAAGGGCCCCCC/AGGAATTTTTTCTACAGCCTGTTGAATAATTTTGATTGATTCCCTCATTTCACCGATTCGTACT	C in clade B
256	TCTAAATTCTCATAAGGGCCCCCAGGAATTTTTTCTACAGCCTGTTGAATAATTTTGATG/TGGATTCCCTCATTTCACCGATTCGTACTAAATAGCGTGCTAATGAATCCCCTTC TTTTTGCCAT	G in clade E4

	Sequence	SNP clade
257	TGTTGAATAATTTTGATTGATTCCCTCATTTCACCGATTCGTACTAAATAGCGC/TGCTAATGAATCCCCTTCTTTTTGCCATTGGACTTTCCAATCGAATTGATTGTAAGACTCG	C in clade G
258	CACATTCAGATCCGTTTTTTGAGTCCATGATATAACCAAATTGGATGGATCTTCCACCCGTTTAGAGCTAAGAAAGAATAGATGCAGAGGTGGATAATAGATCGATATGAAGATCATGAGCTGCCCCATA	Ins in clade F
259	ATTTCGAGTCAAGAGATAGATAAATAGCCACATCCCATTGCACTGATCGG/AGGGCGCTCGTAGTGACTTAGGGG/AGTCGAAGACCAAGAAGTGGCTTATTTATACCAAGCATTCCTCTTATGGCTAGATCCAACCT	G,G in clade G
260	ATTATCAGGGGCGCGCTCTACCACTGAGCTAATAGCCCGTCGCGCGGGCCTCCCAAA/TTTGGGAGGCCTGCTACGCCAAAAGCGAGAAAAACTCCATCCCTTTCCTTTTGACATCCCCATGCCG	AAA in clades C,D
261	ATCTTACCATTTCCGAAGGAACTGGGGCTACATTTCTTTTCAATTTCCATTCC/AAGAGTTTCTATCTGTTTCCACGCCCTTTTTTGAGACCTCGAAACATGAAATGG	C in clade A
262	TACTCGATCTCATAGATCCCTATAGAATTCTGTGGAAAGCCGTATTCGATGAAAGTC/TGTATGTACGGCTTGGAGGGAGATCTTTCCTATCTTTCGAGATCCACCCTACAATATGGGG	C in clade G
263	AACTGGAATAAAAGAATTAGTAGATCTGTTCCGCCCAAAATGGGAATGGGCGCTAGGGTT/AATGAACTTATAATCATGGAATCGACTCGATCATCAGATTATAAGTTCATTCCATACCGGACCAG	T in clade F,G
264	GAAGGGGTCATTCGAGCCTATGGAAATAGGATACTCTGTTTACATAGAAATCCCCACA/GTCCTTACATTCTATTTAGGATTAGGAATAGGTGTAATCAGACCTGCTTTTGACATATCTA	A in clade F
265	CTATGAGTCAGACTCAATAGAATTGGATCCATTCCAAATAGCGAGAATTAGGATTCTT/GGATCCCTCTCAATCTCTCTTTCAATTCGAGGATCCAGAGAGGTGTTTTCATAG	T in clade G

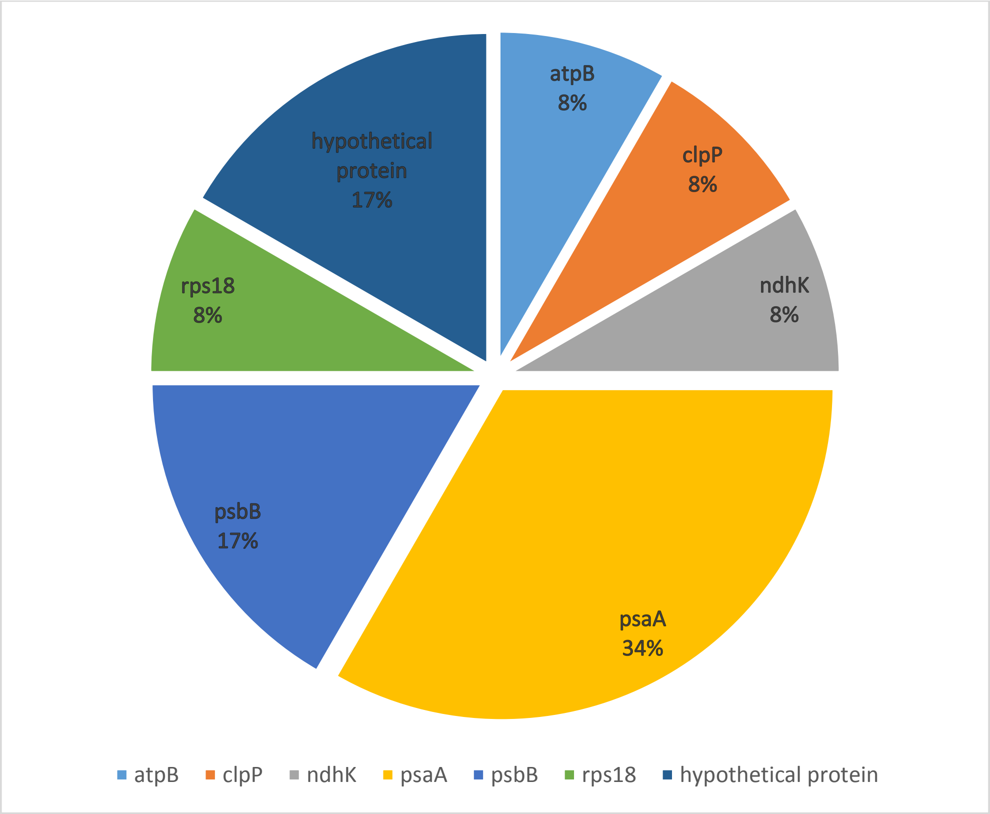


Figure 22 FNPs percentages found in different genes in Asian wild rice chloroplast genomes

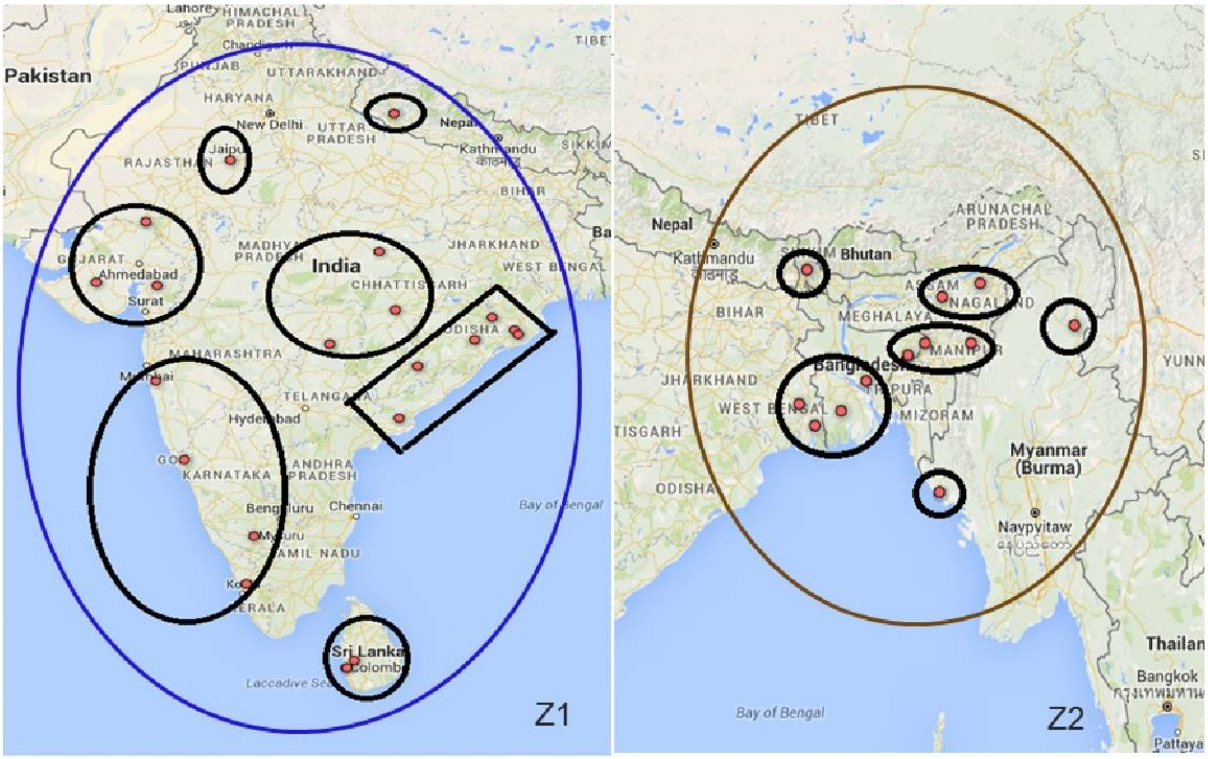


Figure 23 Geographical distribution of Asian wild rice zone 1 including India and zone 2 India and Burma. High coverage samples was selected from each circle.

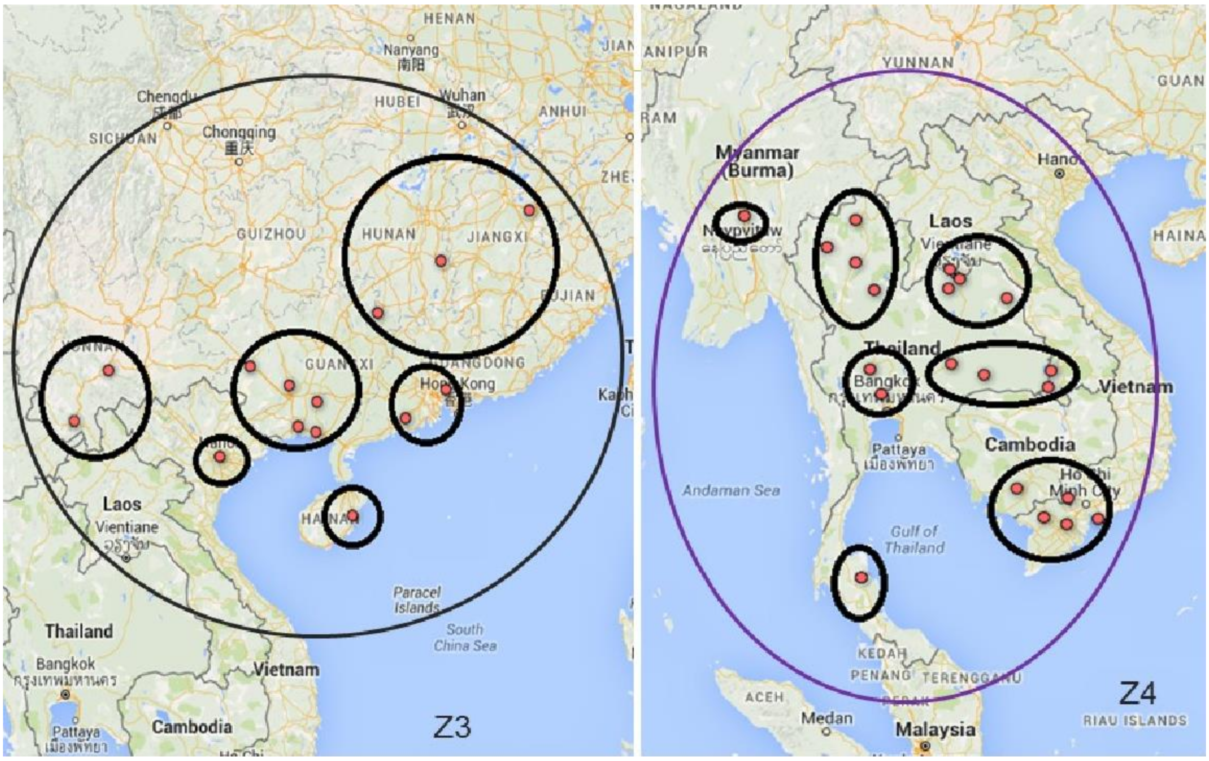


Figure 24 Geographical distribution of Asian wild rice zone 3 including China and zone 4 including Thailand, Vietnam and Cambodia. High coverage samples were selected from each circle

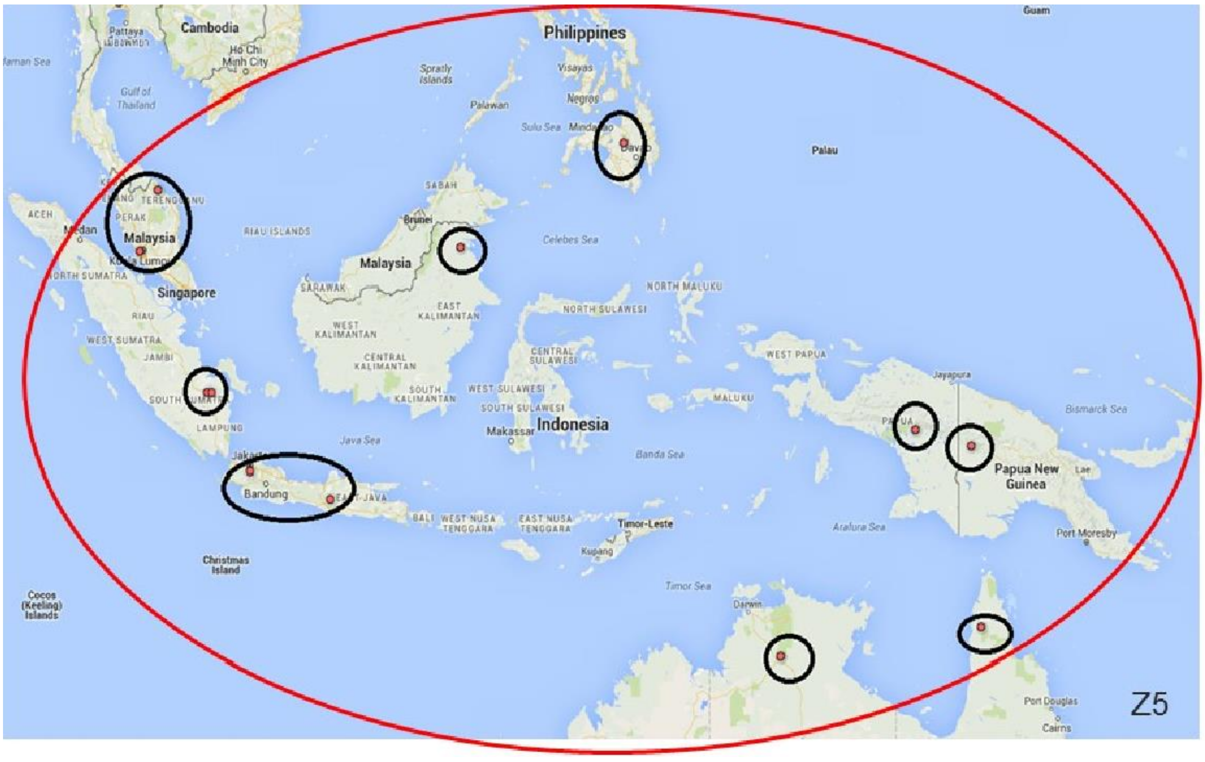


Figure 25 Geographical distribution of Asian wild rice zone 5 including Oceania Australia, Papua New Guinea, Indonesia, Malaysia and Singapore. High coverage samples were selected from each circle.

3 Appendix 3. Chloroplast Assembly Pipeline

3.1 Abbreviations

MA-approach: mapping assembly approach

DA-approach: *de novo* assembly approach

CpN: *Oryza sativa* ssp *japonica* cv Nipponbare (Genbank accession GU592207)

CpW: Australian wild rice Taxa-A (Genbank accession KF428978);

CpL: *Oryza longistaminata* (Genbank accession KM881641)

CpO: *Oryza officianalis* (Genbank accession KM881643);

CpWt: *Triticum aestivum* (Genbank accession NC_002762.1).

CAP: chloroplast assembly pipeline

M-component: Mapping assembly component

MOpt-process: Mapping Optimisation Process

MImp-process: Mapping Improvement process

D-component: *de novo*-assembly-component

D-process: *de novo* assembly process

DImp-process: *de novo* improvement process

OsNipp35bp-PEreads; 35bp paired end Illumina reads of *O. sativa* Nipponbare (GU592207)

R-tool; Read mapping tool

S-tool; Structural variant analysis and Local Realignment" tool

P-tool; Paired-end read extraction and remapping tool

MOpt:R+S; Mapping optimisation using the R and S tool

MOpt:R+S+P; Mapping optimisation using the R, S and P tools

MOpt:R+P: Mapping optimisation using the R and P tools

MOpt:R+P+S; Mapping optimisation using the R, P and S P tools

C, F; Cost and Fraction mapping settings

TaxaA100bp-PE reads; 100 bp paired-end reads of the Australian wild rice Taxa-A

3.2 Sequence data statistics

We used achieved whole genome NGS paired end data available in our research group for all analysis. NGS data of *O. sativa* ssp *japonica* cv Nipponbare consisted of 35 bp paired end reads (**OsNipp35bp-PEreads**), generated on an Illumina GAII analyser, and the Cp sequence for this genotype (accession GU592207) was published (Nock et al., 2011). Summary statistics of the sequence data sets trimmed at a quality score limit of 0.01 (>20 PHRED score) is shown in (Table 22) CLC Bio Genomics Workbench (CLC-GWB, CLC-Bio, QIAGEN, Denmark) was used for the mapping assembly and for *de novo* assembly of Cp genome sequences. Geneious R9 (Biomaters, USA) was used to align Cp sequences and identify number of mismatches and details of the variants. Clone Manager (SciEd, USA) was used to assemble Cp contigs and derive a consensus Cp sequences. Details of CpN and reference Cp sequences used and mismatches between them

Accession numbers of Cp sequences used as reference sequences and sourced from NCBI and GenBank are as follows: CpN, *Oryza sativa* ssp *japonica* cv Nipponbare (Genbank accession GU592207); CpW, Australian wild rice Taxa-A (Genbank accession KF428978); CpL, *Oryza longistaminata* (Genbank accession KM881641); CpO, *Oryza officianalis* (Genbank accession KM881643); CpWt, *Triticum aestivum* (Genbank accession NC_002762.1). Number of mismatches between the publically available CpN and CpW, CpL, CpO and CpWt are 125, 141, 670 and 7,499 mismatches respectively.

3.3 Abbreviations and denotations used to identify assembled Cp sequences

Cp sequences derived from the *de novo* assembly approach were denoted with one identifier, “-D”. In the example CpN-D, *de novo* assembled Cp sequence was generated using reads of *O. sativa* cv Nipponbare (GU592207). Chloroplast sequences derived from the mapping assembly approach are denoted by two identifiers. The first identifier provides details of the reference Cp used while the second identifier provides details of the assembled Cp. In the example CpW/CpN-XXX, the two identifiers are CpW/ and CpN-XXX, indicating that CpW (Australian wild rice Taxa-A, Genbank accession KF428978) was used as a reference Cp sequence to obtain a mapping assembled Cp genome sequence of CpN (*O. sativa* cv Nipponbare, GU592207) with details after the hyphen indicating the process and settings used for the mapping assembly.

3.4 Chloroplast Assembly Pipeline (CAP details)

The CAP is structured to obtain assembled Cp genome sequences using a MA-approach and the DA-approach and are identified as the mapping-assembly-component (M-component) and the *de novo*-assembly-component (D-component) respectively (Figure 26). The M-component consists of two process steps; the Mapping Optimisation Process (MOpt-process) and the Mapping Improvement process (MImp-process). The D-component also consist of two process steps; the *de novo* assembly process (D-process) and the *de novo* improvement process (DImp-process). Both these process steps are designed to sequentially improve the assembly process leading to least number of errors in the assembled Cp genome sequences derived from these two components of the CAP. Consequently, reduced number of mismatches identified between the M-component and the D-component would result in reduced manual curation and increased confidence in the accuracy of the final assembled Cp genome sequence.

3.5 Mapping Optimisation (MOpt) process and the CM-Rule: rationale

The 35 bp paired end Illumina reads of *O. sativa* Nipponbare (GU592207), henceforth referred to as **OsNipp35bp-PEreads,** were shown to be of sufficient quality and length to generate a mapping-assembly-derived Cp sequence of Nipponbare perfectly matched the reference Cp sequence used (another accession of Nipponbare (Genbank accession AY522330.1) (Nock et al., 2011). However, this earlier study (Nock et al., 2011), did not indicate if the reported mapping parameters were optimal settings and if an accurate Cp genome sequence of Nipponbare could be generated using a reference Cp sequence of a closely related species. Hence, the availability of an accurate Cp sequence of *O. sativa* cv Nipponbare (GU592207) and sequence reads of the same genotype prompted us to assess the optimum mapping (MOpt) parameters required when using the CpN as a reference Cp sequence and also after using a Cp sequence of closely related species as reference Cp sequence.

The OsNipp35-PEreads were mapped to the CpN using the read mapping tool (R-tool) at combinations of two “Cost settings” (C-setting) and “Fraction settings (F-setting). Each C-setting consists of mismatch-cost, insertion-cost and deletion-cost settings, while the F-setting consists of length-fraction and similarity-fraction. We used a combination of two C-settings (C1 and C2) and six F-settings (F0, to F5) settings (**C&F-settings**) (Table 23) and these steps using the R-tool are collectively referred to as the MOpt:R steps (Figure 27). As the CpN was used as the reference Cp sequence, the Cp sequences generated at the MOpt:R steps were denoted with a prefix CpN/(e.g. CpN/CpN-C1F3-MOpt:R, Table 23). An accurate mapping-derived Cp sequence was possible only with the most stringent mapping settings of C1F0 and C2F0 but not at any other mapping setting as increased mismatches were observed with reduced stringency in the C and F setting (Figure 28 **i, ii, blue bars).** We attempted to improve the Cp sequences derived from the MOpt:R step using the “Structural variant analysis and Local Realignment” tool (S-tool), referred to as the MOpt:R+S step (Figure 27). However, the corresponding Cp sequences from the MOpt-R+S step also failed to show complete homology to the CpN sequence with either no change, a slight increase or a reduction in the number of mismatches (Figure 28 **i, ii, yellow bars**). Mismatches were comprised predominantly of the T nucleotide at homopolymer regions (Table 24) due to the mapping of single reads from broken paired-end reads to these homopolymer regions (Figure 28). Filtering out all mapped single reads, by implementing the “Paired-end read extraction and remapping tool (P-tool), applied after the MOpt:R step and referred to as the MOpt:R+P step or after the MOpt:R+S step and referred to as the MOpt:R+S+P step (Figure 27), resulted in an accurate consensus Cp sequences with no mismatches when compared to the CpN sequence (Figure 27). Hence, we identified that even when using paired-end reads as input data, reads with homopolymer sequence can map as single reads to corresponding homopolymer regions causing errors in the assembled Cp sequences. We also identified that using the P-tool can completely eliminate these errors leading to an accurate Cp sequence even when using 35bp paired-end Illumina data. Having identified the importance of the P-tool applied in combination with the R-tool and S-tool as the R+P+S-step or R+S+P-step, we tested if these steps could be used to generate an accurate Cp sequence when using the OsNipp35bp-PEreads and a reference Cp sequence of a species closely related to *O. sativa*. Here we chose CpW (Australian wild rice Taxa-A, Genbank accession KF428978) as the reference Cp sequence. All Cp sequences, generated at the various C and F settings (Table 23) with the R-tool, P-tool and S-tool applied (Figure 27), when aligned to the CpN sequence, none showed complete homology, but instead had several mismatches indicating an accurate Cp sequence was not obtained (Figure 29 **i, ii**). Mismatches in all Cp sequences derived from the MOpt:R step, showed a trend in the number of mismatches, with the highest mismatches observed at the most stringent and at the most relaxed C and F settings while the lowest number of 16 mismatches was observed at C and F settings in between (Figure 29 **i, ii, blue-bars or red-line curve**). A similar trend of mismatches was observed in the Cp sequences derived from the MOpt:R+S step but the lowest number of mismatches was 5 (Figure 29 **i, ii, blue-broken-line-bordered-yellow-bars or blue-broken-line-curve**). Implementation of the P-tool as the Mopt:R+P step reduced the mismatches to as low as 14 in some of the assembled Cp sequences (Figure 29 **i, ii, black-broken-line-bordered-brown-bars or black-broken-line-curve**). Implementation of the P-tool as the MOpt:R+S+P step led to further reduction in mismatches with the least mismatches of 3 at C1F3 and C2F3 onwards (Figure 29 **i, ii, blue-bold-line-bordered-brown-bars or blue-line-curve which is super imposed by the black-line-curve**). The implementation of the P-tool as the MOpt:R+P+S step also led to further reduction in mismatches with the least mismatches also of 3 at

C1F3 and C2F3 onwards (Figure 29 **i, ii black-line-bordered-yellow-bars or black-line-curve)**. Results in (Figure 29 **i, ii)** indicate that the P-tool and the S-tool both contribute to reducing the number of mismatches. The P-tool reduced mismatches due to single nucleotide variants (SNPs) and the multi-nucleotide variants (MNVs) (Figure 29 **i**, **ii, insert figures with variant distributions**). The S-tool generally has no impact on reducing variants due to SNPs and MNVs but contributes in reducing the variants due to insertions and deletions. Results in **(**Figure 29 **i, ii)** indicate three key points; 1, the P-tool can be applied before or after the S-tool as the number of mismatches were reduced to 3 in both cases. 2, the application of the MOpt:R+S+P step and the MOpt:R+P+S step led to a consistent number of mismatches in the Cp sequences derived at any given C and F setting. 3, the consistent number mismatches of 3 was represented in Cp sequences at most of the C and F settings (Figure 29 **i, ii**). Mismatches were also determined by comparing the Cp sequences to CpW. As indicated earlier, the CpN when aligned to the CpW sequence had 125 mismatches. Since an accurate Cp sequence was not assembled (Figure 29 **, i, ii**), all of the consensus Cp sequences when compared to the CpW sequence, as expected showed mismatches totalling above or below the expected 125 mismatches (Figure 29 **iii, iv**). The trend in mismatches of Cp sequences derived from the MOpt:R+S+P step and the MOpt:R+P+S step as shown in (Figure 29 **iii, iv)** was similar to that observed in (Figure 29 **i, ii)**. We observed 124 mismatches as the consistent number of mismatches in the consensus Cp sequences derived from most of the C and F (Figure 29 **iii, iv**) and these C and F settings were the same that had 3 mismatches when the Cp sequences was compared to CpN (Figure 29 **i, ii)**. These results indicate that the number of mismatches in the consensus Cp sequences if consistent across most of the C and F settings at the MOpt-R+S+P step and at the MOpt-R+P+S, which we refer to as the "Consistent-Mismatch Rule" (**CM-rule**), can be used as an indicator to identify the optimal C and F settings to obtain a Cp sequences with the least number of mismatches. Having used the CpW as the reference, the CM-rule when applied to the data in (Figure 29 **iii, iv)** alone successfully identified C1F3, C1F4, C1F5, C2F3, C2F4 and C2F5 as the optimum Cost and Fraction settings as the corresponding assembled Cp sequences had the least number of 3 mismatches when compared to CpN. In addition, the 3 mismatches consisted of variants of the same nature in all of the consensus Cp sequences. The consensus Cp sequence, CpW/CpN-C1F3-MOpt:R+P+S, generated at the MOpt process at C1F3 had 3 and 124 mismatches when compared to CpN and CpW respectively (Figure 29).

3.5.1 Mapping Improvement process: rationale

We determined that the Cp sequence obtained from the MOpt-process could be further improved by implementing the R+P+S step, twice in sequence, which we refer to as the Mapping Improvement process (MImp-process) (Figure 26). In the first step referred to as the Mapping

Improvement process step-1 (**MImp-1**), the chosen Cp sequence from the MOpt-process was taken as a reference Cp sequence and subjected to the R+P+S-tool. The Cp sequence from the MIpm-1 step was used as a reference Cp and subjected to the R+P+S-tool and this step is referred to as the Mapping Improvement process step-2 (**MImp-2**). The MImp-process improved the Cp sequence derived from the MOpt-process, as the mismatches were reduced to 2 at the MImp-1 step with no reduction further at the MImp-2 step (Figure 30 **i**). Data in (Figure 30) with mismatches determined by comparing to CpW indicates no change to the number of 124 mismatches because the change in nucleotide corresponding to the reduction in the single mismatch when compared to CpN was still a mismatch when compared to CpW. Hence, data in (Figure 30 **ii)** indicates the MImp-process reduced mismatches and the accuracy of the Cp derived from the MOpt process. The Cp sequence from the M-component (MOpt and MImp steps) was found to be 134,550 bp in length with 2 mismatches when compared to CpN.

3.6 *De novo*-assembly process and *de novo*-improvement process: rationale

The *de novo* assembly process (D-process) is where whole genome sequence reads were subjected to the *de novo* assembly tool in CLC-GWB at various combinations of "Word size" setting (W-setting) and "Bubble size" setting (B-setting), with scaffolding and in the "Fast" mode. Cp-specific contigs, identified by BLAST analysis against the same reference Cp sequence used in the M-component **(**Figure 26), were updated and then aligned to a reference sequence to identify overlaps and gaps. Additional *de novo* assembly at additional W- and B-settings was undertaken to generate additional contigs for closing gaps. The Cp sequence derived from the D-process was generally denoted as CpX-D where X represents the name of the species or genotype.

We determined that the Cp sequence obtained from the D-process can be further improved using the *de novo* improvement (DImp) process, which is similar to the MImp (Figure 26). The Cp sequence generated from the D-process, CpN-D, had 17 mismatches over 96 bases when compared to CpN (Figure 30**, iii red bar**). The application of the 3-Map-tool of R+P+S at the C1F3 setting, the optimal C and F setting from the MOpt-process (Figure 29), was applied to the CpN-D sequence twice and this process is referred to as the *de novo* assembly improvement process (DImp-process) comprising of the DImp1-step and DImp2-steps respectively. The DImp-process led to reducing the 17 mismatches in CpN-D to 6 mismatches and 4 mismatches in the CpN-D/CpN-C1F3DImp1:R+P+S and the CpN-DImp1/CpN-C1F3DImp2:R+P+S sequences obtained from the DImp1-step and the DImp2-step respectively (Figure 30**, iii, black double bordered yellow bars**). Thus, the DImp-process is an important tool which can be applied to reduce errors in the Cp sequence obtained from the D-process. The Cp sequence from the D-component was found to be 134,465 bp in length with 4 mismatches when compared to CpN.

3.7 Manual curation to obtain a Cp-CAP

Cp sequences derived from the M-component and the D-component were aligned, mismatches determined and nucleotide calls revised by manual curation. The process of manual curation of the mismatches involved observing the reads mapped at the mismatch position, recording the mismatch position, the number of nucleotides covering the mismatch and providing appropriate evidence on why a mismatch was considered to be a likely error and warranted correction.

We identified 18 and 5 mismatches on comparing the M-component derived Cp sequence (CpW/CpN-C1F3MImp2:R+P+S, 134,550 bp) and the *de novo* assembly derived CpN-D (134,469 bp) and the D-component derived Cp sequence (CpN-DImp1/CpN-C1F3DImp2:R+P+S , 134,465 bp) respectively. Erroneous mismatches were identified by examining the reads mapped to the mismatch positions of all these sequences (Table 25 and Table 26). In the M-component derived Cp and the CpN-D comparison, with 18 mismatches covering 113 nucleotides, 2 mismatches were due to missing data in CpN-D due to -NN- in the contigs used to generate this sequence. In addition, 17 out of the 18 mismatches were due to errors in the *do novo* generated CpN-D sequence. In the M-component derived Cp and the D-component derived Cp sequence comparison, with 5 mismatches covering 89 nucleotides, 4 mismatches were due to errors in the D-component derived Cp sequence. The manual curation of all 3 sequences led to revised sequences of the same length of 134,550 bp and this was represented as the assembled Cp from the CAP and referred to as CpW/CpN-CAP. The CpW/CpN-CAP differed from CpN by one less T nucleotide in the homopolymer region at 78,440, where CpN is 134,551 bp in length and the homopolymer consists of 17 T nucleotides. This error was not resolved by the CAP due to the absence of mapped reads with sequence spanning both sides of the homopolymer (Figure 31) essentially as the CpW used as the reference had 16 T nucleotides. Using 100bp paired reads would have resolved this short homopolymer region of 17T nucleotides but not extensively longer homopolymer regions as is observed with genome assemblies using illumina reads. However, the CAP is robust enough to generate an almost accurate Cp sequence even when using short 35 bp PE reads and this robustness was tested using 100bp illumina reads of the Australian Wild rice Taxa-A as explained in the results. The implementation of the MOpt-process, the MImp-process and the DImp-process led to the manual curation of 5 mismatches covering 89 bases, a much better option than the non-implementation of these processes resulting in a worst case scenario with the manual curation of 54 mismatches covering 224 bases after comparison of the CpW/CpN-C1F5:R, derived from the worst C1F5 setting, and the *de novo* assembled CpN-D sequence. The CAP is outlined in (Figure 26).

3.7.1 Important steps in the CAP

Availability of an accurate Cp sequence of *Oryza sativa* cv Nipponbare at NCBI (CpN, GenBank accession GU592207), was used to assess the accuracy of CAP-derived Cp sequences using the **OsNipp35bp-PEreads** of the same accession. We identified that use of the **S-tool in CLC** corrected assembly related errors but not occurring in homopolymer regions (Table 24) resulting from spurious non Cp-specific reads mapping to the corresponding homopolymer regions (Figure 28 **iii, iv**), and use of the P-tool resolved these errors by filtering out the single mapped reads (Figure 28 **i, ii**). This analysis provided a key outcomes, where the **P-Tool is essential in reducing** errors due to single mapped reads in any mapping assembly-derived Cp sequence. We also identified the CM-rule, which allows the generation of a highly accurate Cp sequence from the MOpt-process of the M-component of the CAP by identifying the optimal R-tool settings in conjunction with P-tool and the S-tool (Figure 29). We demonstrated that the Cp sequence derived from the MOpt-process can be further improved by applying the R+P+S step in the MImp process (Figure 30 **i**). We demonstrated that the Cp sequence from the D-process can be improved by applying the DImp process, similar to the MImp-process, as it reduced the 17 mismatches covering 96 bases in CpN-D to 4 mismatches covering 86 bases (Figure 30 **iii**). All the steps in the CAP lead to the generation of with reduced errors, if any, to ultimately reduce the manual curation process to generate an accurate Cp sequence. Results from the manual curation of the Cp sequence from the M-component (Table 25 and Table 26**)** clearly indicates some key findings. Mismatches in CpN-D sequence identified when compared to Cp sequence from the M-Component, were overwhelmingly due to errors in the *de novo* generated CpN-D sequence and the DImp-process greatly reduced the mismatches sequence indicating the value of using this improvement process. The Cp sequence derived using CpW as the reference and from the M-component of the process with 2 mismatches over 5 bases (Figure 30 **i**), was more accurate than that derived from the D-component of the process which had 4 errors over 86 bases (Figure 30 **iii**).

3.8 M-Component robustness- OsNipp35bp-PEreads and CpW as a reference

Assessing the robustness of the M-component of the CA-pipeline using the CpW as a reference Cp sequences demonstrates the utility of this process in conjunction with the D-component of the CAP to generate a Cp sequence with least number of errors (Figure 26). In addition, an accurate Cp sequence was generated even when using very short reads, in this case under 35bp PE reads, with the CAP.

Table 22 Summary statistics of the next generation sequence data used for the assembly of chloroplast genome sequence.

Sample details	Sequence data details before trimming		Sequence data details after trimming			
	Paired end reads	Average length bp	Percentage trimmed	Total Number of reads	Paired end reads	Average length bp
Oryza sativa ssp japonica cv Nipponbare GenBank accession GU592207	9,689,084	36	99.80%	9,669,352	9,653,208	32.6

Table 23 Read-mapping parameters and their settings details used in the Mapping-optimisation process at the read mapping step (MOpt:R) and using the 35bp pared-end Illumina reads of *Oryza* sativa cv. Nipponbare.

Cost (C) mapping parameters- Mismatch, Insertion and Deletion Cost	Fraction (F) mapping parameters- Length and Similarity Fraction	Mapping-derived consensus sequences
2,3,3 (C1)	1,1 (F0)	CpN/CpN-C1F0-MOpt:R
	1,0.95 (F1)	CpN/CpN-C1F1-MOpt:R
	1,0.9 (F2)	CpN/CpN-C1F2-MOpt:R
	1,0.8 (F3)	CpN/CpN-C1F3-MOpt:R
	0.8,0.8 (F4)	CpN/CpN-C1F4-MOpt:R
	0.8,0.5 (F5)	CpN/CpN-C1F5-MOpt:R
1,2,2 (C2)	1,1 (F0)	CpN/CpN-C2F0-MOpt:R
	1,0.95 (F1)	CpN/CpN-C2F1-MOpt:R
	1,0.9 (F2)	CpN/CpN-C2F2-MOpt:R
	1,0.8 (F3)	CpN/CpN-C2F3-MOpt:R
	0.8,0.8 (F4)	CpN/CpN-C2F4-MOpt:R
	0.8,0.5 (F5)	CpN/CpN-C2F5-MOpt:R

Increasing C-setting and F-setting values represents decreasing stringency in mapping of reads to a reference Cp. Mapping-derived consensus sequences are denoted with C and F codes, representing the Cost and Fraction settings, and with CpN/ to indicate the publically available chloroplast sequence of *Oryza sativa* Nipponbare (GU592207) used as a reference Cp.

Table 24 Details of mismatches between different CpN/mapping consensus sequences derived when O. sativa Nipponbare 35bp Illumina reads were mapped to the O.sativa chloroplast sequence (GU592207) under various mapping settings.

Reference position	Reference base(s)	Variant nucleotide / variant frequency / Spurious single reads mapped				
		MOpt:R at C1F1 with 1 mismatch	MOpt:R at C1F2 with 1 mismatch	MOpt:R at C1F3 with 5 mismatches	MOpt:R at C1F4 with 7 mismatches	MOpt:R at C1F5 with 8 mismatches
29351-29352	TT				AA / 35 / Yes	
29376	G					
36428	A					
45579	A				del / 53 / Yes	
46061	T					
46065-46069	ACATG					
46086	A					
46090	A					
46094	T					
60137	del					
65707	C					
66336	T					
66352	T					
73151	A					T / 64 / Yes
78410	A					
78414	AA					
78419	AA					
78423	C	T / 62 / Yes	T / 78 / Yes	T / 75 / Yes	T / 81 / Yes	T / 76 / Yes
78441-78442	AA			TT / 90 / Yes	TT / 87 / Yes	TT / 79 /Yes
78444	C			T / 98 / Yes	T / 97 / Yes	T / 92 / Yes
78446-78447	CC			TT / 98 / Yes	TT / 97 / Yes	TT / 93 / Yes
78455	A			T / 97 / Yes	T / 97 / Yes	T / 95 / Yes
78461	C					
102132	T					C / 73 / Yes
102134	G					T / 74 / Yes
104746	A					
Reference position	Reference base/es	MOpt:R at C2F1 with 1 mismatch	MOpt:R at C2F2 with 1 mismatch	MOpt:R at C2F3 with 5 mismatches	MOpt:R at C2F4 with 16 mismatches	MOpt:R at C2F5 with 18 mismatches
					AA / 56 / Yes	
29376	G				del / 58 / Yes	del / 57/ Yes
36428	A					T / 54 / Yes
45579	A				del / 56 / Yes	
46061	T				A / 56 / Yes	A / 53 / Yes

46065-46069	ACATG				GATAT / 66 / Yes	GATAT / 63 / Yes
46086	A				del / 64 / Yes	del / 58 / Yes
46090	A				C / 64 / Yes	
46094	T				A / 48 / Yes	
60137	del					G / 67 / Yes
65707	C					del / 67 / Yes
66336	T					G / 56 / Yes
66352	T					A/ 83 / Yes
73151	A					T / 65 / Yes
78410	A				T / 47 / Yes	
78414	AA				TT / 43 / Yes	
78419	AA				TT / 63 / Yes	
78423	C	T / 60 / Yes	T / 81 / Yes	T / 65 / Yes	T / 65 / Yes	T / 46 / Yes
78441-78442	AA			TT / 74 / Yes	TT / 73 / Yes	TT /67 / Yes
78444	C			T / 93 / Yes	T / 93 / Yes	T / 90/ Yes
78446-78447	CC			TT / 97 / Yes	TT / 96 / Yes	TT / 94 / Yes
78455	A			T / 96 / Yes	T / 95 / Yes	T / 93 / Yes
78461	C					
102132	T					C / 76 / Yes
102134	G					T / 74 / Yes
104746	A					T / 60 / Yes

Table 25 Manual curation of mismatches between chloroplast (Cp) sequences.

Start bp	end bp	Description of mismatches in: CpN-D or the DImp2-process derived Cp sequence		Mismatches between the M-Component derived Cp sequence (CpW/CpN-C1F3-MImp2:R+P+S, 134,550 bases) and the D-process derived Cp sequence (CpN-D, 134,469 bases) and the DImp2-process derived Cp sequence (CpN-D-DImp1/CpN-C1F3-DImp2:R+P+S, 134,465 bases)							
				Outcome of manual curation for				Outcome of manual curation for;			
		type	Length (with gaps)	M-component derived Cp sequence		D-process derived Cp sequence CpN-D		M-component derived Cp sequence		DImp2-process derived Cp sequence	
24687	24707	Deletion	21	Yes	0	No	21	Yes	0	No	21
43019	43019	SNP (transition)	1	Yes	0	No	0				
46096	46153	Deletion	58					Yes	0	No	58
46097	46146	Deletion	58	Yes	0	No	50				
55846	55846	Insertion (tandem repeat)	1	Yes	0	No	-1				
55849	55849	SNP (transition)	1	Yes	0	No	0				
55853	55853	SNP (transition)	1	Yes	0	No	0				
55912	55912	SNP (transition)	1	Yes	0	No	0				
57053	57060	Deletion	13	Yes	0	No	8				
73152	73152	SNP (transversion)	1	Yes	0	No	0				
78424	78424	SNP (transition)	1	Yes	0	No	0				
78439	78442	Deletion	4					Yes	0	No	4
78441	78442	Substitution	2	Yes	0	No	0				
78444	78444	SNP (transition)	1	Yes	0	No	0				
78446	78447	Deletion	2	Yes	0	No	1				
78455	78455	SNP (transversion)	1	Yes	0	No	0				
100654	100655	Deletion	2					Yes	0	No	2
100655	100656	Deletion	2	Yes	0	No	2				
102132	102132	SNP (transition)	1	Yes	0	No	0				
102134	102134	SNP (transversion)	1	Yes	0	No	0				
105791	105794	Substitution	4	No	0	Yes	0	No	0	Yes	0
TOTAL mismatches/variants				1		17		1		4	
Manual curation: change in length (bp)				0		81		0		85	
Manual-curation led Cp final consensus length in bp.				Revised M-component Cp 134,550 + 0 = 134,550 bp		Revised D-process Cp, CpN-D 134,469 + 81 = 134,550 bp		Revised M-component CP 134,550 + 0 = 134,550 bp		Revised DImp2-process Cp 134,465 + 85 = 134,550 bp	

Table 26 Manual curation of chloroplast (Cp) sequences derived from the Mapping Assembly Component (M-Component) and two sequences from the Denovo-Assembly Component (D-component); the Denovo Assembly process (D-process) and from the denovo Improvement process (DImp2-process). All Cp sequences were generated using 35 bp paired end Illumina reads of *O. sativa* Nipponbare (accession GU592207) and the Cp (KF428978) of the Australian Wild rice Taxa-A (CpW) was used as a reference Cp sequence for the M-process. All analysis steps were undertake using CLC genomics Workbench.

Start bp	End bp	Sequence	Mismatches between the M-Component derived Cp sequence (CpW/CpN-C1F3-MImp2:R+P+S, 134,550 bases) and the D-process derived Cp sequence (CpN-D, 134,469 bases) and the DImp2-process derived Cp sequence (CpN-D-DImp1/CpN-C1F3-DImp2:R+P+S, 134,465 bases)						
			Description of mismatches in: CpN-D or the DImp2-process derived Cp sequence			Outcome of manual curation for;			
						M-component derived Cp sequence (134,550 bp)		D-process derived Cp sequence CpN-D (134,469 bases)	
			Type	Length (with gaps)	Sequence	Correct Reason when not correct	Mismatch included (+) or deleted (-)	Correct Reason when not correct	Mismatch included (+) or deleted (-)
24687	24707	AATTGTCGAATTATACTCAGC	Deletion	21		Yes	0	No, The presence of a 8 base sequence TACTCAGC as a repeat is clearly present in the reads. The De novo assembly is erronous as one of this repeat region is represented which can be noticed when the reads overlap at the repeat region in updated contig file .	21
43019	43019	G	SNP (transition)	1	A	Yes	0	No, The presence of spurious single reads in the updated contig file	0
46096	46153	TACGAAAACATAATAAAGAGAACATGCGAATTTCTTGTATTTTCAGTCCATCATTATA	Deletion	58					

46097	46146	TATTATATACGAAAACATAATAAAGAGAACATGCGAATTTCTTGTATTTTCAGTCCAT	Deletion	58		Yes	0	No, Has NNN in this region as no contigs in this region covering 58 bases So 58 base sequence to be insertrd minus the 8Ns = 50 bases difference	50
55846	55846		Insertion (tandem repeat)	1	A	Yes	0	No, Presence of singlespurious reads generating an extra A in the homopolymer A region.	-1
55849	55849	T	SNP (transition)	1	C	Yes	0	No, Presence of singlespurious reads generating an extra A in the homopolymer A region.	0
55853	55853	A	SNP (transition)	1	G	Yes	0	No, Presence of singlespurious reads generating an extra A in the homopolymer A region.	0
55912	55912	T	SNP (transition)	1	C	Yes	0	No, Presence of singlespurious reads generating an extra A in the homopolymer A region.	0

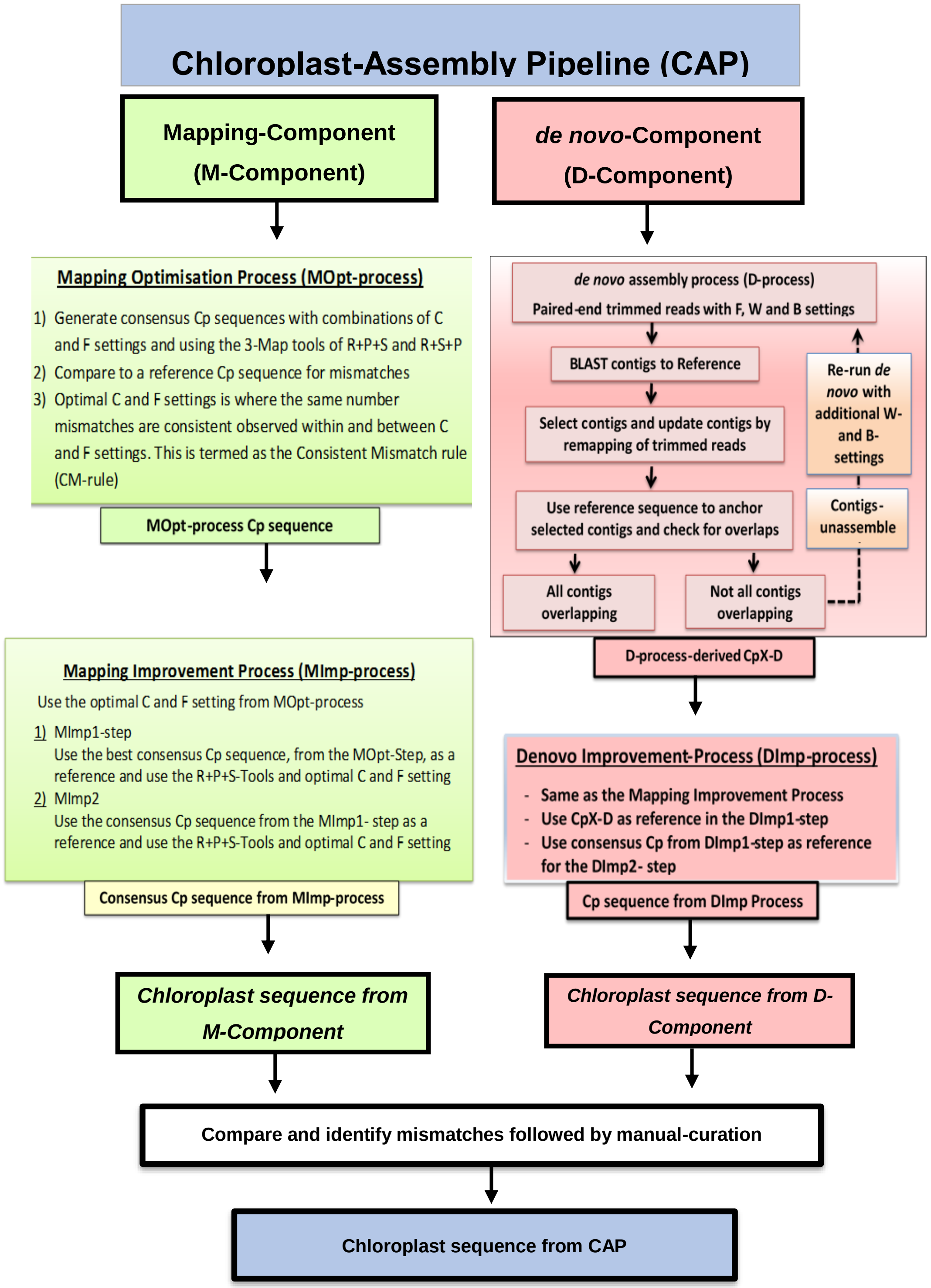
57053	57060	ATATCTA AAGTAT	Deletion	13		Yes	0	No, No contigs in this region leading to NNN. The error is due to presence of a tandem repeat sequence CTTTTTTTTTAGAATA and also a non-tandedem sequence GTATTCT. The sequence spanning the repeat, based on reads is TTCGATTCTTTTTTTTTAGAATACTTTTTTTTTAGAATACTAAAGTATTCTAAAAAAAAAGTATTCTA. The CpD has the following sequence instead CTTTTTTTTTAGAATACTTTTTTTTTAGANNNNTCTAAAAAAAAAGTATTCTA, with the following sequence missing due to the NNNN, ATACTAAAGTAT. So this 13 base sequence ATACTAAAGTAT should be added to the CpD sequence. 12 bases sequence to be insertrs minus 4 Ns = 8 bases	8
73152	73152	A	SNP (transversion)	1	T	Yes	0	No, Correct reads present but mismatch caused by high coverage of of spurious broken reads	0
78424	78424	C	SNP (transition)	1	T	Yes	0	No, Homopolymer region. Correct reads present, but mismatch caused by high coverage of mapping of spurious broken reads	0
78439	78442	TTAAT	Deletion	4					
78441	78442	AA	Substitution	2	TT	Yes	0	No, Part of a T-nucleotide homopolymer region. Mispmatch caused by high coverage of spurious single reads of T-nucleotides.	0
78444	78444	C	SNP (transition)	1	T	Yes	0	No, Part of a T-nucleotide homopolymer region. Mispmatch caused by high coverage of spurious single reads of T-nucleotides.	0

78446	78447	CC	Deletion	2	T	Yes	0	No, Part of a T-nucleotide homopolymer region. Mispmatch caused by high coverage of spurious single reads of T-nucleotides.	1
78455	78455	A	SNP (transversion)	1	T	Yes	0	No, Part of a T-nucleotide homopolymer region. Mispmatch caused by high coverage of spurious single reads of T-nucleotides.	0
100654	100655	AA	Deletion	2					
100655	100656	AA	Deletion	2		Yes	0	No, The deletion is part of GAAA. The specific contig and hence the CpD has the sequence GA. The denovo failed even though there are reads with sequence spanning on either side of the GAAA. The error could be because there are reads ending at G or at GA.	2
102132	102132	T	SNP (transition)	1	C	Yes	0	No, Presence of partly mapped single reads causing the transition error.	0
102134	102134	G	SNP (transversion)	1	T	Yes	0	No, Presence of partly mapped single reads causing the transversion error.	0
105791	105794	GCTT	Substitution	4	AAGC	No, one read matched the reference even though most of the reads had the AAGC sequence	0	Yes	0
TOTAL mismatches/variants						1		17	
Manual curation: change in length (bp)						0		81	
Manual-curation led final Cp consensus length in bp						Revised M-component Cp: 134,550 + 0 = 134,550 bp		Revised D-process Cp, CpN-D: 134,469 + 81 = 134,550 bp	

			Mismatches between the M-Component derived Cp sequence (CpW/CpN-C1F3-MImp2:R+P+S, 134,550 bases) and the D-process derived Cp sequence (CpN-D, 134,469 bases) and the DImp2-process derived Cp sequence (CpN-D-DImp1/CpN-C1F3-DImp2:R+P+S, 134,465 bases)					
						Outcome of manual curation for;			
Start bp	End bp	Sequence	Description of mismatches in: CpN-D or the DImp2-process derived Cp sequence			M-component derived Cp sequence (134,550 bp)		DImp2-process derived Cp sequence (134,465 bases)	
			Type	Length (with gaps)	Sequence	Correct Reason when not correct	Mismatch included (+) or deleted (-)	Correct Reason when not correct	Mismatch included (+) or deleted (-)
24687	24707	AATTGTC GAATTAT ACTCAGC	Deletion	21		Yes	0	No, The presence of a 8 base sequence TACTCAGC as a repeat is clearly present in the reads. The De novo assembly is erronous as one of this repeat region is represented which can be noticed when the reads overlap at the repeat region in updated contig file	21
43019	43019	G	SNP (transition)	1	A				
46096	46153	TACGAAA ACATAAT AAAGAG AACATGC GAATTTC TTGTATT TTCAGTC CATCATT ATA	Deletion	58		Yes	0	No, No coverage by contigs in this region. NNN were inserted which in the DMP process removed the Ns and this led to the the false deletion	58

46097	46146	TATTATATACGAAAACATAATAAAGAGAACATGCGAATTTCTTGTATTTTCAGTCCAT	Deletion	58					
55846	55846		Insertion (tandem repeat)	1	A				
55849	55849	T	SNP (transition)	1	C				
55853	55853	A	SNP (transition)	1	G				
55912	55912	T	SNP (transition)	1	C				
57053	57060	ATATCTAAAGTAT	Deletion	13					
73152	73152	A	SNP (transversion)	1	T				
78424	78424	C	SNP (transition)	1	T				
78439	78442	TTAAT	Deletion	4		Yes	0	No, Contigs mis-assembled due to T homopolymer. The error in the Contigs was not corrected by the DMP analysis.	4
78441	78442	AA	Substitution	2	TT				

78444	78444	C	SNP (transition)	1	T				
78446	78447	CC	Deletion	2	T				
78455	78455	A	SNP (transversion)	1	T				
100654	100655	AA	Deletion	2		Yes	0	No, The deletion is part of GAAA. The specific contig has the sequence GA. The denovo failed even though there are reads with sequence spanning on either side of the GAAA. The error could be because there are reads ending at G or at GA.	2
100655	100656	AA	Deletion	2					
102132	102132	T	SNP (transition)	1	C				
102134	102134	G	SNP (transversion)	1	T				
105791	105794	GCTT	Substitution	4	AAGC	No, one read matched the reference even though most of the reads had the AAGC sequence	0	Yes	0
TOTAL mismatches/variants						1		4	
Manual curation: change in length (bp)						0		85	
Manual-curation led final Cp consensus length in bp						Revised M-component Cp: 134,550 + 0 = 134,550 bp		Revised DImp2-process Cp: 134,465 + 85 = 134,550 bp	



Chloroplast-Assembly Pipeline (CAP)

Mapping-Component (M-Component)

de novo-Component (D-Component)

Mapping Optimisation Process (MOpt-process)

1) Generate consensus Cp sequences with combinations of C and F settings and using the 3-Map tools of R+P+S and R+S+P
2) Compare to a reference Cp sequence for mismatches
3) Optimal C and F settings is where the same number mismatches are consistent observed within and between C and F settings. This is termed as the Consistent Mismatch rule (CM-rule)

MOpt-process Cp sequence

de novo assembly process (D-process)

Paired-end trimmed reads with F, W and B settings

BLAST contigs to Reference

Select contigs and update contigs by remapping of trimmed reads

Use reference sequence to anchor selected contigs and check for overlaps

Re-run *de novo* with additional W- and B- settings

Contigs-unassemble

All contigs overlapping

Not all contigs overlapping

D-process-derived CpX-D

Mapping Improvement Process (MImp-process)

Use the optimal C and F setting from MOpt-process

1) MImp1-step
Use the best consensus Cp sequence, from the MOpt-Step, as a reference and use the R+P+S-Tools and optimal C and F setting
2) MImp2
Use the consensus Cp sequence from the MImp1- step as a reference and use the R+P+S-Tools and optimal C and F setting

Consensus Cp sequence from MImp-process

Denovo Improvement-Process (DImp-process)

- Same as the Mapping Improvement Process
- Use CpX-D as reference in the DImp1-step
- Use consensus Cp from DImp1-step as reference for the DImp2- step

Cp sequence from DImp Process

Chloroplast sequence from M-Component

Chloroplast sequence from D-Component

Compare and identify mismatches followed by manual-curation

Chloroplast sequence from CAP

Figure 26 Details of the Chloroplast Assembly Pipeline (CAP) R; read mapping tool, P; extract paired end mapped reads and remapping, S; structural variant plus realignment tools. **i,** The CAP pipeline consists of two distinct components the M-Component and the D-Component. **ii,** All mappings steps included Cost (C) settings of C1 and C2 comprising of 2, 3, 3 and 1, 2, 2 for Mismatch Cost, Insertion Cost and Deletion Cost respectively. Each of the Cost Settings had a combination of five Fraction (F) settings of 1.0, 1.0 and 1.0, 0.95 and 1.0, 0.8 and 0.8, 0.8 and 0.8, 0.5 for Length Fraction and Similarity Fraction respectively. **iii**, All *de novo* assembly steps were undertaken using the “Fast” (F) mode and at various “Word” (W) and “Bubble” (B) settings. The DImp-process involved subjecting the CpX-D to the 3-Map-Tool of R+P+S-Tool. X, code name of the genotype whose Cp is being generated. All analysis are designed to be undertaken in the CLC Genomics Workbench (CLCBio, Qiagen, Denmark).

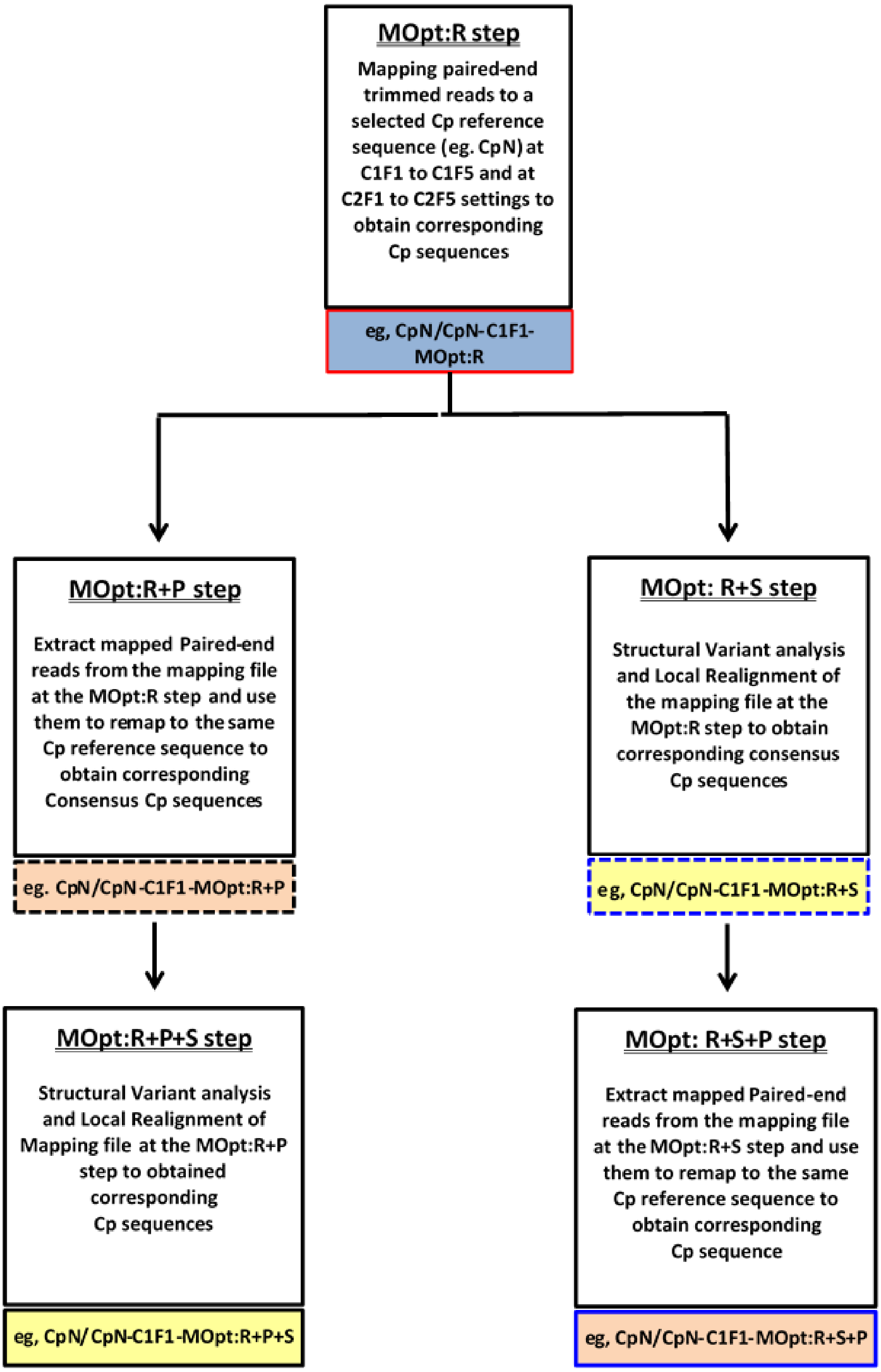

MOpt:R step
Mapping paired-end trimmed reads to a selected Cp reference sequence (eg. CpN) at C1F1 to C1F5 and at C2F1 to C2F5 settings to obtain corresponding Cp sequences
eg, CpN/CpN-C1F1-MOpt:R
MOpt:R+P step
Extract mapped Paired-end reads from the mapping file at the MOpt:R step and use them to remap to the same Cp reference sequence to obtain corresponding Consensus Cp sequences
eg. CpN/CpN-C1F1-MOpt:R+P
MOpt: R+S step
Structural Variant analysis and Local Realignment of the mapping file at the MOpt:R step to obtain corresponding consensus Cp sequences
eg, CpN/CpN-C1F1-MOpt:R+S
MOpt:R+P+S step
Structural Variant analysis and Local Realignment of Mapping file at the MOpt:R+P step to obtained corresponding Cp sequences
eg, CpN/CpN-C1F1-MOpt:R+P+S
MOpt: R+S+P step
Extract mapped Paired-end reads from the mapping file at the MOpt:R+S step and use them to remap to the same Cp reference sequence to obtain corresponding Cp sequence
eg, CpN/CpN-C1F1-MOpt:R+S+P

Figure 27 Steps of the Mapping Optimisation process and the “Cost” and “Fraction” mapping settings used to obtain an accurate mapping-derived chloroplast assembled sequence. N/, abbreviation of the reference Cp genome sequence used a suffix to denote the reference Cp sequence used; C, Cost mapping setting used; F, Fraction mapping settings used; MOpt, Mapping Optimisation process; R, read mapping tool; P, extracting of mapped paired-end reads and remapping tool: S, structural variant analysis plus local realignment tool. Mapping assembled Chloroplast sequences were progressively passed through various MOpt steps. Increasing C- and F-values represents decreasing stringency in mapping of sequence reads to the Cp reference used. All analyses were conducted using the CLC Genomics Workbench analysis software.

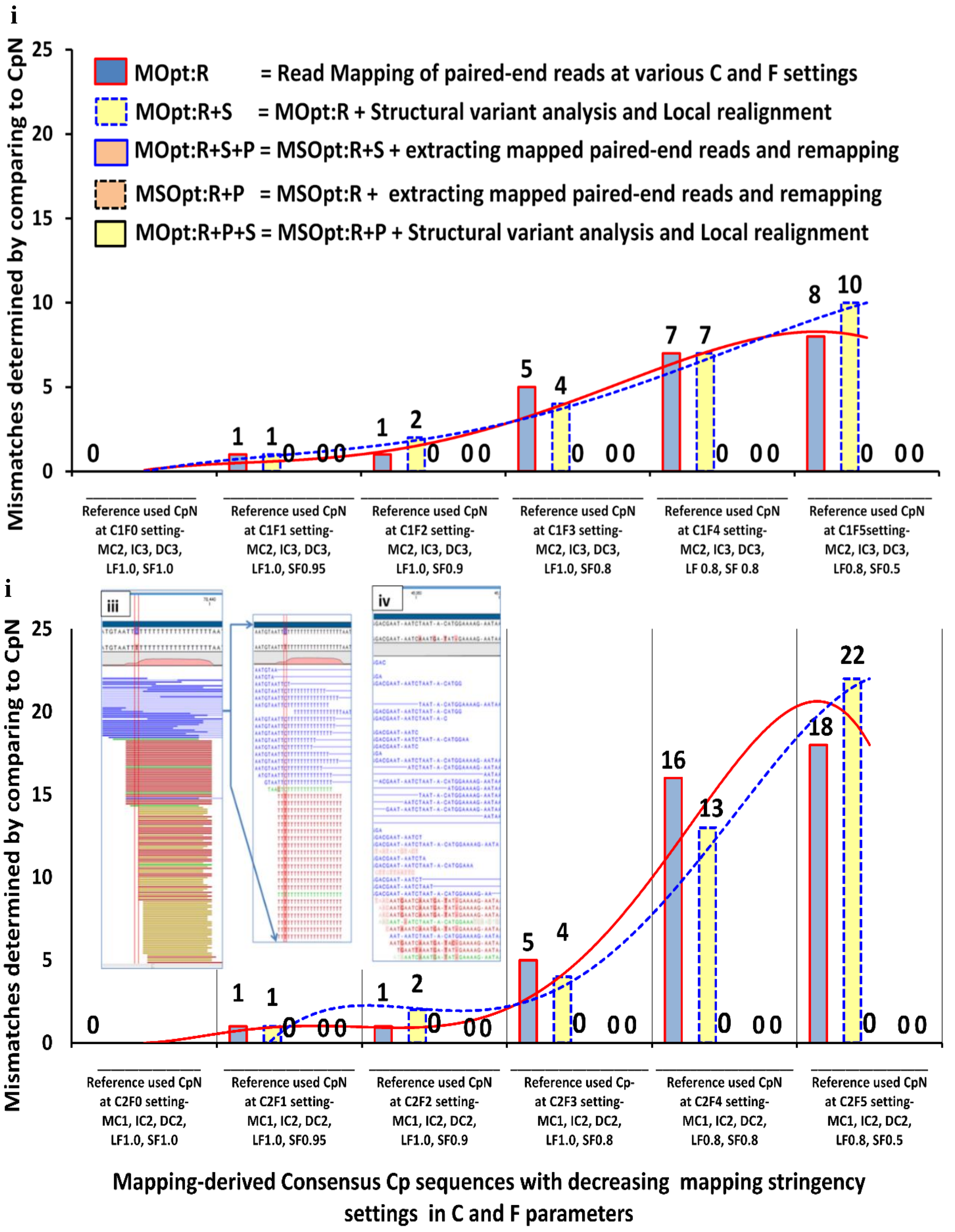


Figure 28 Mapping Optimisation (MOpt) process-derived chloroplast genome (Cp) sequences using CpW as a reference and mismatches when compared to CpN. CpW, Cp sequence of the Australian Wild rice Taxon-A (Genbank accession KF428978); CpN, Cp sequence of *Oryza sativa* Nipponbare (CpN, Genbank accession GU592207); LF, SF, length and similarity fraction; MC, IC, DC, mismatch, insertion and deletion cost. **i, ii, iii, iv**, data related to the MOpt process-derived Cp sequences derived using a fixed setting for MC, IC and DC of 2, 3, 3 and of 1, 2, 2 respectively and within these six combinations of LF and SF Fraction settings. Y-axis indicates mismatches in the MOpt process-derived Cp genome sequences when compared to the CpN (**i, ii**) and when compared to CpW (**iii, iv**). Read mapping was carried out using 35 bp paired-end Illumina reads of *O. sativa* Nipponbare (Genbank accession GU592207) and using the publically available CpW. MOpt process involves the read mapping tool (R), extracting the mapped paired-end reads and remapping tool (P) and the structural variant plus local realignment tool (S), implemented is sequence as the R+P+S or R+S+P with the aim of reducing the mismatches in the Cp sequences obtained from the preceding step. Number of mismatches when compared to CpN, shown above each bar, is a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions. Consistent number of mismatches in consensus Cp sequences derived from the R+P+S step and the R+S+P step, at each of the C and F setting used, are highlighted in blue. All mapping analysis was carried out using CLC Genomics Workbench V7.5.1.

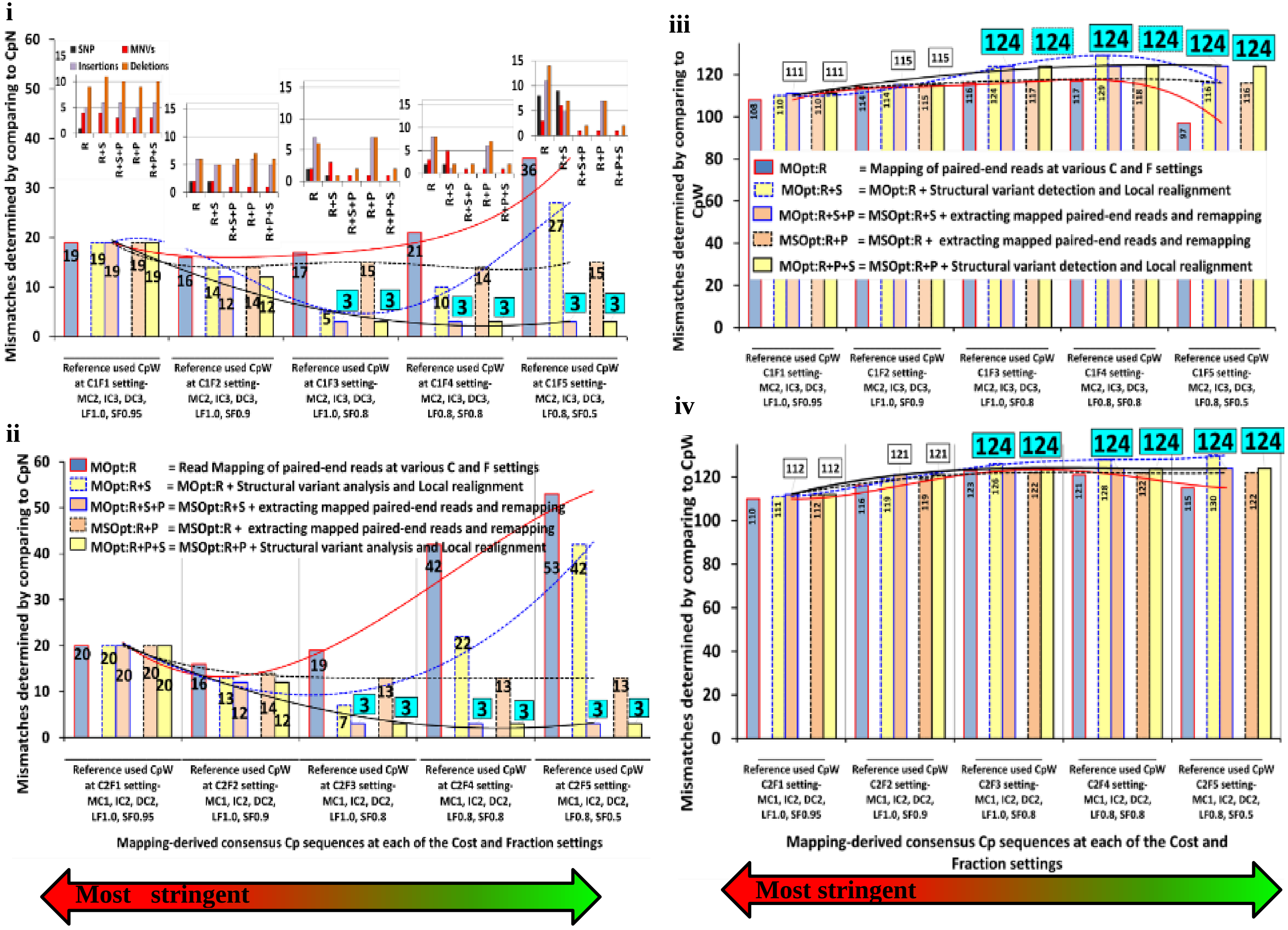

i
Mismatches determined by comparing to CpN
SNP
MNVs
Insertions
Deletions
R
R+S
R+S+P
R+P
R+P+S
Reference used CpW at C1F1 setting- MC2, IC3, DC3, LF1.0, SF0.95
Reference used CpW at C1F2 setting- MC2, IC3, DC3, LF1.0, SF0.9
Reference used CpW at C1F3 setting- MC2, IC3, DC3, LF1.0, SF0.8
Reference used CpW at C1F4 setting- MC2, IC3, DC3, LF0.8, SF0.8
Reference used CpW at C1F5 setting- MC2, IC3, DC3, LF0.8, SF0.5
ii
Mismatches determined by comparing to CpN
MOpt:R = Read Mapping of paired-end reads at various C and F settings
MOpt:R+S = MOpt:R + Structural variant analysis and Local realignment
MOpt:R+S+P = MSOpt:R+S + extracting mapped paired-end reads and remapping
MSOpt:R+P = MSOpt:R + extracting mapped paired-end reads and remapping
MOpt:R+P+S = MSOpt:R+P + Structural variant analysis and Local realignment
Reference used CpW at C2F1 setting- MC1, IC2, DC2, LF1.0, SF0.95
Reference used CpW at C2F2 setting- MC1, IC2, DC2, LF1.0, SF0.9
Reference used CpW at C2F3 setting- MC1, IC2, DC2, LF1.0, SF0.8
Reference used CpW at C2F4 setting- MC1, IC2, DC2, LF0.8, SF0.8
Reference used CpW at C2F5 setting- MC1, IC2, DC2, LF0.8, SF0.5
Mapping-derived consensus Cp sequences at each of the Cost and Fraction settings
Most stringent
iii
Mismatches determined by comparing to CpW
MOpt:R = Mapping of paired-end reads at various C and F settings
MOpt:R+S = MOpt:R + Structural variant detection and Local realignment
MOpt:R+S+P = MSOpt:R+S + extracting mapped paired-end reads and remapping
MSOpt:R+P = MSOpt:R + extracting mapped paired-end reads and remapping
MOpt:R+P+S = MSOpt:R+P + Structural variant detection and Local realignment
Reference used CpW C1F1 setting- MC2, IC3, DC3, LF1.0, SF0.95
Reference used CpW C1F2 setting- MC2, IC3, DC3, LF1.0, SF0.9
Reference used CpW C1F3 setting- MC2, IC3, DC3, LF1.0, SF0.8
Reference used CpW C1F4 setting- MC2, IC3, DC3, LF0.8, SF0.8
Reference used CpW C1F5 setting- MC2, IC3, DC3, LF0.8, SF0.5
iv
Mismatches determined by comparing to CpW
Reference used CpW C2F1 setting- MC1, IC2, DC2, LF1.0, SF0.95
Reference used CpW C2F2 setting- MC1, IC2, DC2, LF1.0, SF0.9
Reference used CpW C2F3 setting- MC1, IC2, DC2, LF1.0, SF0.8
Reference used CpW C2F4 setting- MC1, IC2, DC2, LF0.8, SF0.8
Reference used CpW C2F5 setting- MC1, IC2, DC2, LF0.8, SF0.5
Mapping-derived consensus Cp sequences at each of the Cost and Fraction settings
Most stringent

Figure 29 Mapping Optimisation (MOpt) process-derived chloroplast genome (Cp) sequences using CpW as a reference and mismatches when compared to CpN. CpW, Cp sequence of the Australian Wild rice Taxon-A (Genbank accession KF428978); CpN, Cp sequence of *Oryza sativa* Nipponbare (CpN, Genbank accession GU592207); LF, SF, length and similarity fraction; MC, IC, DC, mismatch, insertion and deletion cost. i, ii, iii, iv, data related to the MOpt process-derived Cp sequences derived using a fixed setting for MC, IC and DC of 2, 3, 3 and of 1, 2, 2 respectively and within these six combinations of LF and SF Fraction settings. Y-axis indicates mismatches in the MOpt process-derived Cp genome sequences when compared to the CpN (i, ii) and when compared to CpW (iii, iv). Read mapping was carried out using 35 bp paired-end Illumina reads of *O. sativa* Nipponbare (Genbank accession GU592207) and using the publically available CpW. MOpt process involves the read mapping tool (R), extracting the mapped paired-end reads and remapping tool (P) and the structural variant plus local realignment tool (S), implemented is sequence as the R+P+S or R+S+P with the aim of reducing the mismatches in the Cp sequences obtained from the preceding step. Number of mismatches when compared to CpN, shown above each bar, is a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions. Consistent number of mismatches in consensus Cp sequences derived from the R+P+S step and the R+S+P step, at each of the C and F setting used, are highlighted in blue. All mapping analysis was carried out using CLC Genomics Workbench V7.5.1.

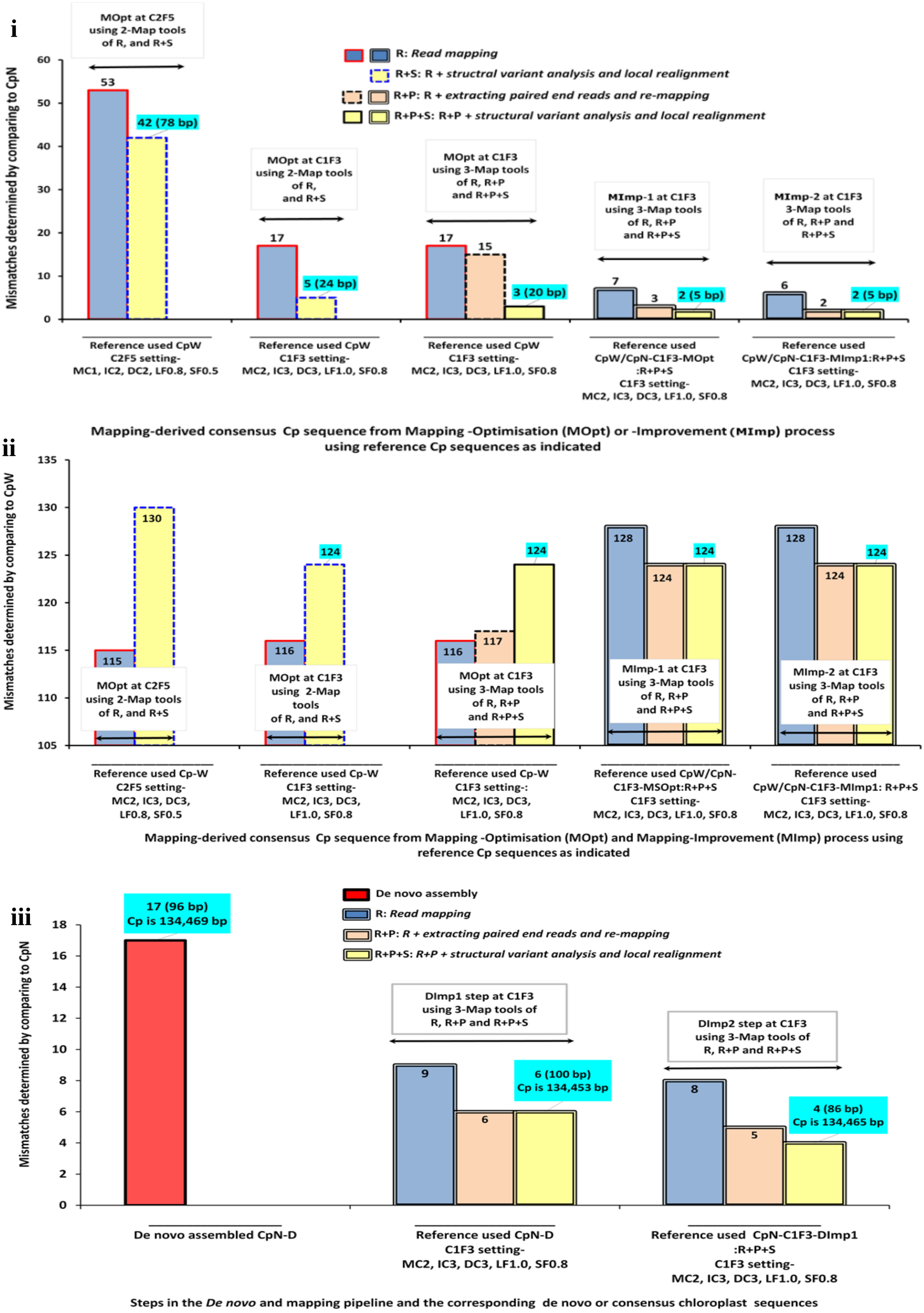

i
MOpt at C2F5
using 2-Map tools
of R, and R+S
R: Read mapping
R+S: R + structral variant analysis and local realignment
R+P: R + extracting paired end reads and re-mapping
R+P+S: R+P + structural variant analysis and local realignment
Mismatches determined by comparing to CpN
60
50
40
30
20
10
0
53
42 (78 bp)
MOpt at C1F3
using 2-Map tools
of R,
and R+S
17
5 (24 bp)
MOpt at C1F3
using 3-Map tools
of R, R+P
and R+P+S
17
15
3 (20 bp)
MImp-1 at C1F3
using 3-Map tools
of R, R+P
and R+P+S
7
3
2 (5 bp)
MImp-2 at C1F3
3-Map tools
of R, R+P and
R+P+S
6
2
2 (5 bp)
Reference used CpW
C2F5 setting-
MC1, IC2, DC2, LF0.8, SF0.5
Reference used CpW
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Reference used CpW
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Reference used
CpW/CpN-C1F3-MOpt
:R+P+S
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Reference used
CpW/CpN-C1F3-MImp1:R+P+S
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Mapping-derived consensus Cp sequence from Mapping -Optimisation (MOpt) or -Improvement (MImp) process
using reference Cp sequences as indicated
ii
Mismatches determined by comparing to CpW
135
130
125
120
115
110
105
115
130
MOpt at C2F5
using 2-Map tools
of R, and R+S
116
124
MOpt at C1F3
using 2-Map
tools
of R, and R+S
116
117
124
MOpt at C1F3
using 3-Map tools
of R, R+P
and R+P+S
128
124
124
MImp-1 at C1F3
using 3-Map tools
of R, R+P
and R+P+S
128
124
124
MImp-2 at C1F3
using 3-Map tools
of R, R+P
and R+P+S
Reference used Cp-W
C2F5 setting-
MC2, IC3, DC3,
LF0.8, SF0.5
Reference used Cp-W
C1F3 setting-
MC2, IC3, DC3,
LF1.0, SF0.8
Reference used Cp-W
C1F3 setting-:
MC2, IC3, DC3,
LF1.0, SF0.8
Reference used CpW/CpN-
C1F3-MSOpt:R+P+S
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Reference used
CpW/CpN-C1F3-MImp1: R+P+S
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Mapping-derived consensus Cp sequence from Mapping -Optimisation (MOpt) and Mapping-Improvement (MImp) process using
reference Cp sequences as indicated
iii
De novo assembly
R: Read mapping
R+P: R + extracting paired end reads and re-mapping
R+P+S: R+P + structural variant analysis and local realignment
Mismatches determined by comparing to CpN
18
16
14
12
10
8
6
4
2
0
17 (96 bp)
Cp is 134,469 bp
DImp1 step at C1F3
using 3-Map tools of
R, R+P and R+P+S
9
6
6 (100 bp)
Cp is 134,453 bp
DImp2 step at C1F3
using 3-Map tools of
R, R+P and R+P+S
8
5
4 (86 bp)
Cp is 134,465 bp
De novo assembled CpN-D
Reference used CpN-D
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Reference used CpN-C1F3-DImp1
:R+P+S
C1F3 setting-
MC2, IC3, DC3, LF1.0, SF0.8
Steps in the De novo and mapping pipeline and the corresponding de novo or consensus chloroplast sequences

Figure 30 Mapping Improvement (MImp) and De novo Improvement (DImp) process reduces mismatches in the Cp sequence from MOpt-process and CpN-D sequence respectively. Cp; chloroplast sequence; MOpt, Mapping optimisation process; CpN-D, *de novo* assembly-derived Cp sequence; MC, IC and DC, mismatch, insertion and deletion cost (C) setting; LF and SF, length and similarity fraction (F) setting ; R, read mapping tool; P, extracting mapped paired-end reads and remapping tool; S, structural variant analysis and local realignment tool. Read mapping and *de novo* assembly was carried out using 35bp Illumina Paired end reads of *Oryza sativa* cv Nipponbare. The Cp sequence of the Australian Wild rice Taxon-A (CpW, KF428978) was used as a reference for read mapping assembly at C1F3 mapping settings representing a C setting of 2, 3, 3 for MC, IC and DC respectively, and a F setting of 1.0 and 0.8 for LF and SF respectively. The X-axis indicates the various Cp sequences and the mapping settings used. Mismatches in mapping-derived Chloroplast (Cp) sequences when compared to the publically available Cp sequence of *O. sativa* Nipponbare (CpN, GU592207) (**i, iii**) and to the Australian Wild rice Taxa-A (CpW, KF428978) (**ii**). Number of mismatches are a sum of single nucleotide variants, multi-nucleotide variants, insertions and deletions and are shown at top of each bar while those in blue highlight represent mismatches and bases covered. **i**; The Cp sequence from the MOpt process at the C2F5 had the highest mismatches of 42 over 78 bases, while at the C1F3 setting had 5 over 24 bases to 3 mismatches over 5 bases when using the 3-Map tools and further reduced at the MImp process to 2 over 5 bases. **ii**; All of the consensus Cp sequences discussed above had 124 mismatches when compared to CpW. iii, The *de novo* assembly derived CpN-D sequence was also improved when passed through the DImp process with 17 mismatches over 96 bases reduced to 4 over 86 bases.

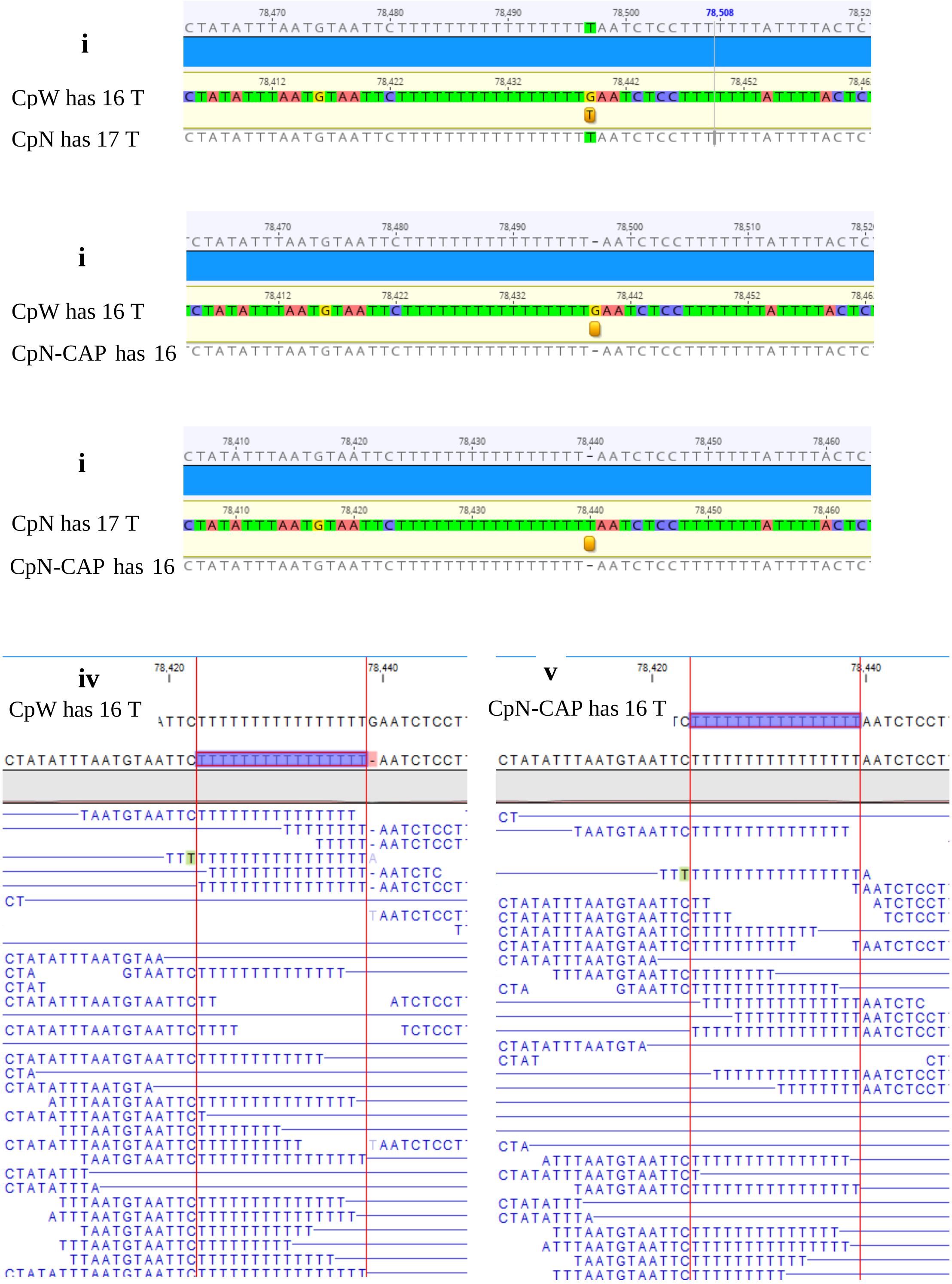

i
CpW has 16 T
CpN has 17 T
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTTTAATCTCCTTTTTTTATTTTACTC
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTTTGAATCTCCTTTTTTTATTTTACTC
i
CpW has 16 T
CpN-CAP has 16
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTTT-AATCTCCTTTTTTTATTTTACTC
i
CpN has 17 T
CpN-CAP has 16
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTTT-AATCTCCTTTTTTTATTTTACTC
iv
CpW has 16 T
ATTCTTTTTTTTTTTTTTTTGAATCTCCT
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTT-AATCTCCT
v
CpN-CAP has 16 T
TCTTTTTTTTTTTTTTTTAATCTCCT
CTATATTTAATGTAATTCTTTTTTTTTTTTTTTTAATCTCCT

Figure 31 Comparisons of Chloroplast (Cp) sequences at a T nucleotide homopolymer sequence between CpN-CAP sequence of *Oryza* sativa cv Nipponbare to CpN and CpW respectively. CpN-CAP, chloroplast sequence assembled using the Chloroplast assembly pipeline (CAP) using paired end Illumina reads (35 bp) of *Oryza sativa*; CpW, chloroplast sequences of Australian Wild rice Taxa-A (KF428978); CpN, Chloroplast sequence of *O. sativa* (GU592207). CpW has 16 T and one of the T is replaced with a G nucleotide while CpN has 17 T nucleotide (i). The assembled CpN-CAP has 16 T nucleotides (B) and its alignment to CpW is shown in (ii) and to CpN is shown in (iii). The mapping of the paired end Illumina reads (35 bp) to CpW is shown in (iv) and to CpN-CAP is shown in (v). All alignments were undertaken Geneious V 9 and all mapping of reads to Cp sequences using CLC genomics Work Bench V 9.0.

4 Appendix 4.

Table 27 Details of collections of wild rice from north Queensland made in 2015, 2016 and 2017. Including site description, GPS coordinates, panicle shape, awn and anther length for wild populations from each collection site.

Site #	Sample #	Site description	GPS location and elevation	Likely Species*	Panicles	Awn length** (mm)	Awn SD±	Anther length (mm)***	Anthers SD ±
1	WR-8	Mareeba Wetlands (Clancy Lagoon)	S:16.92661° E:145.35620° Elevation: 410 m	Taxon A	Open	4.6	1.3	4.60	0.21
2	WR-20B	Mareeba Wetlands (Pandanus lake)	S:16.93795° E:145.35077° Elevation: 422 m	Taxon B	Closed	9.5	1.3	2.09	0.13
3	WR-24B	Abbatoir Swamp (Mossman-Mt Molloy Road)	S:16.63574° E:145.32603° Elevation: 422 m	Taxon A	Open	5	2.1	-	-
4	WR-31	small roadside swamp, cnr Bethel Road and Mulligan Hwy	S:16.57874° E:145.18906° Elevation: 363 m	Taxon B (classic) or *O. meridionals*	Closed	10	1.8	2.28	0.07

5	WR-44,	Lakeland-Cook Town section, Mulligan Hwy	S:15.758640° E:144.99924° Elevation: 159 m	Mixed Taxon A and B	Open	6.9	2.1	4.45 3.82 3.5	0.18 (2015 collection) 0.10 (2017collection) large lake 0.17 (2017collection) small lake
	WR-52			Taxon B	Closed	9	1.3	4.26 2.81	0.23 (2017collection) large lake 0.14 (2017collection) small lake
	WR-65			Taxon B+ *O. australiensis*	Partially open	11.8	1.1		-
6	WR-74	Barretts Road, near Cook Town Airport. Wetland/Swamp	S:15.43399° E:145.17816° Elevation: 25 m	*O. meridionalis*, Taxon B, Taxon B+	Closed	10.1	2.2	2.34	0.21
7	WR-83	Unnamed marshland/wetland	S:15.53078° E:144.38336° Elevation: 95 m	*O. meridionalis*	Closed	8	2.1	2.08	0.16
8	WR-91	Lakefield National Park	S:15.20969° E:144.38966° Elevation: 58 m	*O. meridionalis*, Taxon B, Taxon B+	Closed	11.6	2.6		-
9	WR-103	Lakefield National Park	S:14.85996° E:144.16586° Elevation: 32 m	*O. meridionalis,* Taxon B, Taxon B+	Closed	8.9	1.3		-

10	WR-111	Jpn11 site (Sotowa et al., 2013)	S:14.84947° E:144.16811° Elevation: 21 m	*O. meridionalis*, Taxon B, Taxon B+	Closed	9.1	0.9		-
11	WR-121	Lakefield National Park	S;15.14672° E:144.32773° Elevation: 57 m	*O. meridionalis* , Taxon B, Taxon B+	Closed	7.8	1.1		-
12	WR-133	Jpn2 site (Sotowa et al., 2013)	S:15.43943° E:144.21111° Elevation: 148 m	*O. meridionalis*, Taxon B, Taxon B+	Closed	7.5	1.6	2.05	0.09
13	WR-141B	Balurga Road (off Musgrave to Pormpurraw road)	S:14.83915° E:142.56808° Elevation: 88 m	Taxon B, *O. meridionalis*	Closed	14.7	2.2	1.94	0.21
14	WR-153	Balurga Road (off Musgrave to Pormpurraw road)	S:14.90241° E:142.49919° Elevation: 75 m	Taxon B, *O. meridionalis*	Closed	9.5	1.7	2.41	0.08
15	WR-162	Merluna	S: 13.05811° E:142.61964° Elevation: 137 m	*O. meridionalis* , Taxon B, Taxon B+	Closed	8.1	1.7		-
16	WR-172	Andoom Road, Weipa	S:12.61513° E:141.89191° Elevation : 8 m	*O. meridionalis*, Taxon B, Taxon B+	Closed	9.9	1.5	2.21	0.19
17	WR-182	Lydia Creek, Batavia Downs Road	S:12.66010° E:142.66843° Elevation: 68	Taxon B, *O. meridionalis*	Closed	10.9	2.6	2.26	0.18

18	WR-193	Development road to Bamaga, Moreton.	S:12.45885° E:142.63562° Elevation: 39	Taxon, *O. meridionalis*	Closed	9.7	1.9	2.68	0.14
19	WR-207	Telegraph Road (Weipa turnoff to Batavia Downs).	S:12.88274° E:142.73929° Elevation: 93	*O. meridionalis* Taxon B, Taxon B+	Closed	7.4	1.7	1.66	0.15
20	WR-213	Peninsular Development Road (Between Archer River Road to Weipa turnoff)	S:13.29167° E:142.84729° Elevation: 148	*O. meridionalis*, Taxon B, Taxon B+	Closed	11.2	1.6	1.54	1.03
21	WR-221	Peninsula Development Road (Between Coen and Musgrave)	S:14.005117° E:143.1903607° Elevation: 208	*O. meridionalis*, Taxon B, Taxon B+	Closed	8.6	1.7	2.19	0.18
22	WR-231	Peninsula Development Road (Between Musgrave to Laura)	S:14.785617° E:143.504467° Elevation: 76	*O. meridionalis*, Taxon A+ or *O. officinalis*	Open /completely open	9	1.7	2.51	0.11
23	WR-242	Peninsula Development Road	S:15.00745° E:143.640993° Elevation: 59	Taxon, *O. meridionalis*	Closed	6.6	1.1	1.69	0.14

24	WR-260	Townsville Site-1, Bruce Highway 30 km south of Townsville	S:19.395962 E:147.004486	Taxon B	Closed	9.3	1.4	2.03	0.12
25	WR-261	Townsville Site-2, Woodstock-Giru Road	S:19.599657 E:146.882965	Taxon A, ?	Open / Closed	5.9	1.3	1.83	0.07
26	WR-271	Townsville Site-3, Charters Towers-Townsville road	S:19.397224 E:146.723831	Taxon A, ?	Open / Closed	6.0	2.7	1.90	0.08
27	WR-285	Townsville Site-4, Town Common Wetlands, Townsville	S:19.25445 E:146.725586	Taxon B, *O. meridionalis*	Closed	10.5	1.7	3.60	0.17

*Designation in field: Taxon A *Oryza rufipogon*-like (open panicles), Taxon B *O. meridionalis* (closed panicles and short anthers) and Taxon B+ different to both Taxon A and B.

**Awn length average in cm for 10 seeds from 10 different plants from the population sampled randomly. Not representing the sequenced sample

***this is the average of ten anthers from the same plant. Not representing the sequenced sample

±standard deviation

‡ this site contains three different taxa

Table 28 Details of sequence coverage of Australian wild rice samples. Including whole genome coverage with total number of reads, and minimum, maximum and mean coverage of the chloroplast genome.

	Sample number	Site number	Whole genome		Chloroplast genome		
			Sequencing coverage	Total reads	Minimum coverage	Maximum coverage	Mean coverage
1	WR-8	1	7.33	16,581,166	10	649	388.07
2	WR-20B	2	9.21	20,821,128	16	620	364.43
3	WR-24B	3	10.2	23,069,168	24	1008	646.98
4	WR-31	4	9.35	21,140,048	17	659	446.06
5	WR-44	5	8.11	18,332,596	17	503	310.81
6	WR-52	5	8.61	19,462,696	15	610	370.77
7	WR-65	5	8.79	19,873,876	23	1082	718.52
8	WR-74	6	9.84	22,243,622	21	863	579.19
9	WR-83	7	15.42	34,862,816	34	1054	685.87
10	WR-91	8	13.3	30,070,336	35	1337	922.73
11	WR-103	9	12.24	27,683,838	47	1507	1088.64
12	WR-111	10	13.62	30,802,742	41	1314	961.82
13	WR-121	11	11.22	25,377,232	35	1011	686.18
14	WR-133	12	7.74	17,509,322	19	608	408
15	WR-141B	13	14.48	32,739,902	24	1128	700.56
16	WR-153	14	10.88	24,591,888	33	1330	906.44
17	WR-162	15	5.63	12,732,082	20	587	400.76
18	WR-172	16	6.9	15,604,400	12	476	278.41
19	WR-182	17	8.42	19,030,168	47	1268	901.65
20	WR-193	18	13.22	29,898,648	46	1541	1104
21	WR-207	19	8.71	19,686,052	21	821	577.54
22	WR-213	20	8.42	19,029,062	33	914	575.42
23	WR-221	21	10.37	23,450,552	56	1283	930.66
24	WR-231	22	6.29	14,225,150	22	646	412.28
25	WR-242	23	11.42	25,826,240	56	2063	1444.48
26	WR-260	24	3.95	8,934,498	15	478	320.6
27	WR-261	25	11.47	25,936,014	50	1341	932.98
28	WR-271	26	13.89	31,419,206	58	1518	1015.09
29	WR-285	27	9.55	21,591,668	48	1085	738.28

Table 29 Variants in chloroplast genomes insertions, deletions and SNPs compared with the O. sativa subsp. *japonica* Nipponbare GU592207.1 reference genome. Abbreviations are as follows: Del: deletion, Del.T.R.: deletion tandem repeat, Ins.: insertion, Ins.T.R.: insertion tandem repeat, SNP Tr.: SNP transition, SNP Trv.:SNP transversion and Subs.: substitution.

Sample number	Deletion	Deletion tandem repeat	Insertion	Insertion tandem repeat	SNP trans-etion	SNP trans-version	Sub-stitution	Total
WR-8	12	7	5	11	48	41	4	128
WR-20B	11	8	6	10	49	43	3	130
WR-24B	12	7	4	11	47	42	4	127
WR-31	11	7	6	11	50	44	4	133
WR-44	12	7	4	12	48	41	4	128
WR-52	12	7	4	12	48	41	4	128
WR-65	12	8	6	10	49	42	3	130
WR-74	12	7	6	10	50	43	4	132
WR-83	12	7	6	11	49	43	3	131
WR-91	12	8	6	10	49	43	2	130
WR-103	12	7	6	10	50	43	4	132
WR-111	12	7	6	10	49	42	3	129
WR-121	12	7	6	10	49	43	3	130
WR-133	11	8	6	11	49	43	3	131
WR-141B	13	7	6	10	50	43	3	132
WR-153	12	7	4	11	46	39	3	122
WR-162	13	8	5	11	48	39	5	129
WR-172	11	7	6	11	49	40	4	128
WR-182	12	7	6	10	49	42	4	130
WR-193	12	7	6	10	49	42	4	130
WR-207	13	7	6	12	49	42	2	131
WR-213	12	6	6	10	44	40	5	123
WR-221	12	7	6	10	49	42	3	129
WR-231	12	7	6	10	49	42	4	130
WR-242	12	7	6	10	50	43	3	131
WR-260	12	7	6	10	49	43	4	131
WR-261	12	7	6	11	49	43	4	132
WR-271	12	7	6	11	49	43	4	132
WR-285	12	7	6	10	49	42	3	129

Table 30 Chloroplast functional nucleotide polymorphisms (FNPs) in Australian wild rice populations. including position, gene name, gene product, amino acid substitution and codon change.

	Site	Gene	Gene product	Protein ID	Amino acid change	CDS	CDS codon number	CDS position	CDS position within codon	Change	Codon change	Polymorphism type	Protein effect
1	8,593		hypothetical protein	NP_039365.1	G -> E	hypothetical protein CDS	82	245	2	G -> A	GGA -> GAA	SNP (transition)	Substitution
2	8,599		hypothetical protein	NP_039365.1	G -> E	hypothetical protein CDS	84	251	2	G -> A	GGG -> GAG	SNP (transition)	Substitution
3	8,622		hypothetical protein	NP_039365.1	S -> P	hypothetical protein CDS	92	274	1	T -> C	TCC -> CCC	SNP (transition)	Substitution
4	24,178	rpoC1	RNA polymerase beta' subunit	NP_039374.1	N -> S	rpoC1 CDS	567	1,700	2	A -> G	AAT -> AGT	SNP (transition)	Substitution
5	24,756	rpoC2	RNA polymerase beta" subunit	NP_039375.1	Q ->H	rpoC2 CDS	10	30	3	G -> T	CAG -> CAT	SNP (transversion)	Substitution
6	25,897	rpoC2	RNA polymerase beta" subunit	NP_039375.1	H ->D	rpoC2 CDS	391	1,171	1	C -> G	CAT -> GAT	SNP (transversion)	Substitution
7	27,695	rpoC2	RNA polymerase beta" subunit	NP_039375.1	G ->D	rpoC2 CDS	990	2,969	2	G -> A	GGT -> GAT	SNP (transition)	Substitution
8	28,019	rpoC2	RNA polymerase beta" subunit	NP_039375.1	W ->L	rpoC2 CDS	1,098	3,293	2	G -> T	TGG -> TTG	SNP (transversion)	Substitution
9	29,113	rpoC2	RNA polymerase beta" subunit	NP_039375.1	N ->D	rpoC2 CDS	1,463	4,387	1	A -> G	AAC -> GAC	SNP (transition)	Substitution

10	29,138	rpoC2	RNA polymerase beta" subunit	NP_039375.1	Q -> P	rpoC2 CDS	1,471	4,412	2	A -> C	CAA -> CCA	SNP (transversion)	Substitution
11	30,699	atpI	ATP synthase CF0 A subunit	NP_039377.1	D ->N	atpI CDS	67	199	1	G -> A	GAT -> AAT	SNP (transition)	Substitution
12	40,251	psaA	photosystem I P700 chlorophyll a apoprotein A1	NP_039383.1	R -> G	psaA CDS	334	1,000	2	G -> C	CGC -> CCC	SNP (transversion)	Substitution
13	56,665		acetyl-CoA carboxylase beta subunit	NP_039394.1	S -> Y	acetyl-CoA carboxylase beta subunit CDS	38	113	2	C -> A	TCT -> TAT	SNP (transversion)	Substitution
14	66,104	rps18	ribosomal protein S18	NP_039408.1	T -> N	rps18 CDS	155	464	2	C -> A	ACC -> AAC	SNP (transversion)	Substitution
15	70,278	psbB	photosystem II 47 kDa protein	NP_039411.1	A -> T	psbB CDS	494	1,480	1	G -> A	GCA -> ACA	SNP (transition)	Substitution
16	70,281	psbB	photosystem II 47 kDa protein	NP_039411.1	I -> F	psbB CDS	495	1,483	1	A -> T	ATC -> TTC	SNP (transversion)	Substitution
17	105,906	ccsA	cytochrome c biogenesis protein	NP_039443.1	Y -> S	ccsA CDS	224	671	2	A -> C	TAT -> TCT	SNP (transversion)	Substitution
18	124,775		hypothetical protein	NP_039456.1	M ->L	hypothetical protein CDS	34	100	1	A -> C	ATG -> CTG	SNP (transversion)	Substitution

Table 31 Comparison of the SNPs, FNPs and the unique FNPs in Australian wild rice populations.

Accession	SNP	FNP	FNPs %	Common FNPs	Unique FNPs	Unique FNPs %
WR-8	93	11	11.83	6	5	46
WR-20B	95	12	12.63	6	6	50
WR-24B	93	11	11.83	6	5	46
WR-31	98	12	12.24	6	6	50
WR-44	93	11	11.83	6	5	46
WR-52	93	11	11.83	6	5	46
WR-65	94	12	12.77	6	6	50
WR-74	97	12	12.37	6	6	50
WR-83	95	12	12.63	6	6	50
WR-91	94	12	12.77	6	6	50
WR-103	97	12	12.37	6	6	50
WR-111	94	12	12.77	6	6	50
WR-121	95	12	12.63	6	6	50
WR-133	95	12	12.63	6	6	50
WR-141B	96	12	12.5	6	6	50
WR-153	88	10	11.36	6	4	40
WR-162	92	10	10.87	6	4	40
WR-172	93	14	15.05	6	8	57
WR-182	95	12	12.63	6	6	50
WR-193	95	12	12.63	6	6	50
WR-207	93	12	12.9	6	6	50
WR-213	89	10	11.24	6	4	40
WR-221	94	12	12.77	6	6	50
WR-231	95	12	12.63	6	6	50
WR-242	96	13	13.54	6	7	54
WR-260	96	12	12.5	6	6	50
WR-261	96	12	12.5	6	6	50
WR-271	96	12	12.5	6	6	50
WR-285	94	12	12.77	6	6	50

Table 32 Phylogenetic analysis tools applied to chloroplast genome analysis.

	Program	Analysing method	Substitution model	Rate variation	Bootstrapping	Out group
1	PAUP	Maximum Parsimony		Gamma	1000	*O. officinalis*
2	PHYLM	Maximum likelihood	GTR	Gamma	1000	-
3	MrBayes	Bayesian	GTR	Gamma	2000	*O. officinalis*

We compared methods and found that GTR was the best method for comparing diverse *Oryza* genomes (Brozynska et al., 2014a; Brozynska et al., 2014b) giving results consistent with known relationships at different genetic distances.

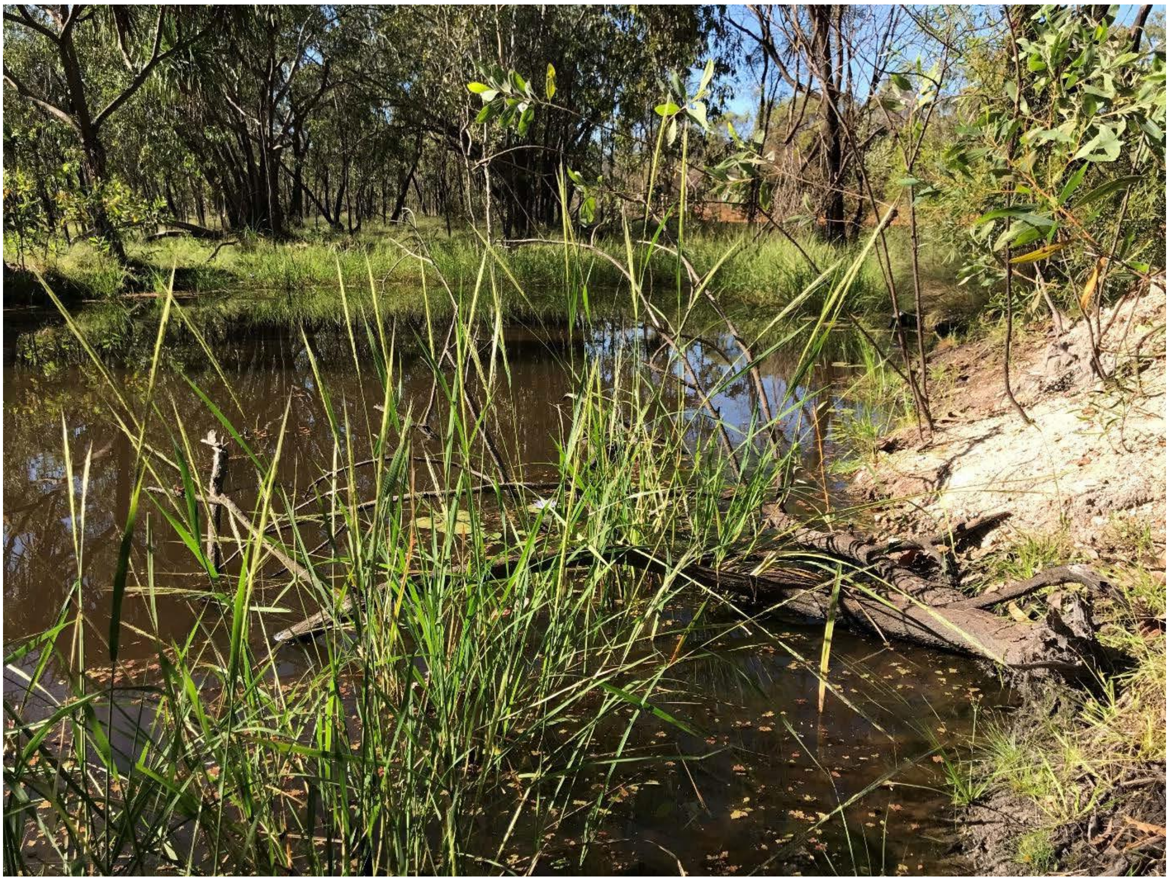


Figure 32 Wild rice habitat in northern Queensland Jpn2 site S:15.43943° E:144.21111°

Table 33 Unique chloroplast SNPs found in the Australian taxa.

	Sequence	SNPs
1	CACTAATAGGTTTCATGTTACGTCAATTTGAACTTGCTCGGTCTGTTCAATTGCGA/GCCTTATAATGCAATTTCATTCTCTGGCCCAATCGCTGTTTTTGTTTCCGTATTCCTGATTT	A Australian new taxa clade
2	GTCTTTCTGGTAGCTATTCTAAATTCTCTCATTTCTTAAATGTGTTTAGTAG/TTTAGTAGCCCGC/ATACAAAATAAAAAAGGGCCGTTTATTCGGATTGTGAGACGCATTAAAATGCAATTTGCG	G,C Australian new taxa clade
3	GCGAAGCAGGGGGGTGTAAATTGCAAAAAAGAAATTGGACTCTTTTTCCTATTAGATCAC/ATCAAATCACTACCCGTACTGAACTAATATAGAATCCCTTTTATTAATCTATTCTTATTCCATATCCTTT	C Australian new taxa clade
4	GTATTAACGATTGGAAACCGTCGAGGTATTTGTGCAAATAGATATAATAGTTGCGGAAACTATCCAAACCAAAAAGTAAG/ATTACAATAATAATAATCCTAAGTATACGAAAGATAAAGAATCTCTTTTTTCTAGTTCCTATGATGCACTGGGAGCTTATAGACAGAAACAAAT	G Australian new taxa clade
5	CCCGCAACCCCACGGTTATGAGCCTTGTCAGCTACCAAACTGTTCTATCCTGTTAAACTAAAGAGAGGGGAACTAGTGGATAAAAA/GGGGGGTTGAATACGCCCCTCTACCATATCTATACAAATAGAATAGTCCATTTATACAGAATGGTAAAGAGGGCTCTTCTACGATCATCAATTCCAGAAATCCAT	A Australian new taxa clade
6	AAGATTTCTCAATTTTCATTAAATCTTATAGAAAGAGGTAGAATTTCTTCTTTTTTTCAGGGATTTTAGGGAAAC/ATAAGGCTCTTGTCATTTTTTATTCTATTACTGAACAGAATGGGAAGACAGGGTTGGTTATTCTTCGTCTACGAATATCCAAATTTTAAC	C Australian new taxa clade
7	TTCGTAAAAATCTTTGGAAGAAAAAGACTTATTTTTCCATAGTACAATCTTATTCTTTAGCAAAATCAAGATCATTTTCTGGCGTCAGCGAGCAC/TCCAAAACCAAAGGGTTTTTCTCGGCAACAAACAAACAAATAATAGGGTTTTGGGATAATATGAATTGACCTATCCCCAAAAAATTCCAATTATTTAATATGAATAATTAG	C Australian new taxa clade
8	TCTTTTTGCCATTGGACTTTCCAATCGAATTGATTGTAAGACTCGTAAAGATCAACTTTACGAAGATCCCATTGTATTCCAGAAGCTCGTAACATGGGA/GCCCGATAAGCCCCAATTTACAGCTTCTTCTCCGCTAATAAAACCAACTCCCTCAACTCGTTCCAAAAAAATGGGATTCTGTGTAATAAGTTGTTGATATTCAA	A Australian new taxa clade

Table 34 Chromosomes phylogenetic analysis topology agreement.

Chromosome	Maximum Likelihood (ML) Vs Maximum Parsimony (MP)	Maximum Likelihood (ML) Vs Bayesian Inference (BI)	Maximum Parsimony (MP) Vs Bayesian Inference (BI)	Agreement among approaches
1	100%	-	-	-
2	100%	100%	100%	100%
3	100%	100%	100%	100%
4	100%	100%	100%	100%
5	90%	100%	100%	97%
6	95%	100%	100%	98%
7	95%	95%	100%	97
8	100%	100%	100%	100%
9	90%	-	100%	-
10	-	-	-	-
11	90%	100%	100%	97%
12	100%	90%	-	-

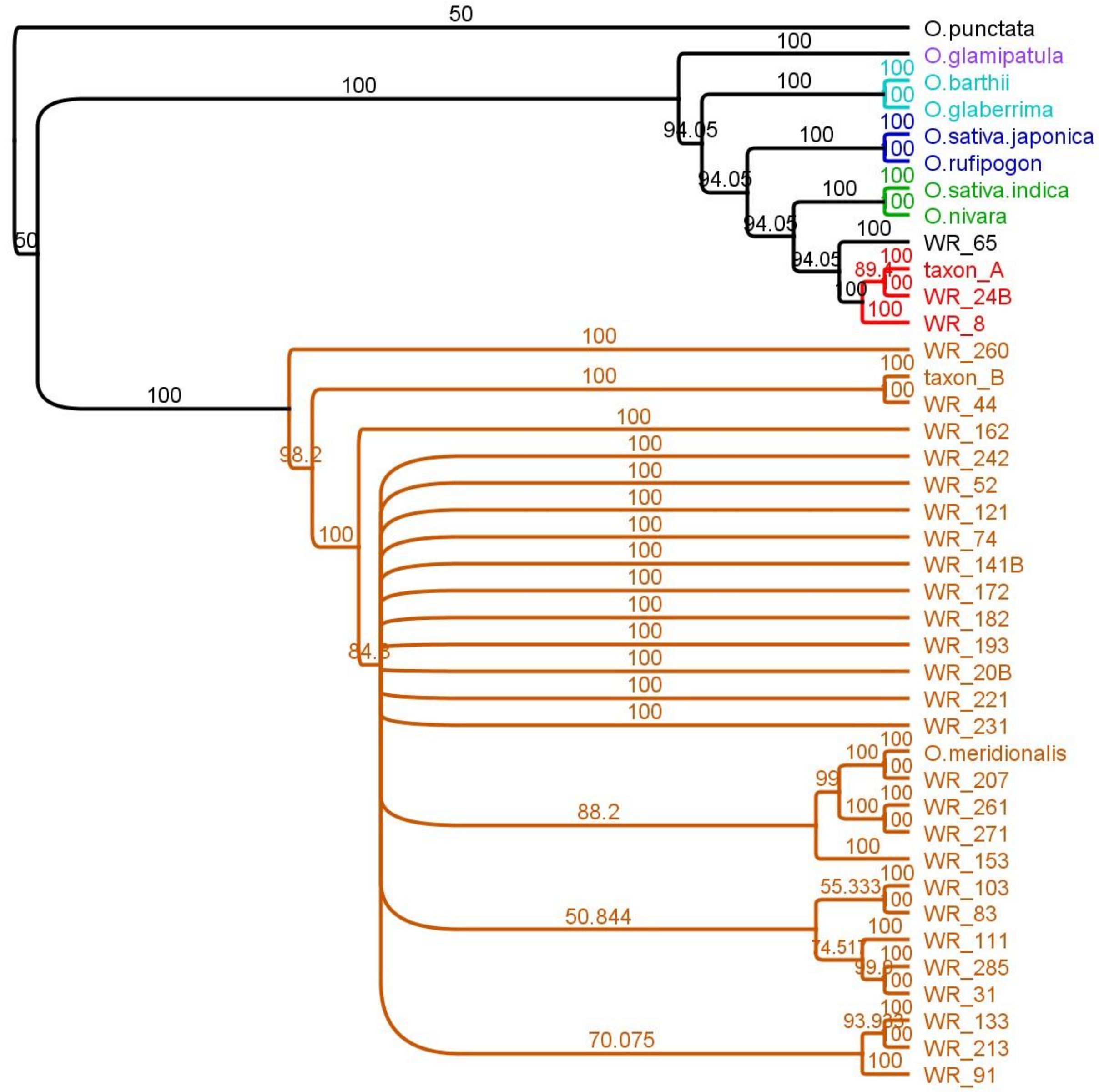


Figure 33 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 1 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches

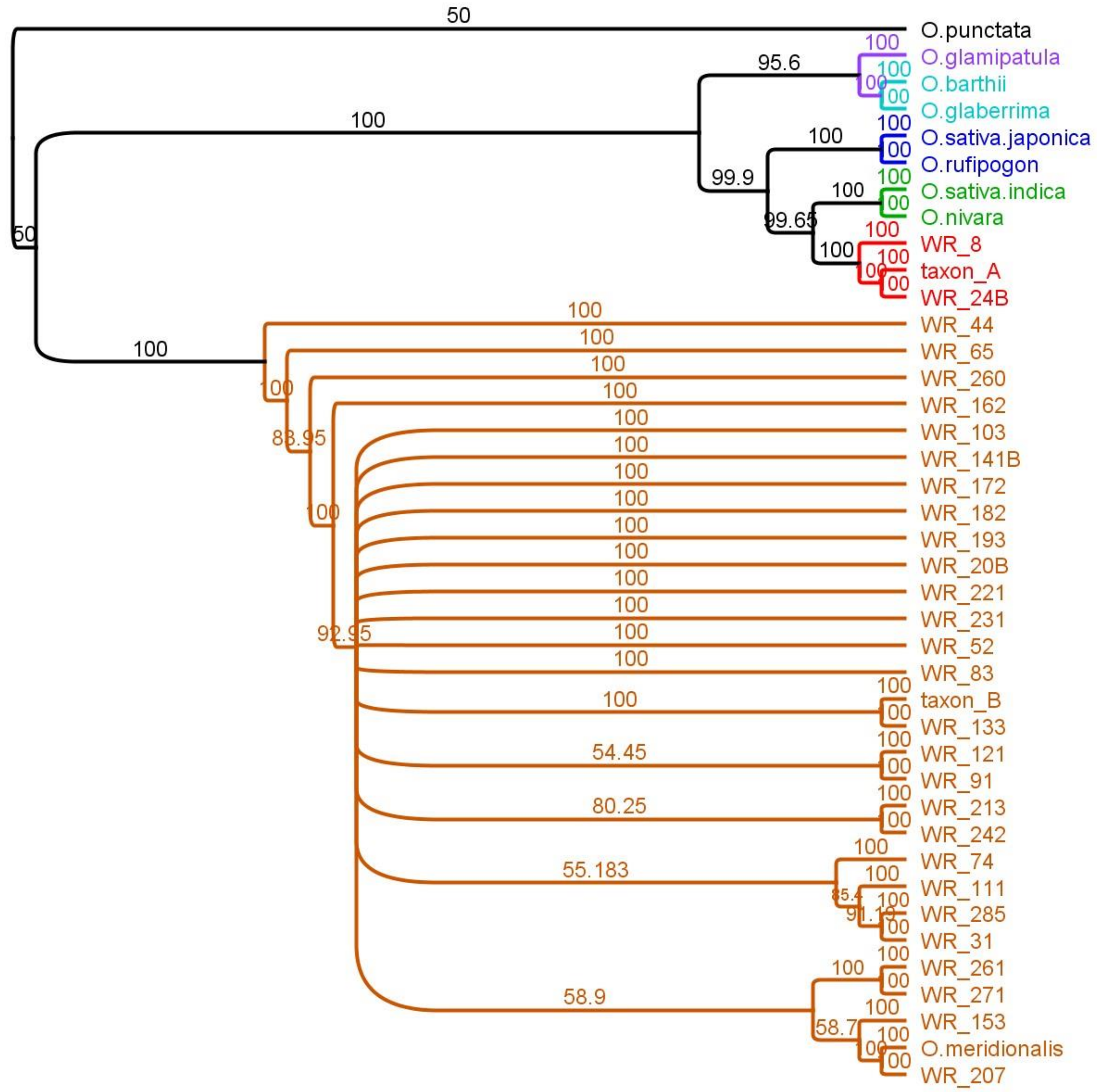


Figure 34 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 2 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches

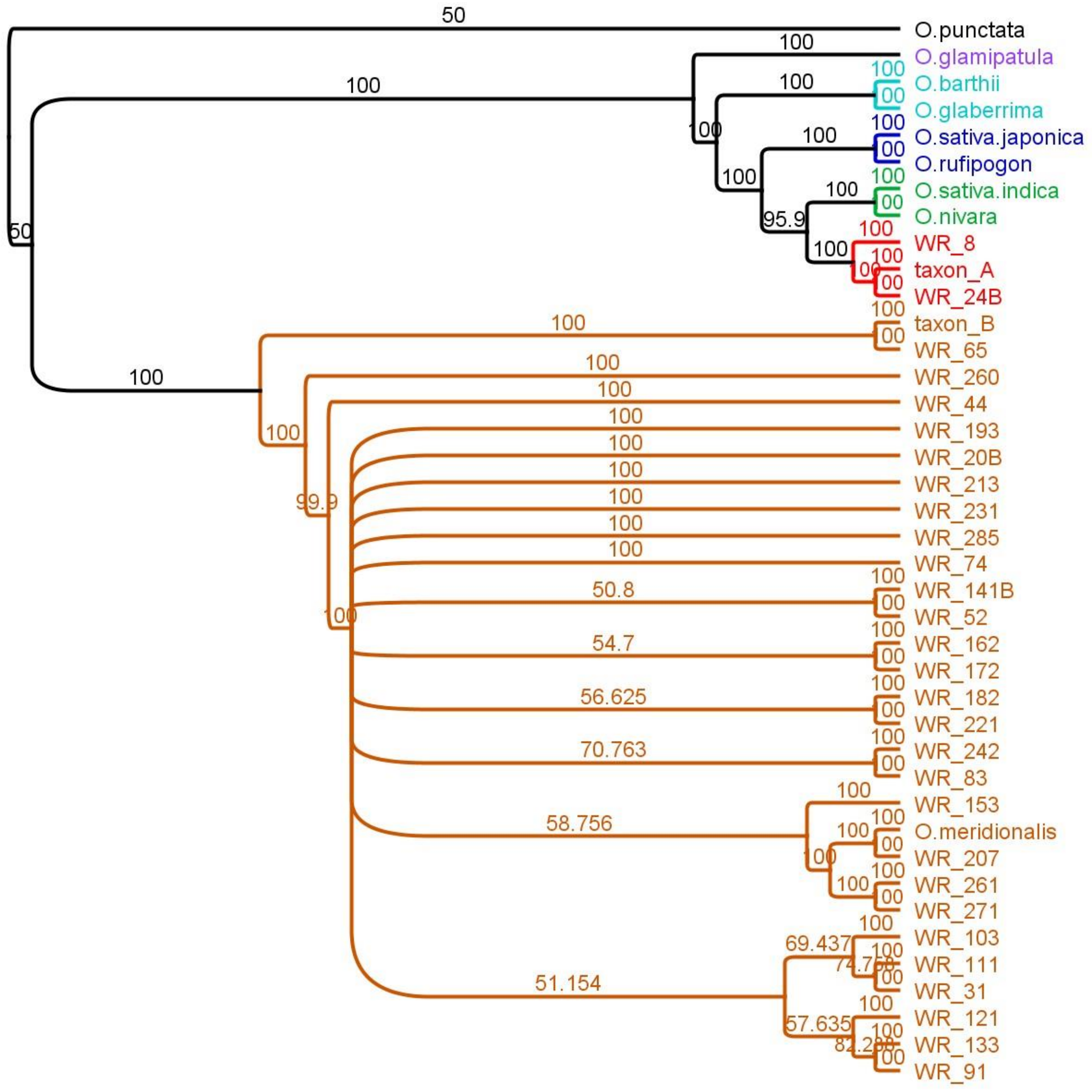


Figure 35 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 3 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches

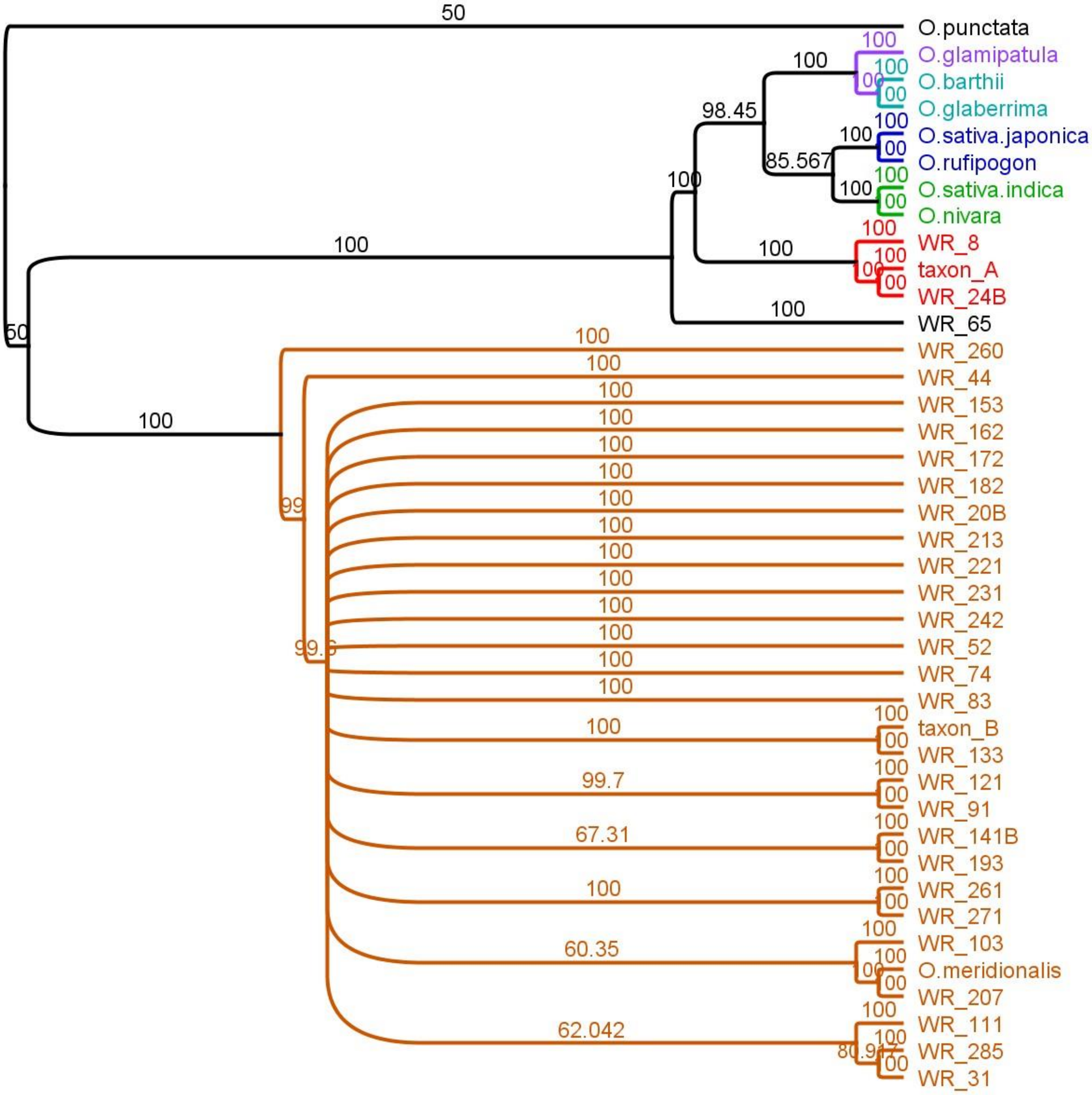


Figure 36 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 4 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches

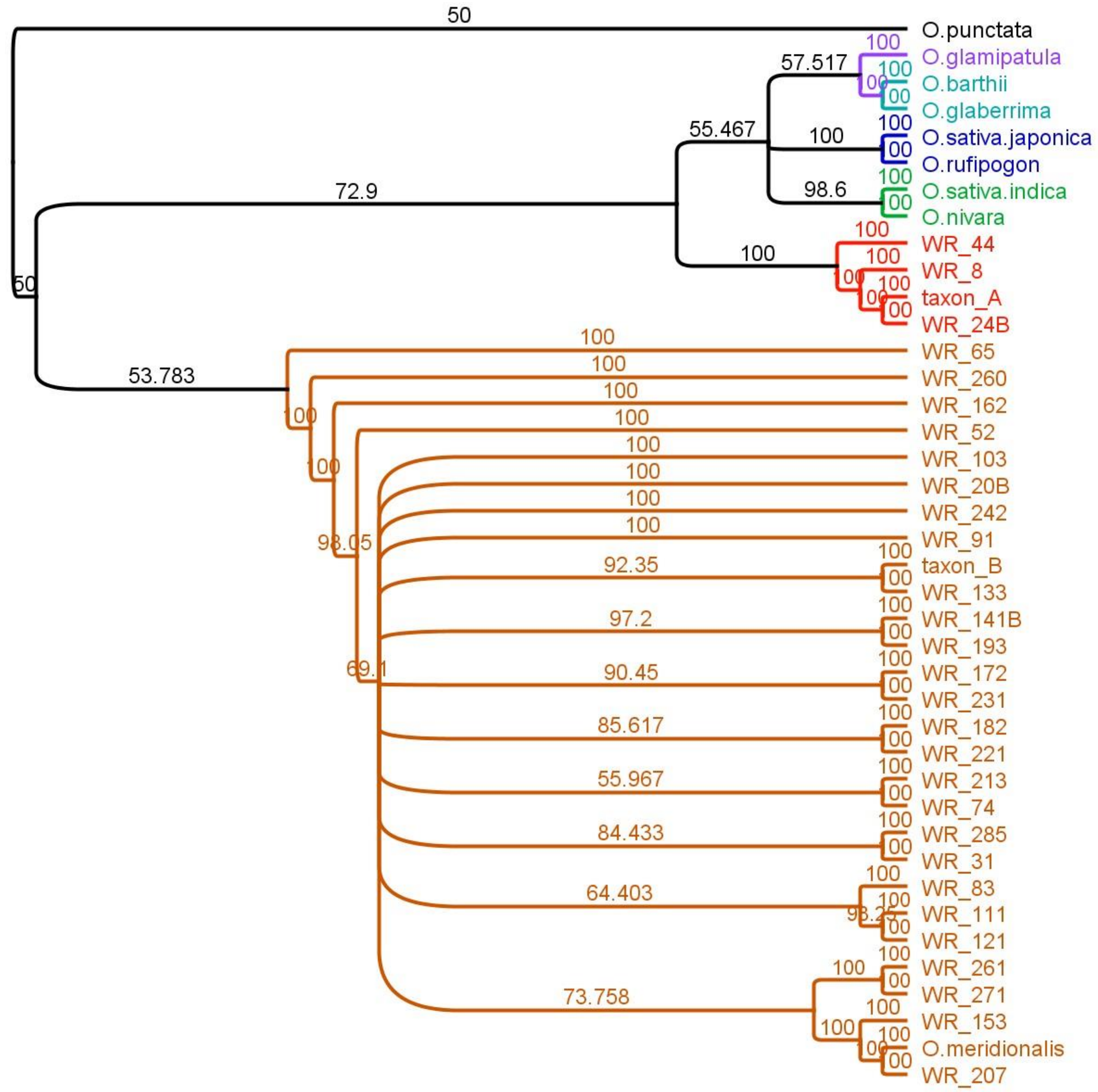


Figure 37 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 5 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

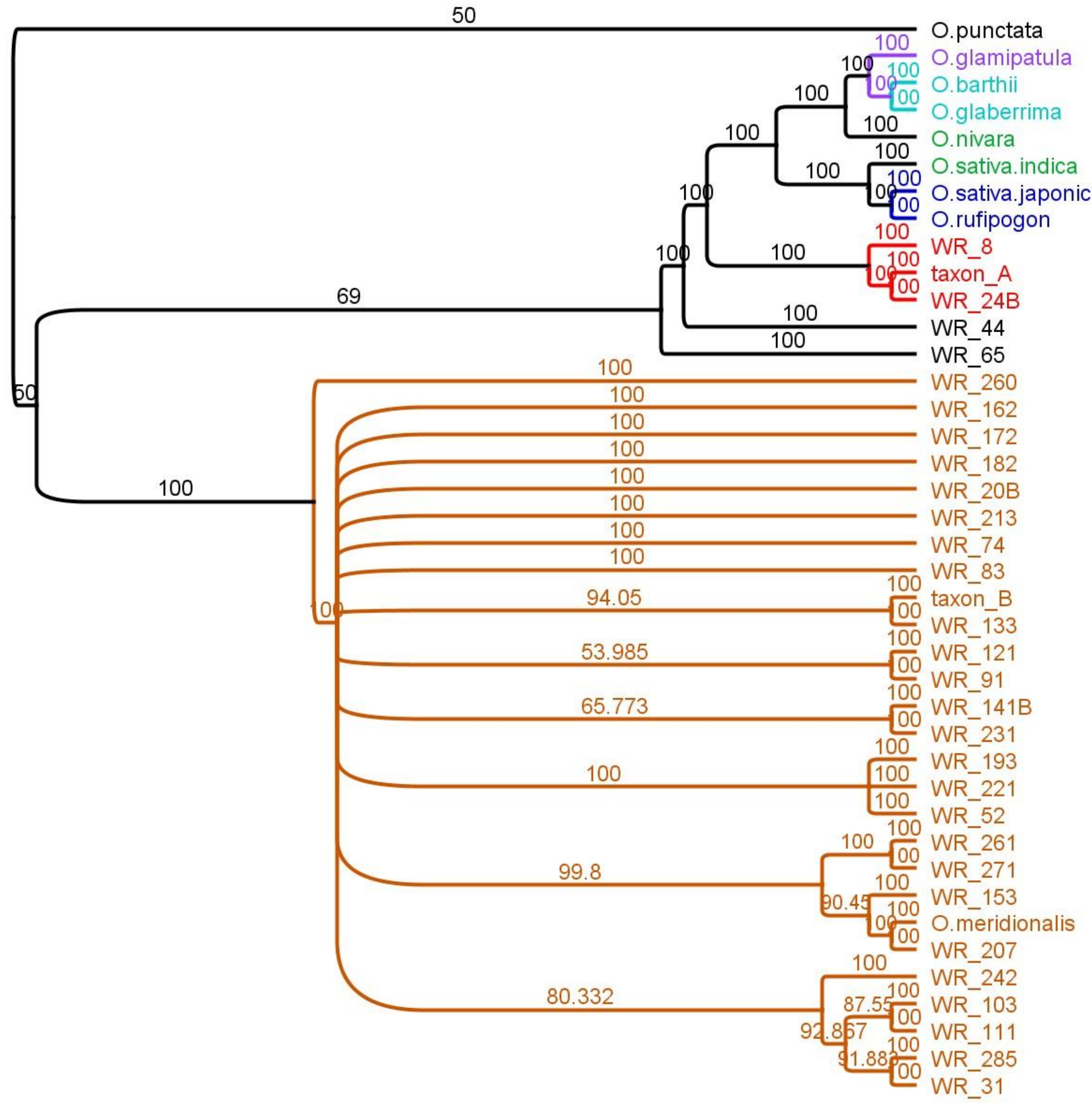


Figure 38 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 6 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

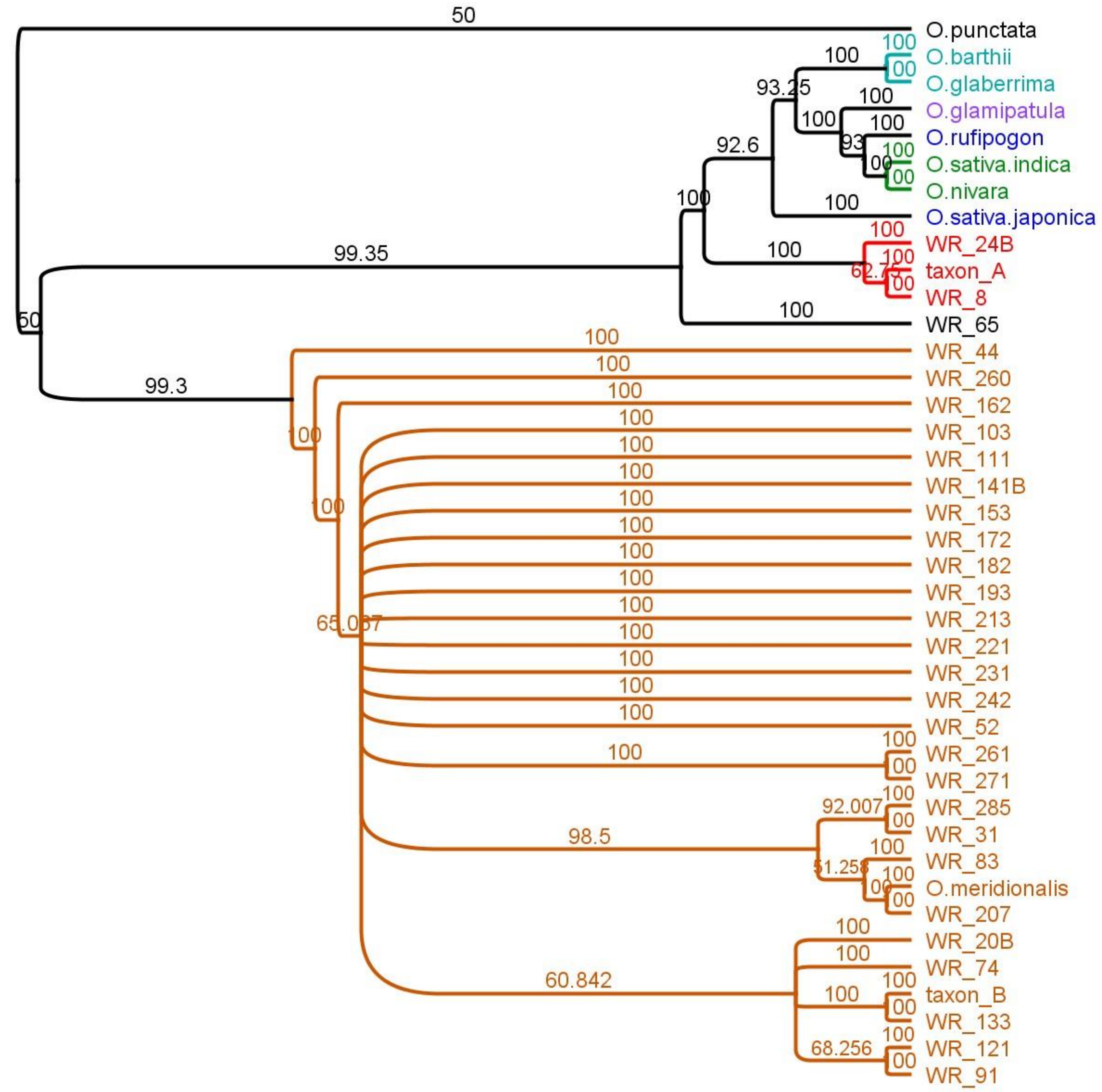


Figure 39 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 7 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

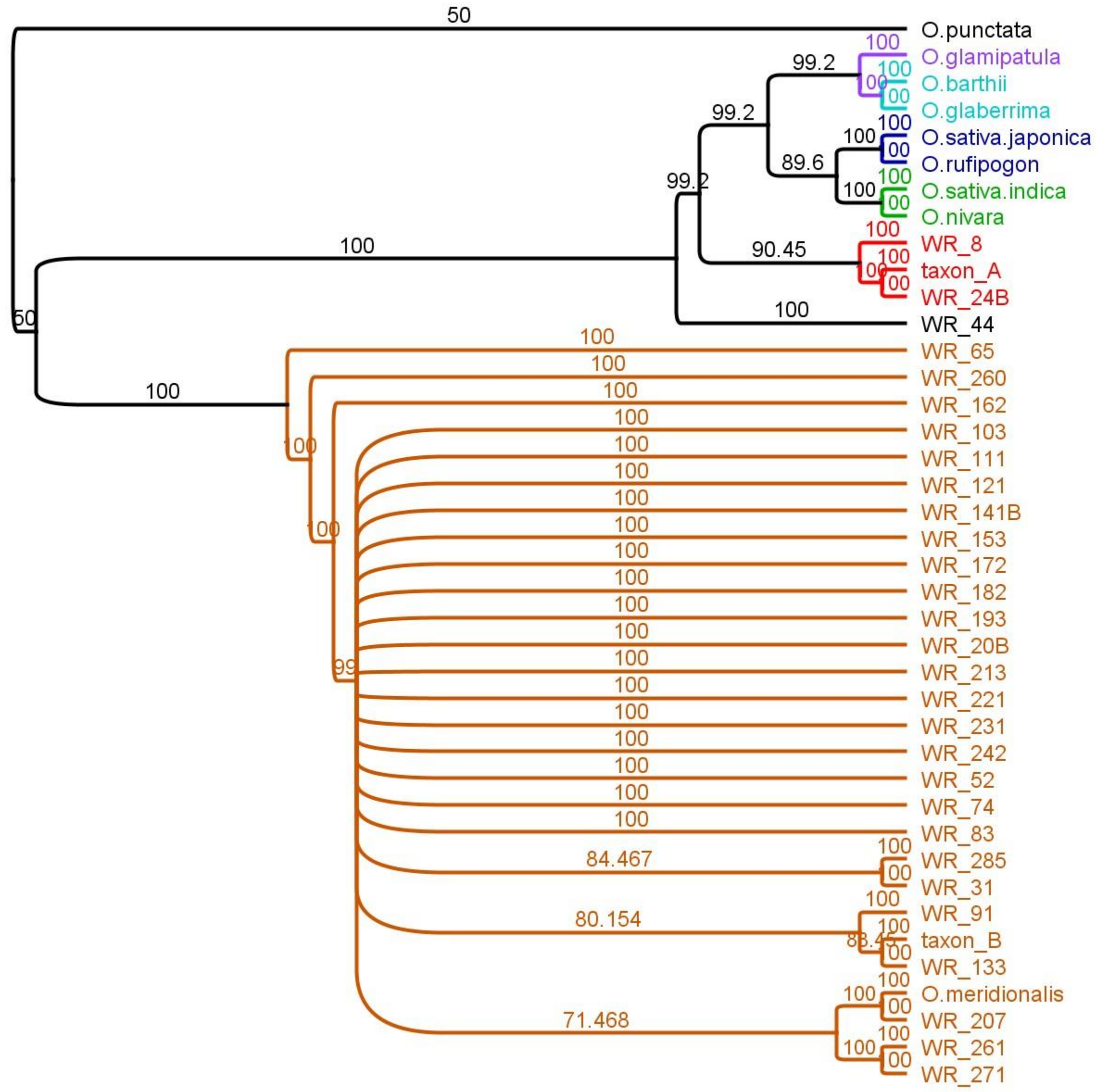


Figure 40 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 8 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

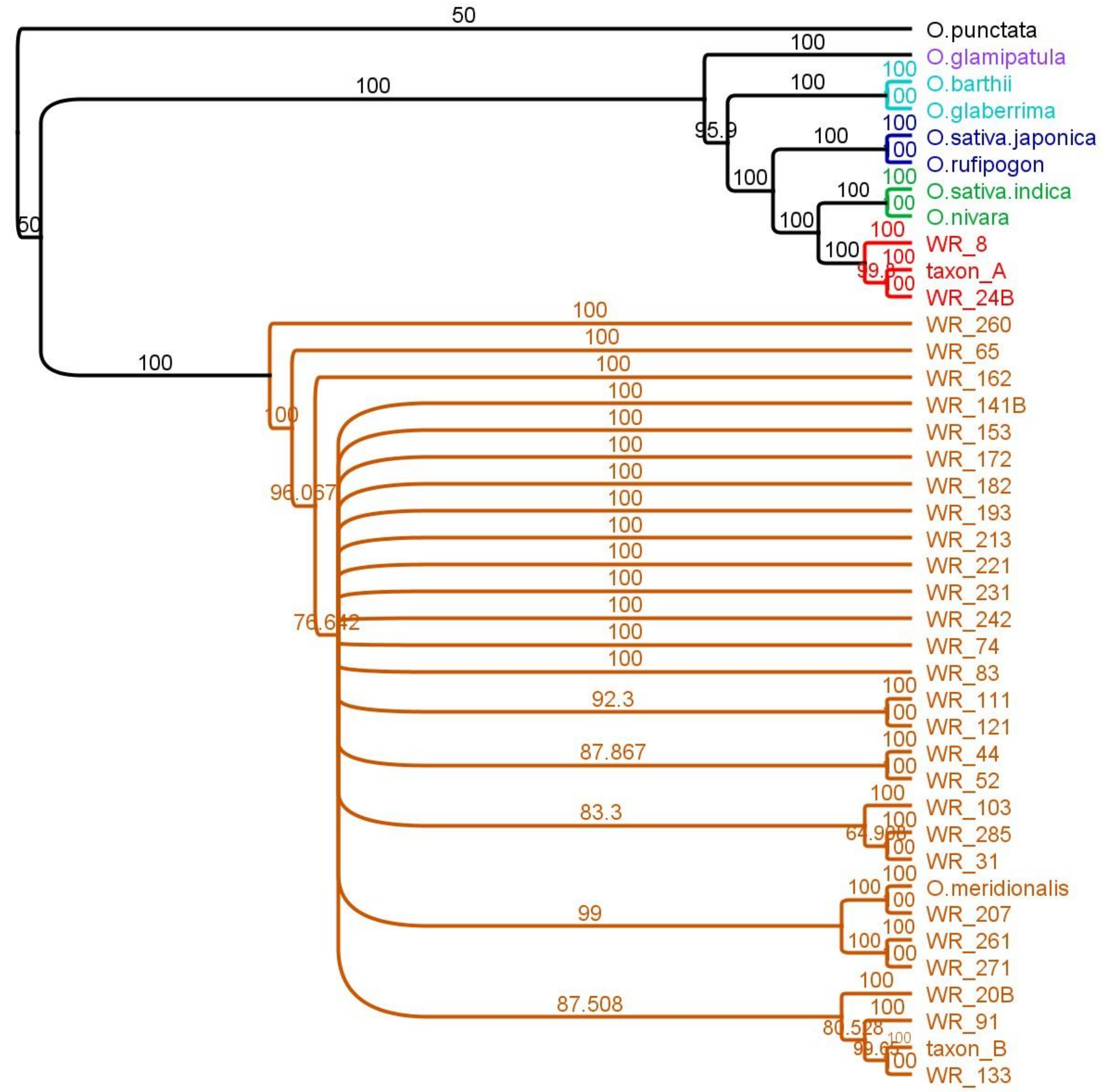


Figure 41 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 9 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

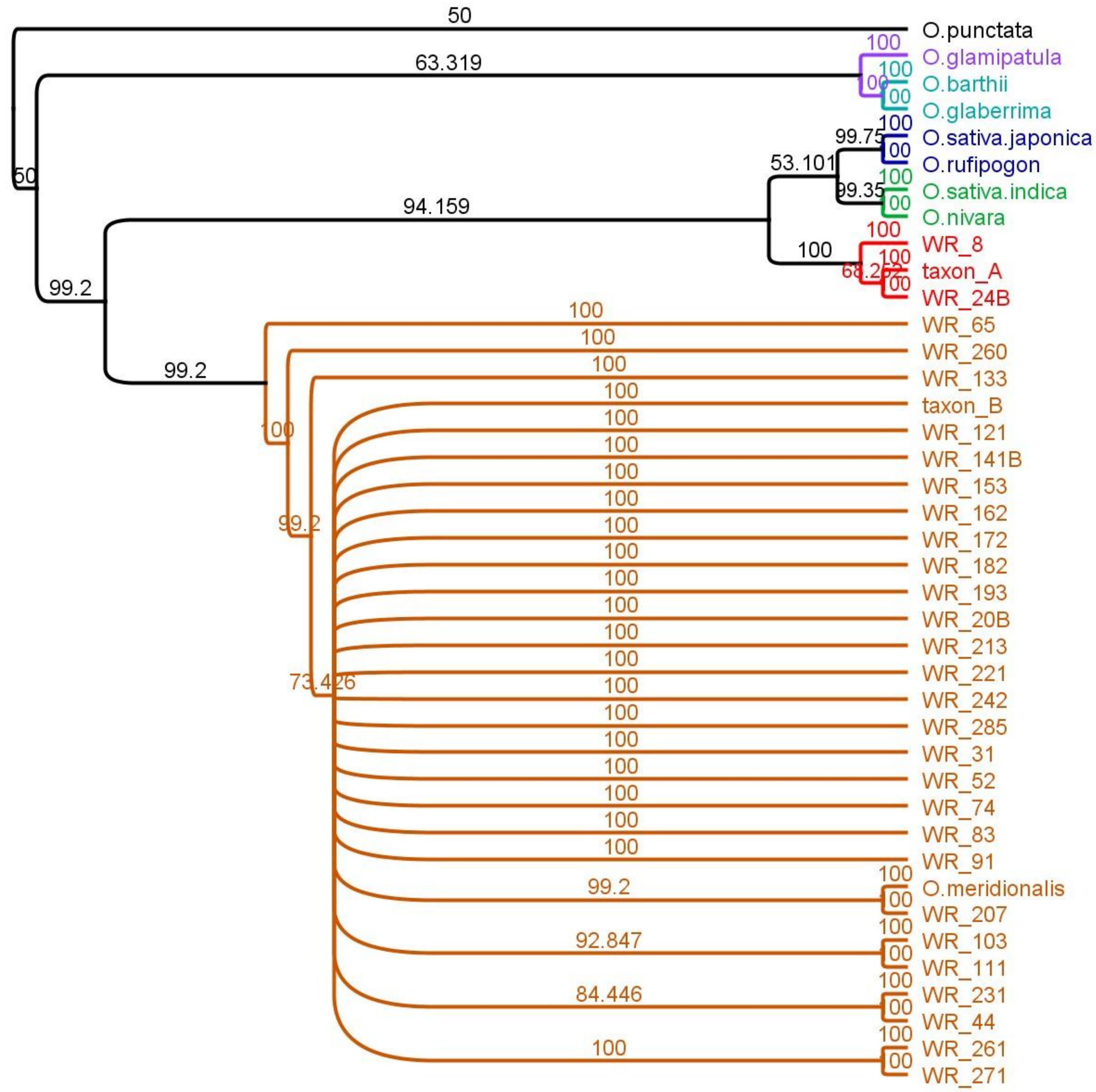


Figure 42 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 10 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

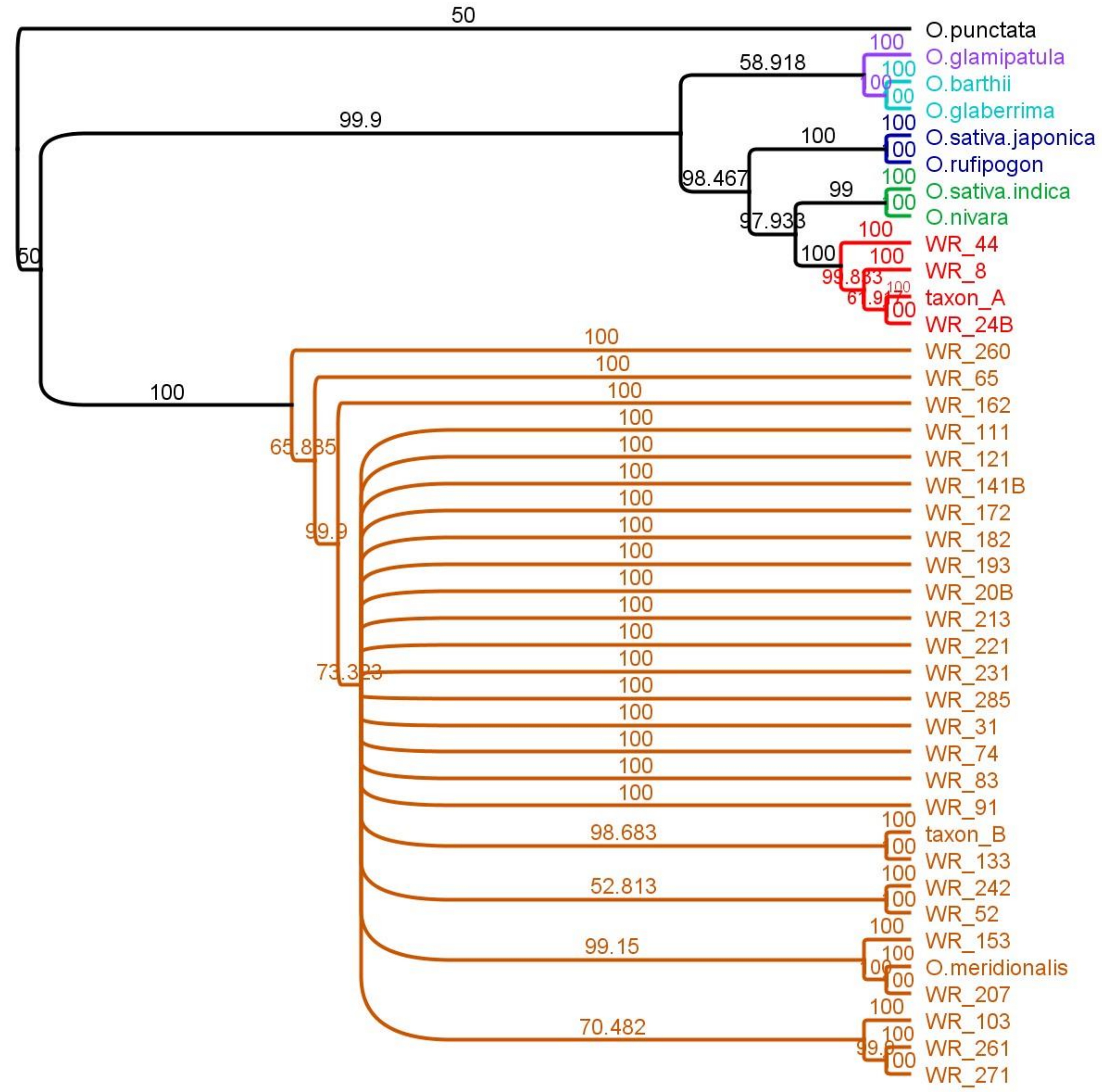


Figure 43 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 11 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

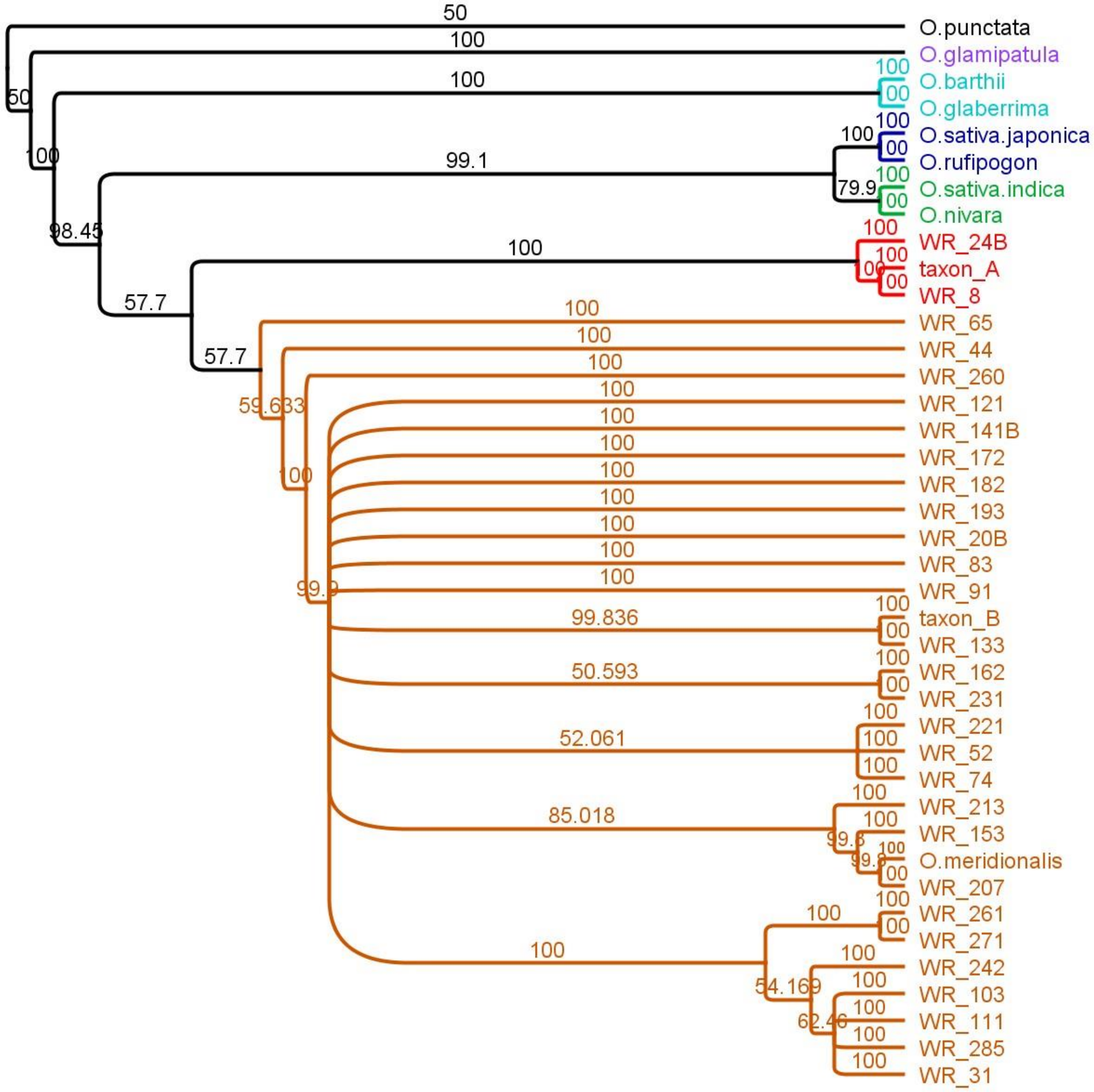


Figure 44 Maximum Parsimony phylogenetic tree analysis of the concatenated alignment of chromosome 12 genes. Colours relate to the main clades. Red and Brown clades are from Australia. Bootstrap value of 1000 replicates are shown on the branches.

5 Appendix 5

Table 35 Non synonyms nucleotide polymorphism in 13 starch related gene. Gene, protein and amino acid substitutions are shown. Colours: Green taxon A, orange taxon B yellow in both

GBSS-I													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	FREQUENCY
G	A	2,054	LOC4340018	XP_015644486.1	I -> V	165	493	1	A -> G	ATC -> GTC	SNP (transition)	Substitution	2
ISA3													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
C	T	1,719	LOC4347328	XP_015612255.1	I -> T	167	500	2	T -> C	ATA -> ACA	SNP (transition)	Substitution	27
A	G	1,703	LOC4347328	XP_015612255.1	D -> N	162	484	1	G -> A	GAT -> AAT	SNP (transition)	Substitution	27
A	G	7,377	LOC4347328	XP_015612255.1	C -> Y	560	1,679	2	G -> A	TGT -> TAT	SNP (transition)	Substitution	27
SBE1													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
A	G	2,600	LOC4342117	XP_015643111.1	V -> I	50	148	1	G -> A	GTC -> ATC	SNP (transition)	Substitution	24
A	G	3,570	LOC4342117	XP_015643111.1	R -> H	190	569	2	G -> A	CGC -> CAC	SNP (transition)	Substitution	26
A	G	7,215	LOC4342117	XP_015643111.1	R -> H	762	2,285	2	G -> A	CGT -> CAT	SNP (transition)	Substitution	23
C	G	7,293	LOC4342117	XP_015643111.1	G -> A	788	2,363	2	G -> C	GGG -> GCG	SNP (transversion)	Substitution	25
SBE3													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
G	A	6,502	LOC4329532	XP_015627503.1	T -> A	525	1,573	1	A -> G	ACC -> GCC	SNP (transition)	Substitution	1
A	C	8,471	LOC4329532	XP_015627503.1	S -> Y	569	1,706	2	C -> A	TCT -> TAT	SNP (transversion)	Substitution	1
G	A	1,288	LOC4329532	XP_015627503.1	E -> G	120	359	2	A -> G	GAA -> GGA	SNP (transition)	Substitution	25

SBE4													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
T	G	825	LOC4335763	XP_015634245.1	V -> L	62	184	1	G -> T	GTG -> TTG	SNP (transversion)	Substitution	26
G	A	1,151	LOC4335763	XP_015634245.1	N -> D	93	277	1	A -> G	AAT -> GAT	SNP (transition)	Substitution	27
A	G	1,867	LOC4335763	XP_015634245.1	A -> T	213	637	1	G -> A	GCT -> ACT	SNP (transition)	Substitution	24
SS-I													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
A	G	5,406	LOC9269493	XP_015644241.1	G -> E	500	1,499	2	G -> A	GGG -> GAG	SNP (transition)	Substitution	25
A	G	401	LOC9269493	XP_015644241.1	A -> T	72	214	1	G -> A	GCG -> ACG	SNP (transition)	Substitution	24
S-II-1													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
A	G	2,250	LOC4348711	XP_015614561.1	R -> H	215	644	2	G -> A	CGT -> CAT	SNP (transition)	Substitution	3
T	G	2,366	LOC4348711	XP_015614561.1	A -> S	254	760	1	G -> T	GCT -> TCT	SNP (transversion)	Substitution	2
C	T	1,107	LOC4348711	XP_015614561.1	V -> A	115	344	2	T -> C	GTT -> GCT	SNP (transition)	Substitution	27
A	G	6,643	LOC4348711	XP_015614561.1	G -> E	498	1,493	2	G -> A	GGG -> GAG	SNP (transition)	Substitution	27
A	G	2,093	LOC4348711	XP_015614561.1	A -> T	163	487	1	G -> A	GCA -> ACA	SNP (transition)	Substitution	27
SS-II-2													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
C	T	1,085	LOC4330709	XP_015627452.1	S -> P	130	388	1	T -> C	TCT -> CCT	SNP (transition)	Substitution	3
T	C	322	LOC4330709	XP_015627452.1	P -> L	9	26	2	C -> T	CCG -> CTG	SNP (transition)	Substitution	3
G	A	4,373	LOC4330709	XP_015627452.1	N -> S	653	1,958	2	A -> G	AAC -> AGC	SNP (transition)	Substitution	2
G	A	522	LOC4330709	XP_015627452.1	T -> A	76	226	1	A -> G	ACG -> GCG	SNP (transition)	Substitution	29
C	A	1,014	LOC4330709	XP_015627452.1	Y -> S	106	317	2	A -> C	TAC -> TCC	SNP (transversion)	Substitution	29

T	C	3,866	LOC4330709	XP_015627452.1	T -> M	484	1,451	2	C -> T	ACG -> ATG	SNP (transition)	Substitution	27
G	C	983	LOC4330709	XP_015627452.1	H -> D	96	286	1	C -> G	CAT -> GAT	SNP (transversion)	Substitution	27
T	C	1,149	LOC4330709	XP_015627452.1	A -> V	151	452	2	C -> T	GCT -> GTT	SNP (transition)	Substitution	27
C	A	3,896	LOC4330709	XP_015627452.1	E -> A	494	1,481	2	A -> C	GAG -> GCG	SNP (transversion)	Substitution	27
SS-II-3													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
T	A	1,058	LOC4340567	XP_015644246.1	K -> M	244	731	2	A -> T	AAG -> ATG	SNP (transversion)	Substitution	2
A	G	1,190	LOC4340567	XP_015644246.1	G -> D	288	863	2	G -> A	GGC -> GAC	SNP (transition)	Substitution	2
G	A	4,394	LOC4340567	XP_015644246.1	M -> V	737	2,209	1	A -> G	ATG -> GTG	SNP (transition)	Substitution	28
G	A	3,995	LOC4340567	XP_015644246.1	S -> G	604	1,810	1	A -> G	AGC -> GGC	SNP (transition)	Substitution	27
C	A	889	LOC4340567	XP_015644246.1	T -> P	188	562	1	A -> C	ACG -> CCG	SNP (transversion)	Substitution	24
T	G	894	LOC4340567	XP_015644246.1	K -> N	189	567	3	G -> T	AAG -> AAT	SNP (transversion)	Substitution	24
G	A	413	LOC4340567	XP_015644246.1	D -> G	72	215	2	A -> G	GAT -> GGT	SNP (transition)	Substitution	23
SS-III													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
C	A	1,392	LOC4337056	XP_015636215.1	K -> N	207	621	3	A -> C	AAA -> AAC	SNP (transversion)	Substitution	2
C	T	5,259	LOC4337056	XP_015636215.1	W -> R	858	2,572	1	T -> C	TGG -> CGG	SNP (transition)	Substitution	2
G	C	6,923	LOC4337056	XP_015636215.1	L -> V	1,103	3,307	1	C -> G	CTT -> GTT	SNP (transversion)	Substitution	3
A	G	4,097	LOC4337056	XP_015636215.1	V -> I	762	2,284	1	G -> A	GTT -> ATT	SNP (transition)	Substitution	3
T	G	4,080	LOC4337056	XP_015636215.1	S -> I	756	2,267	2	G -> T	AGT -> ATT	SNP (transversion)	Substitution	29
G	A	1,561	LOC4337056	XP_015636215.1	R -> G	264	790	1	A -> G	AGG -> GGG	SNP (transition)	Substitution	27
G	A	1,588	LOC4337056	XP_015636215.1	K -> E	273	817	1	A -> G	AAA -> GAA	SNP (transition)	Substitution	27
G	A	2,035	LOC4337056	XP_015636215.1	N -> D	422	1,264	1	A -> G	AAT -> GAT	SNP (transition)	Substitution	27

A	G	6,449	LOC4337056	XP_015636215.1	R -> K	1,011	3,032	2	G -> A	AGG -> AAG	SNP (transition)	Substitution	27
A	G	6,478	LOC4337056	XP_015636215.1	V -> I	1,021	3,061	1	G -> A	GTT -> ATT	SNP (transition)	Substitution	27
SS-IV													
Base	Reference	Reference position	Gene ID	Protein ID	Amino Acid Change	CDS Codon Number	CDS Position	CDS Position Within Codon	Change	Codon Change	Polymorphism Type	Protein Effect	Frequency
C	T	5,407	LOC4331078	XP_015626202.1	C -> R	510	1,528	1	T -> C	TGT -> CGT	SNP (transition)	Substitution	3
T	C	354	LOC4331078	XP_015626202.1	R -> W	50	148	1	C -> T	CGG -> TGG	SNP (transition)	Substitution	3
T	C	1,126	LOC4331078	XP_015626202.1	S -> L	131	392	2	C -> T	TCG -> TTG	SNP (transition)	Substitution	3
T	C	6,528	LOC4331078	XP_015626202.1	P -> S	624	1,870	1	C -> T	CCC -> TCC	SNP (transition)	Substitution	2
G	C	860	LOC4331078	XP_015626202.1	Q -> E	103	307	1	C -> G	CAG -> GAG	SNP (transversion)	Substitution	29
T	C	8,015	LOC4331078	XP_015626201.1	S -> N	23	68	2	C -> T	AGT -> AAT	SNP (transition)	Substitution	17
T	C	8,039	LOC4331078	XP_015626201.1	G -> E	15	44	2	C -> T	GGA -> GAA	SNP (transition)	Substitution	17
A	G	525	LOC4331078	XP_015626202.1	A -> T	61	181	1	G -> A	GCT -> ACT	SNP (transition)	Substitution	16

*For more details please see the excel file

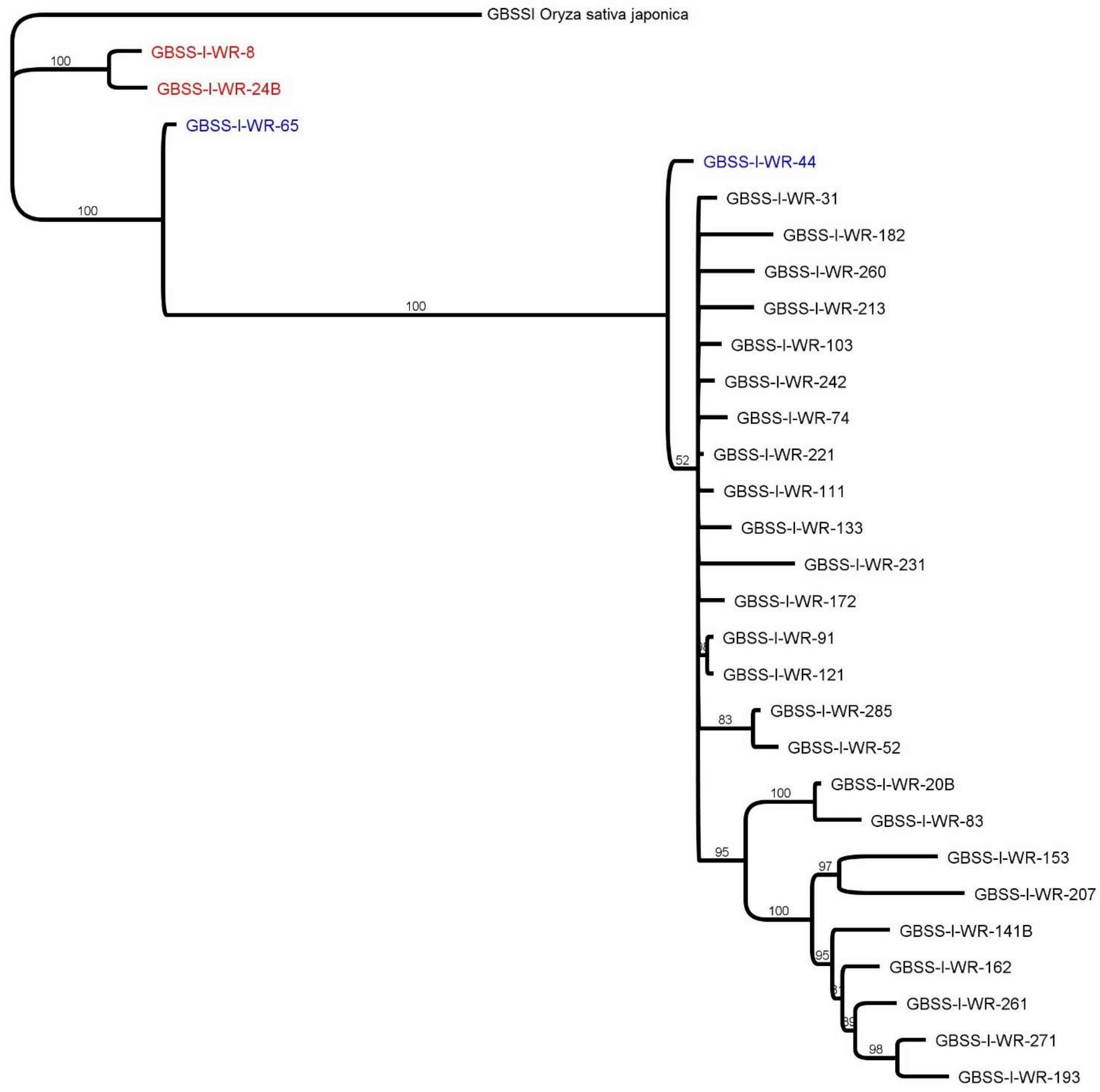


Figure 45 Phylogenetic tree based on Bayesian analysis of GBSSI gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids

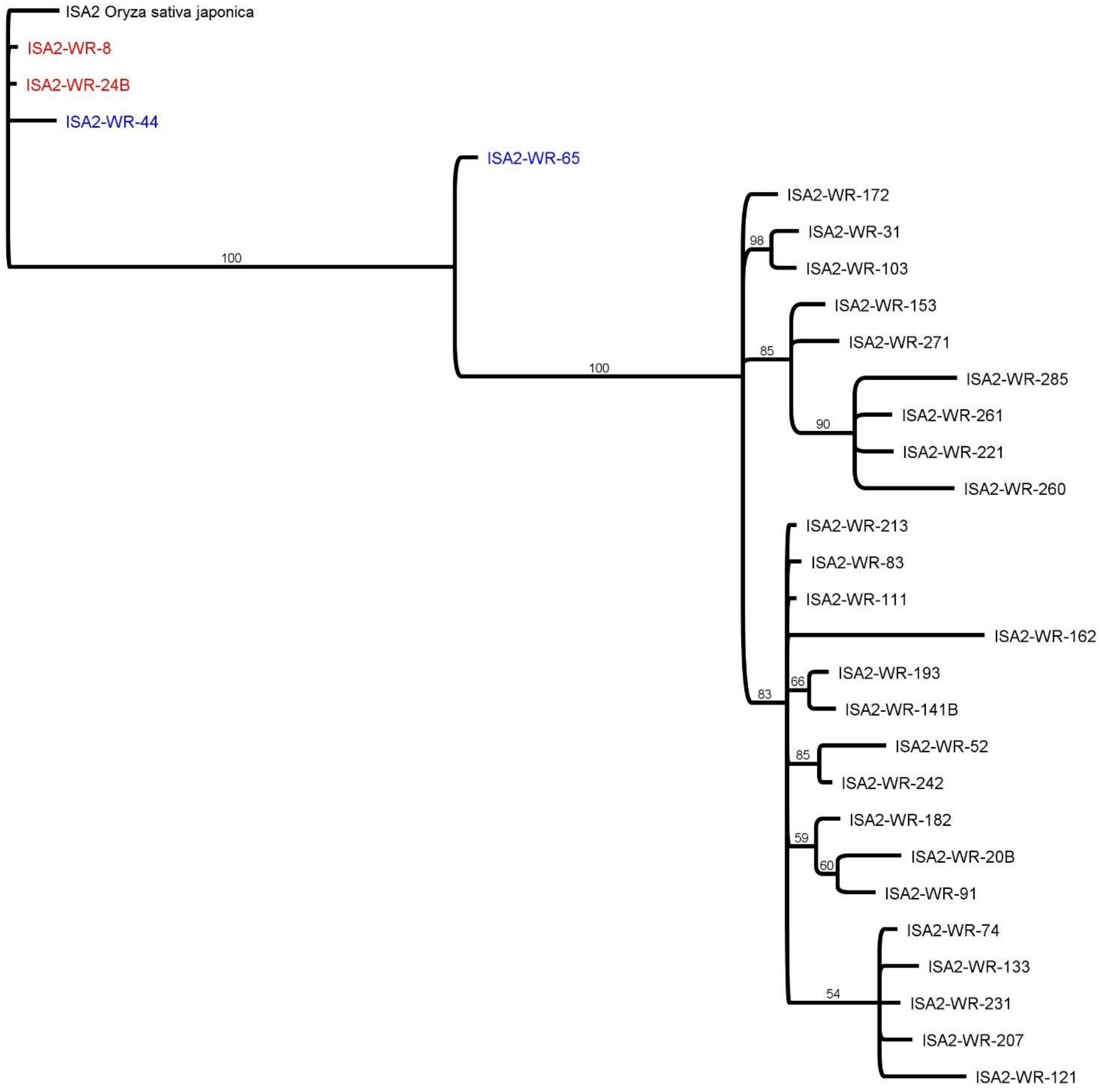


Figure 46 Phylogenetic tree based on Bayesian analysis of ISA2 gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids.

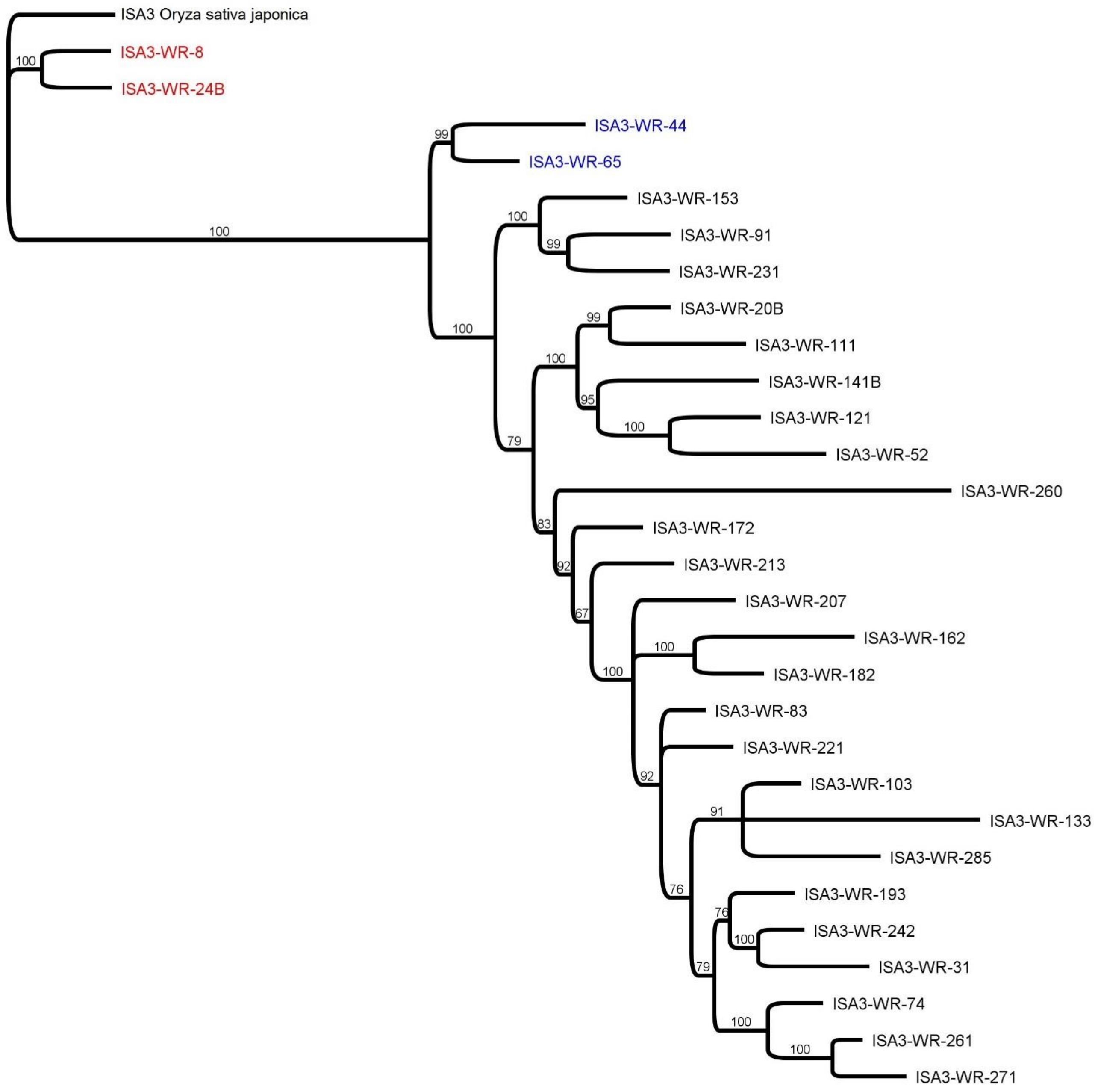


Figure 47 Phylogenetic tree based on Bayesian analysis of ISA3 gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids.

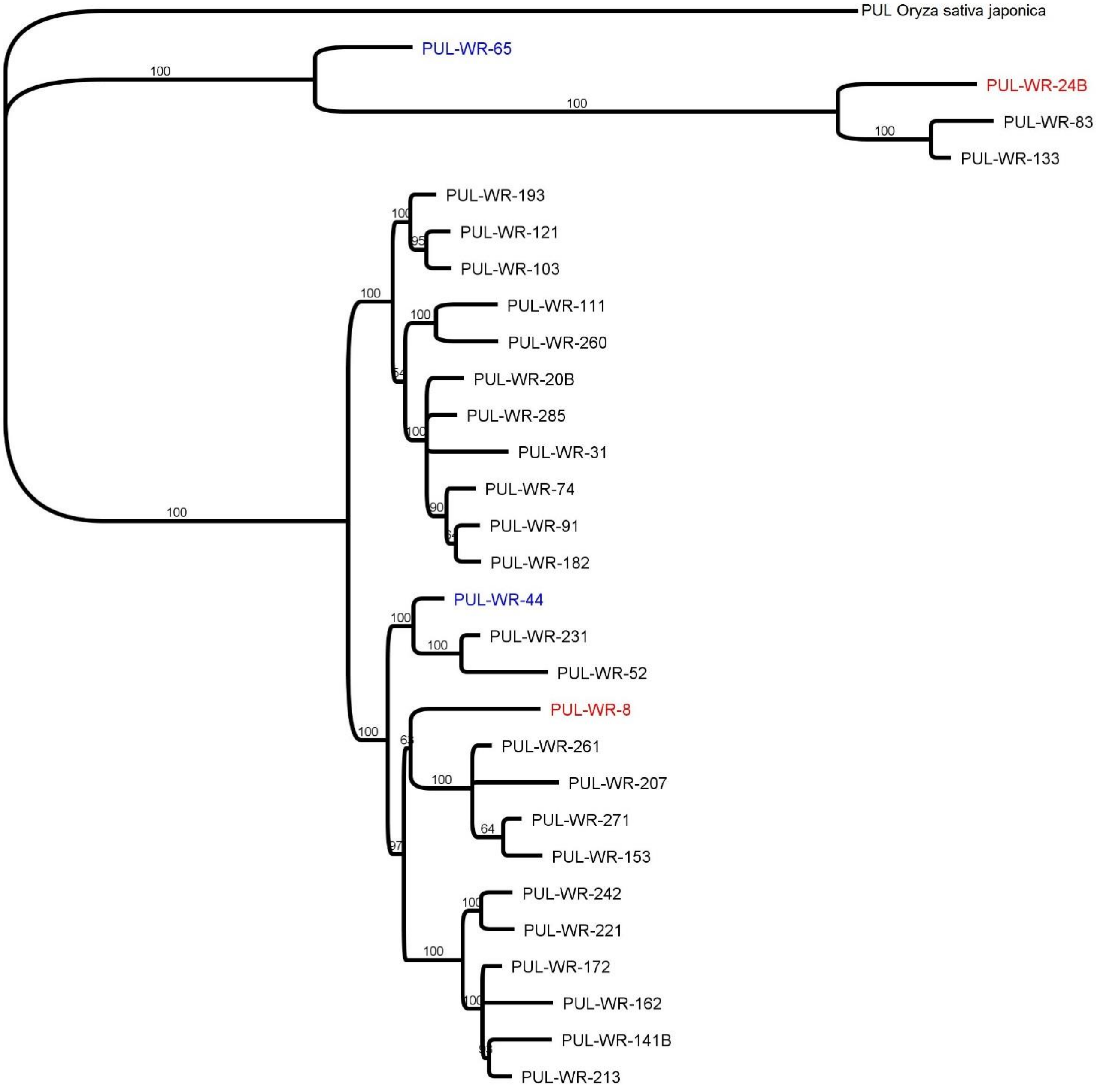


Figure 48 Phylogenetic tree based on Bayesian analysis of *PUL* gene. Bootstrap values (1000 replicates) are shown on the branches.

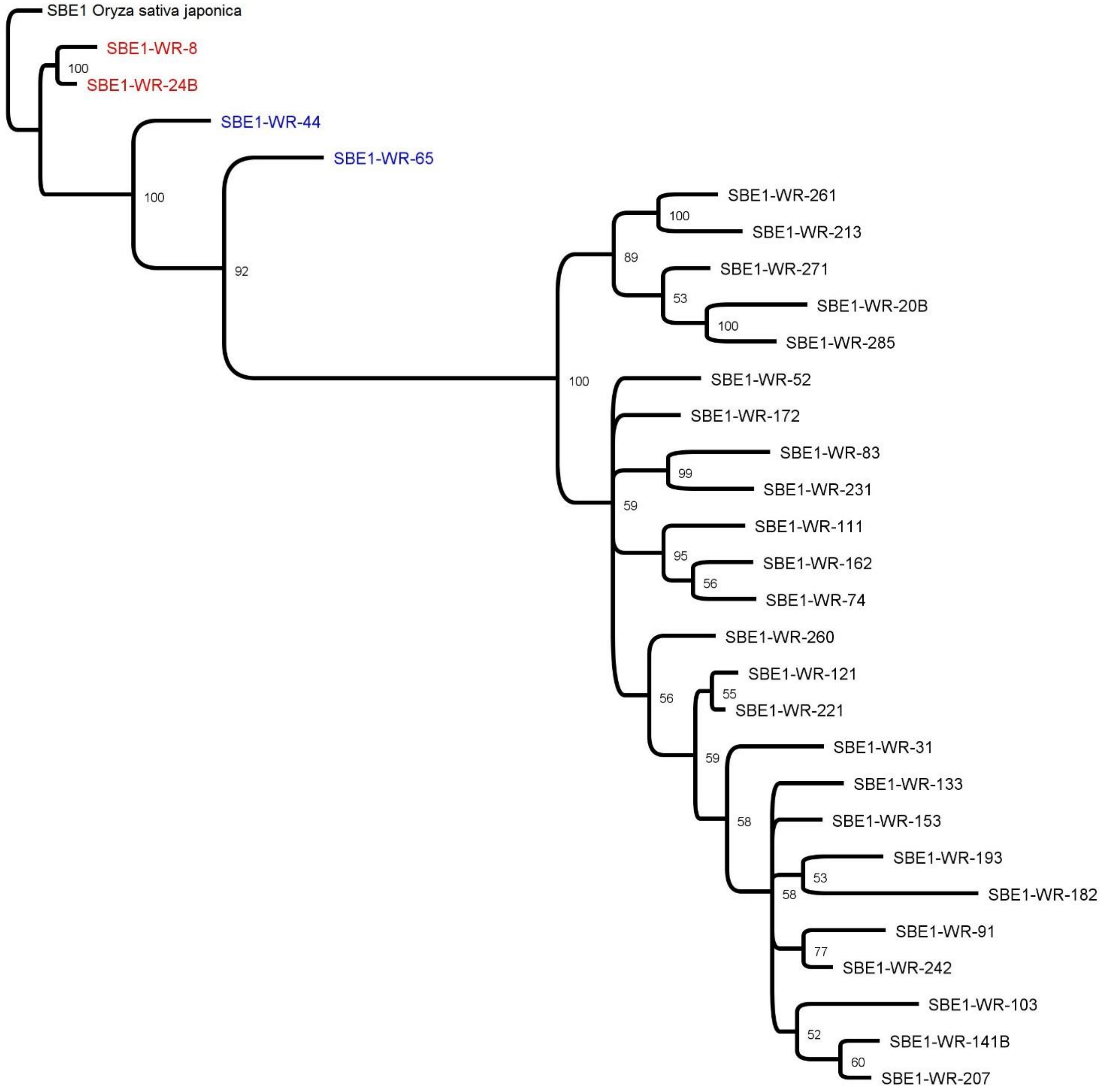


Figure 49 Phylogenetic tree based on Bayesian analysis of *SBE1* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids.

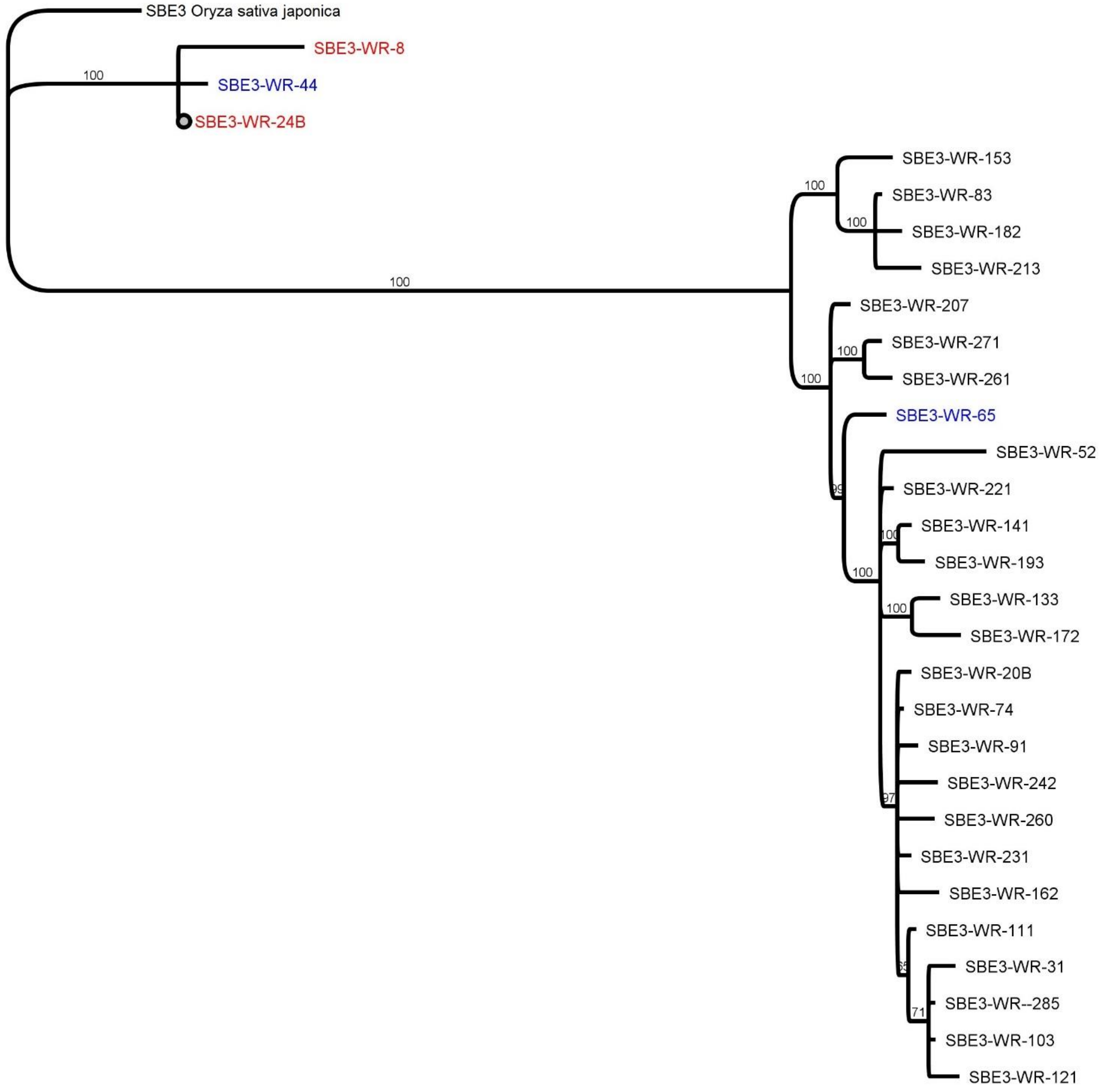


Figure 50 Phylogenetic tree based on Bayesian analysis of *SBE3* gene. Bootstrap values (1000 replicates) are shown on the branches.

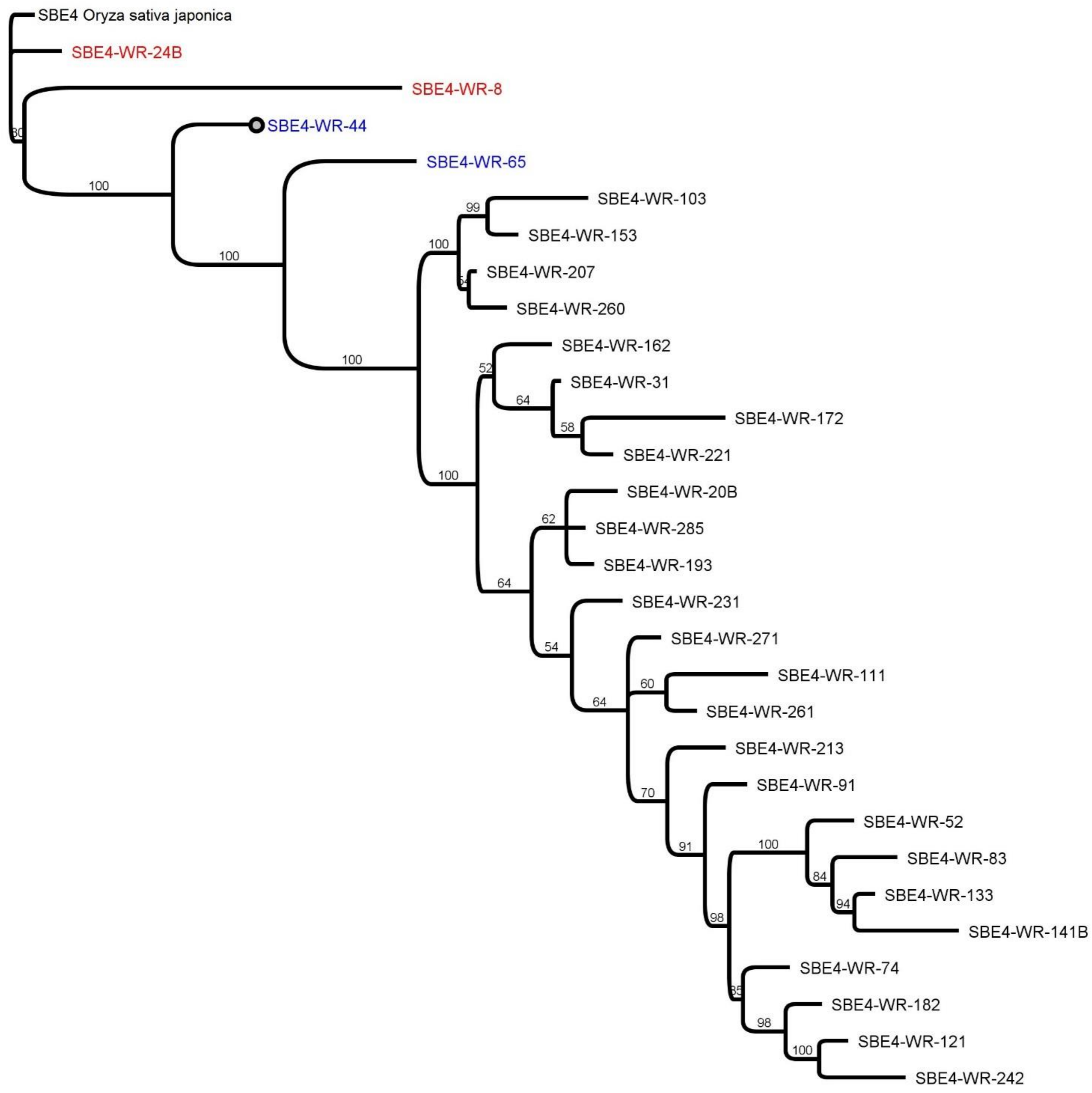


Figure 51 Phylogenetic tree based on Bayesian analysis of *SBE4* gene. Bootstrap values (1000 replicates) are shown on the branches.

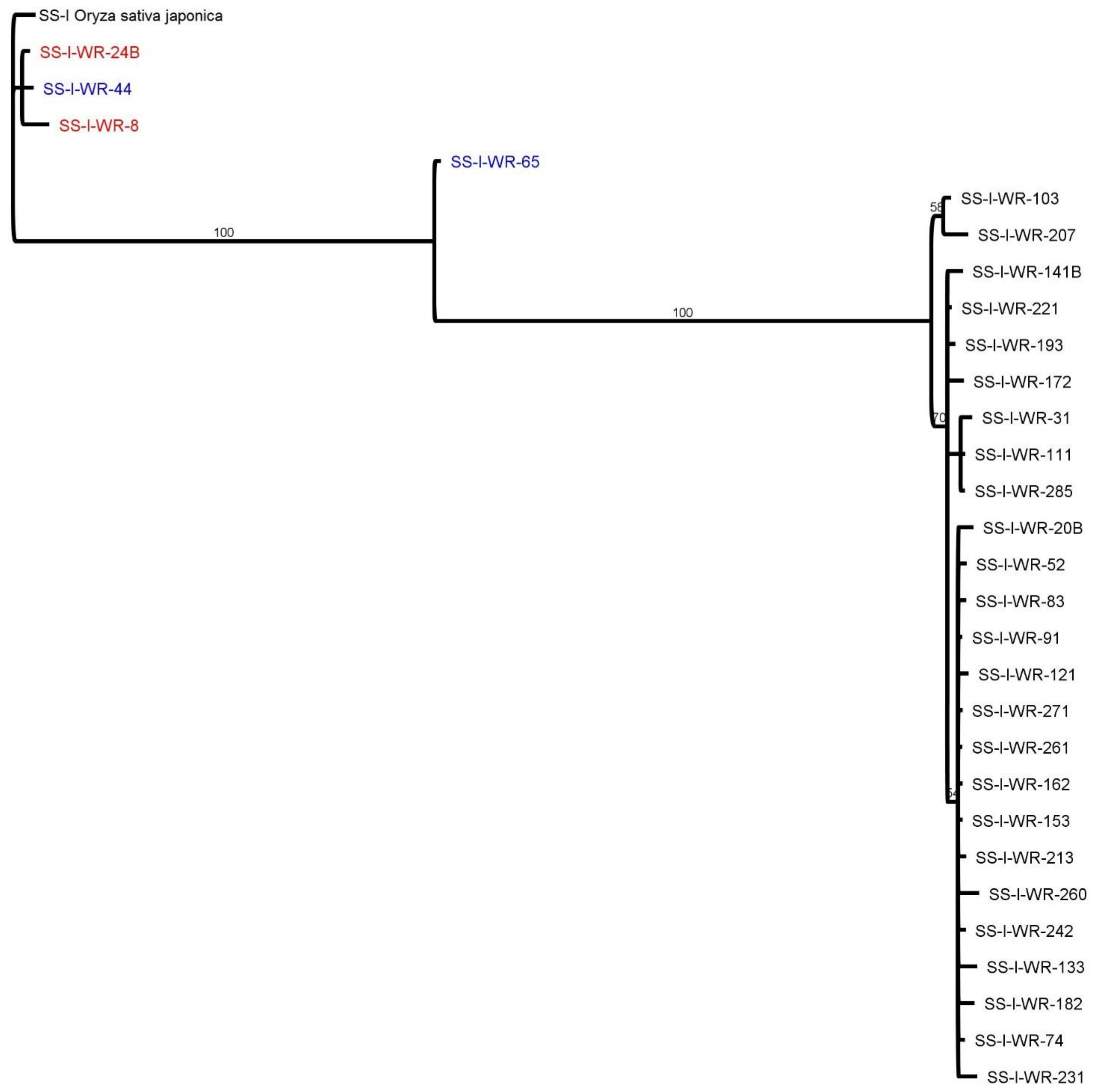


Figure 52 Phylogenetic tree based on Bayesian analysis of *SSI* gene. Bootstrap values (1000 replicates) are shown on the branches.

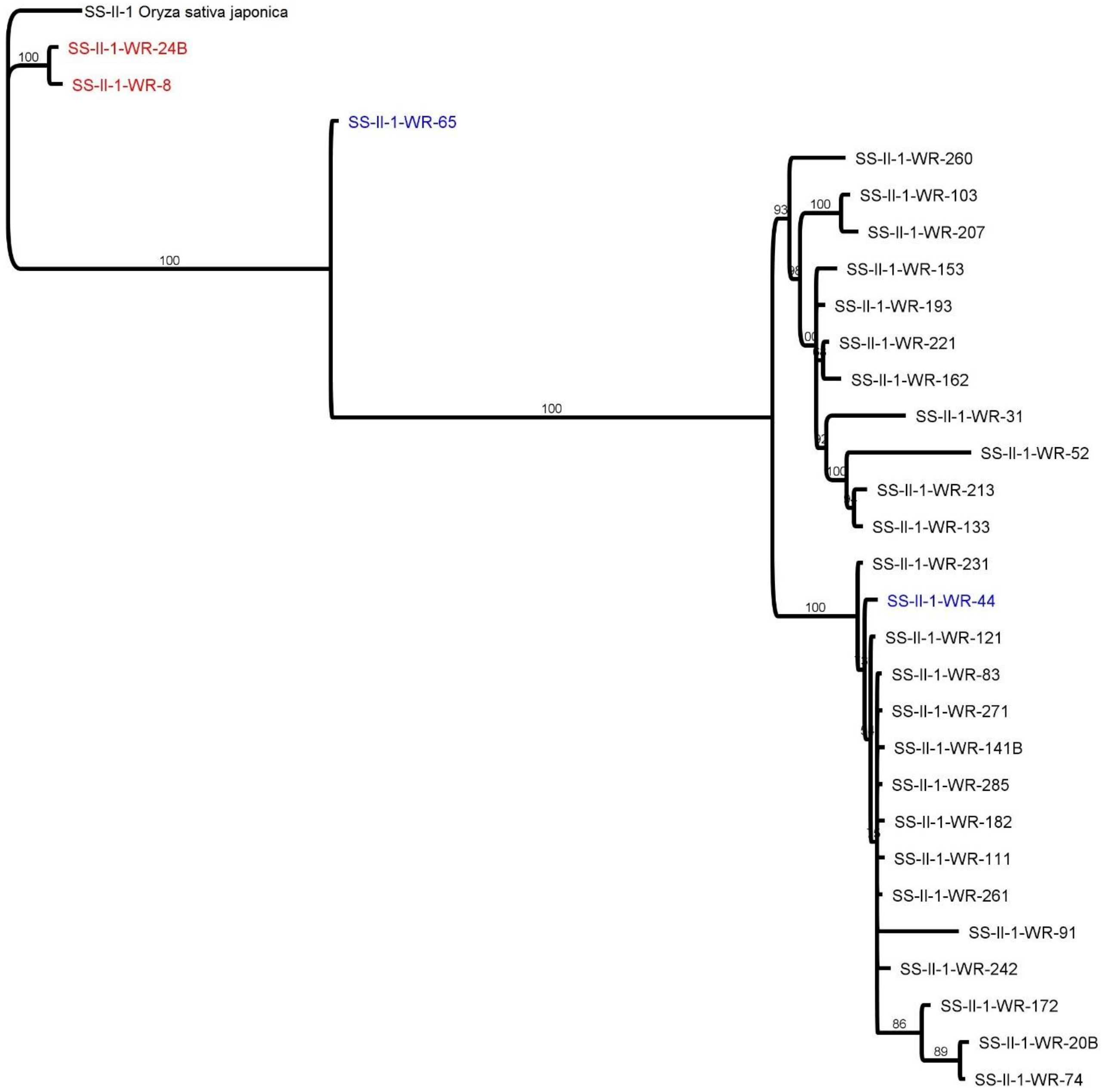


Figure 53 Phylogenetic tree based on Bayesian analysis of *SSII-1* gene. Bootstrap values (1000 replicates) are shown on the branches.

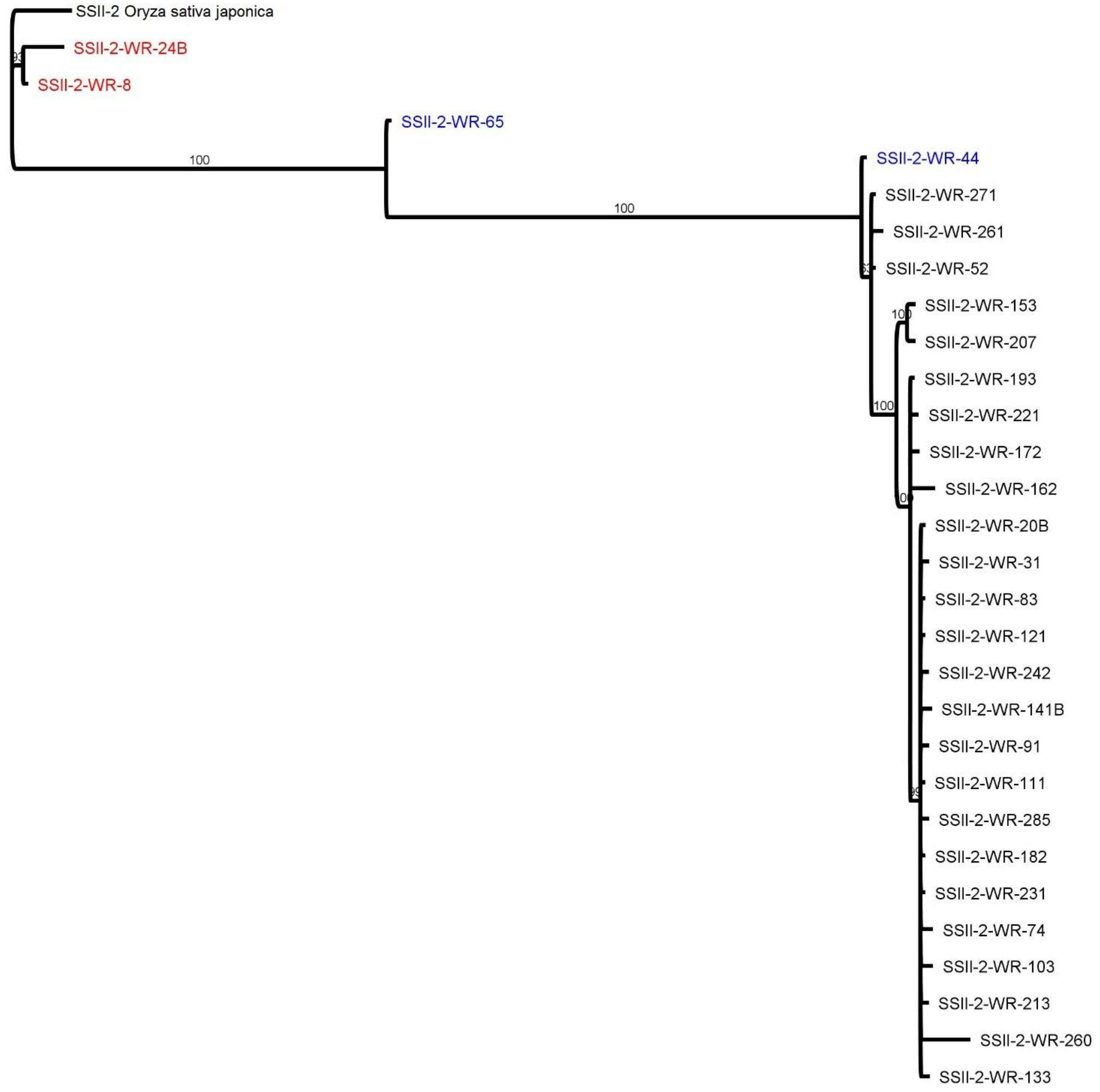


Figure 54 Phylogenetic tree based on Bayesian analysis of *SSII-2* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids.

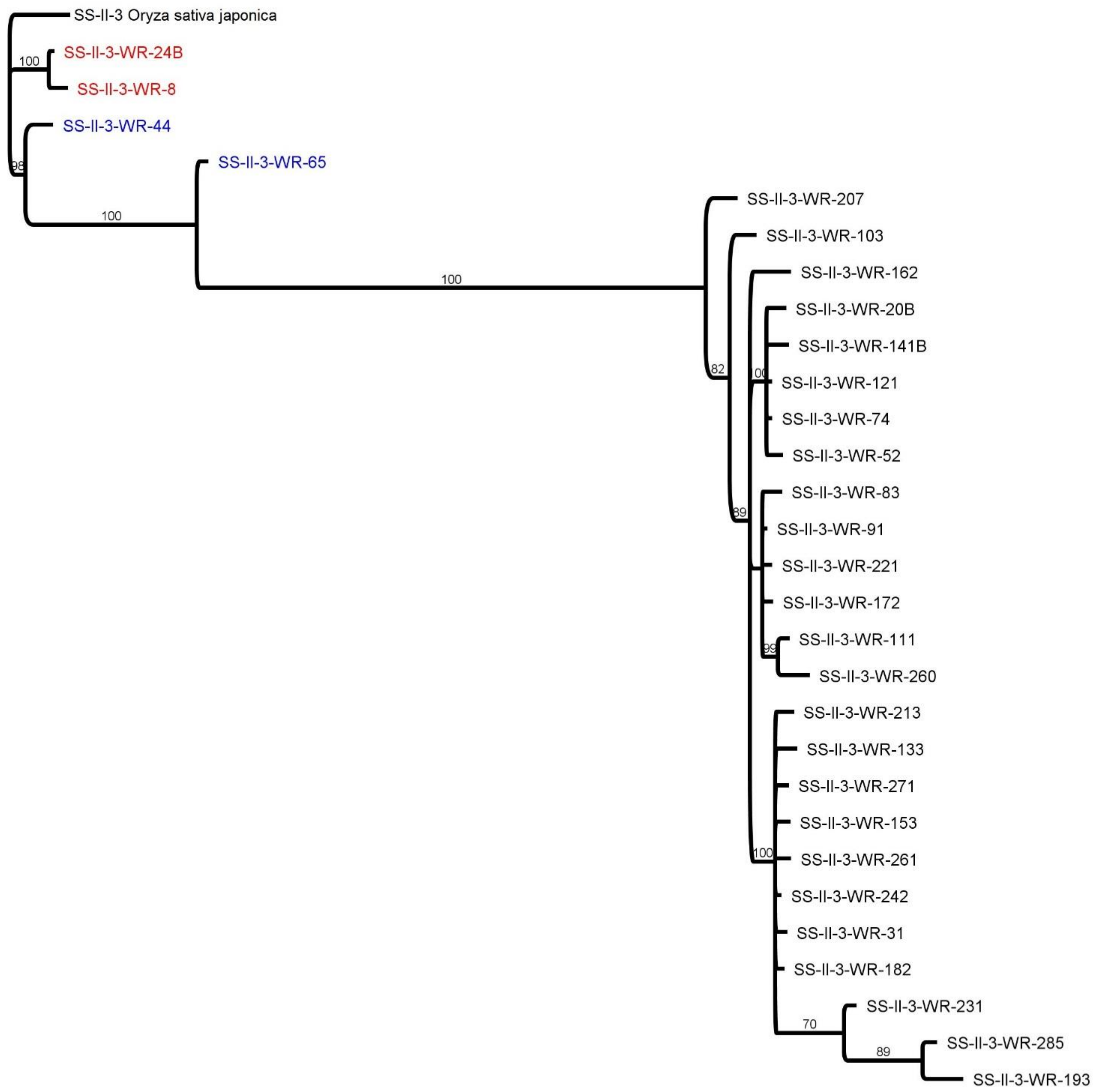


Figure 55 Phylogenetic tree based on Bayesian analysis of *SSII-3* gene. Bootstrap values (1000 replicates) are shown on the branches. Taxa A accessions grouped with domesticated rice while Taxa B accessions grouped together as a separate clade. WR-65 and WR-44 were in between those two clades indicating they were hybrids.

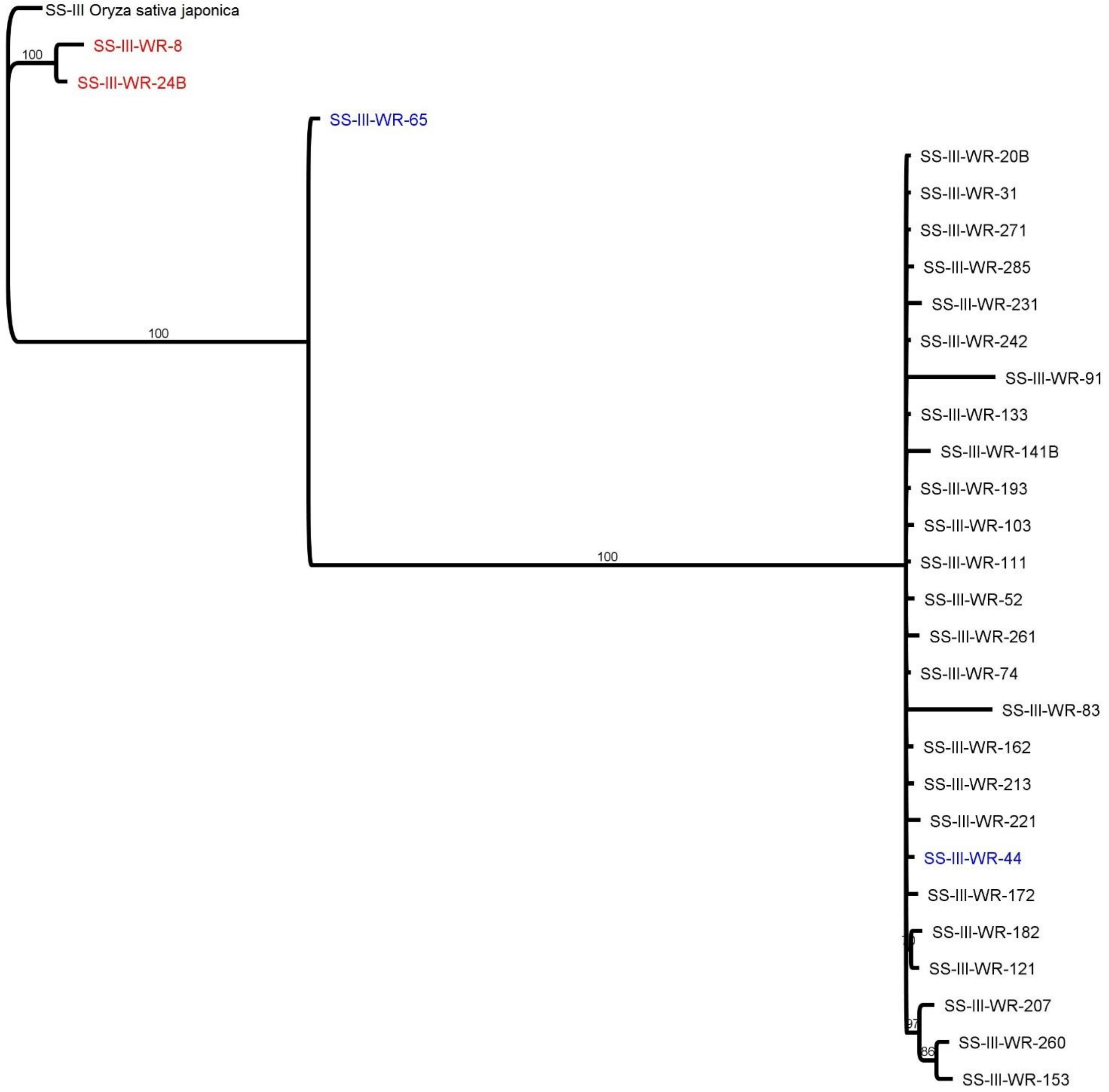


Figure 56 Phylogenetic tree based on Bayesian analysis of *SSIII* gene. Bootstrap values (1000 replicates) are shown on the branches.

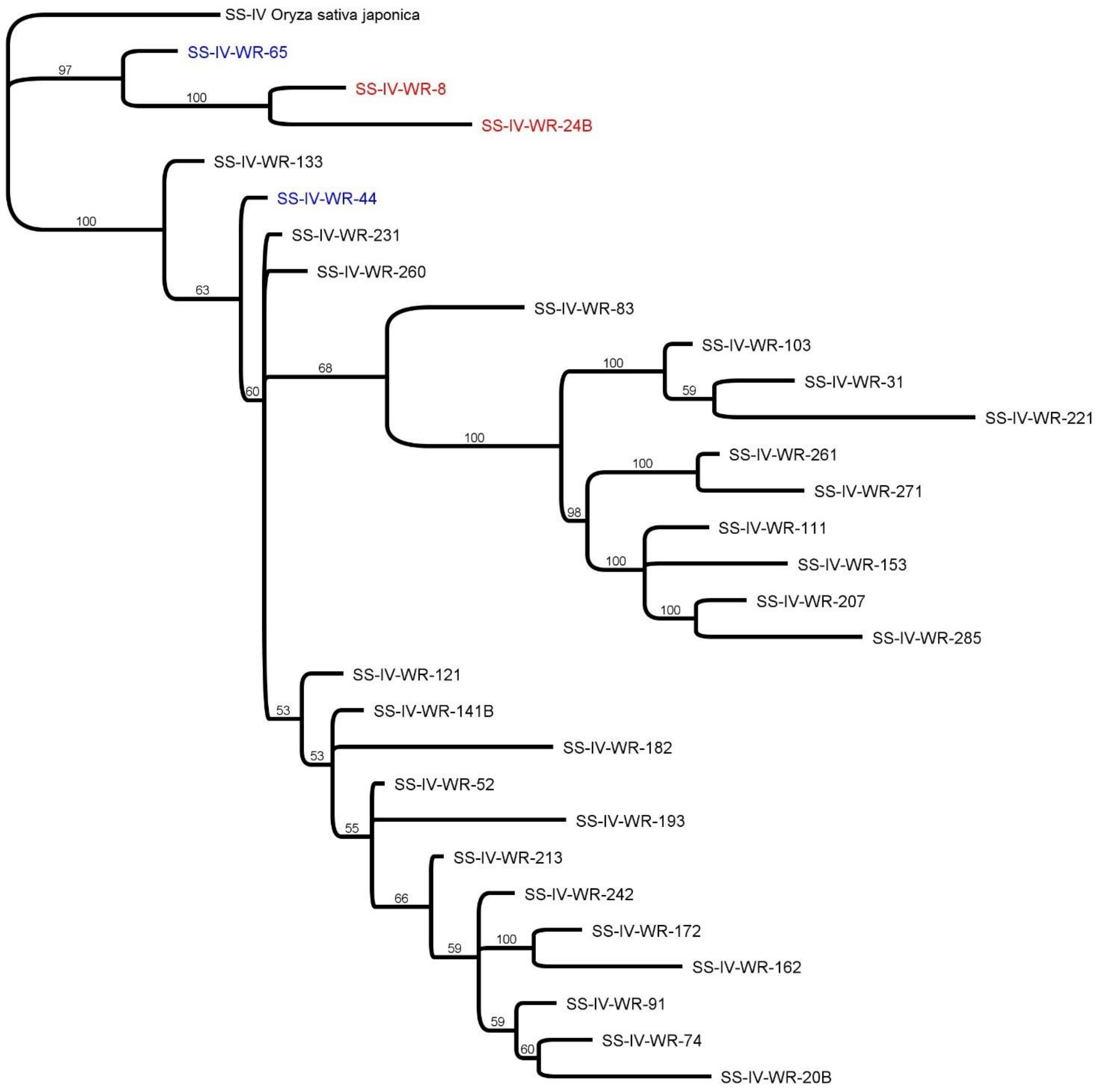


Figure 57 Phylogenetic tree based on Bayesian analysis of *SSIV* gene. Bootstrap values (1000 replicates) are shown on the branches.