Biosecurity New Zealand
Tiakitanga Pūtaiao Aotearoa

MPI 18608 Project Report

Topic 2.1 — Austropuccinia psidii De Novo sequencing

Biosecurity New Zealand Technical Paper No: 2019/39

Prepared for Ministry for Primary Industries
By Chagné D[1], Deng C[2], Wu C[2], Templeton M[2], Smith G[3]

Plant & Food Research: [1]Palmerston North, [2]Mt Albert; [3]Lincoln

ISBN No: 978-1-99-000854-2 (o) (online)
ISSN No: 2624-0203 (online)

June 2019

New Zealand Government

Ministry for Primary Industries
Manatū Ahu Matua

Disclaimer

While every effort has been made to ensure the information in this publication is accurate, the Ministry for Primary Industries does not accept any responsibility or liability for error of fact, omission, interpretation or opinion that may be present, nor for the consequences of any decisions based on this information.

Requests for further copies should be directed to:

Publications Logistics Officer
Ministry for Primary Industries
PO Box 2526
WELLINGTON 6140

Email: brand@mpi.govt.nz
Telephone: 0800 00 83 33
Facsimile: 04-894 0300

This publication is also available on the Ministry for Primary Industries website at http://www.mpi.govt.nz/news-and-resources/publications/

PFR SPTS No. 18040-2.1

Topic 2.1 — *Austropuccinia psidii* De Novo sequencing

Chagné D[1], Deng C[2], Wu C[2], Templeton M[2], Smith G[3]
Plant & Food Research [1]Palmerston North, [2]Mt Albert, [3]Lincoln

June 2019

Contents

1 Executive Summary

As part of consortium with a University of Sydney-led Australian collective we have sequenced and assembled the whole genome of the myrtle rust fungal pathogen *Austropuccinia psidii* using state-of-the-art DNA sequencing technologies. In total, we obtained a genome assembly of 3187 contigs spanning a total of just over 1 billion base pairs. This is the biggest fungal pathogen genome sequenced to date, which created some unique challenges during the bioinformatics analysis. We are attempting to organise these sequences into a set of 14 putative chromosomes. RNA sequence of genes expressed during the early stages of infection was generated and will be used, together with the genome assembly, to identify putative pathogenicity genes.

2 Recommendations

The whole genome assembly of the myrtle rust pathogen lays the foundation for in-depth understanding of how *A. psidii* infects and interacts with its hosts, and to investigate the \population structure of the pathogen. The genome and RNA will be used in studies planned in the Beyond Myrtle Rust programme to investigate if the fungus uses the same molecular mechanism(s) to infect its many susceptible host species.

3 Introduction

How *A. psidii* causes disease is unknown. Sequencing and analysing pathogen genomes has revealed potential mechanisms of pathogenicity that can be targeted by breeding or other technologies. The only publically available *A. psidii* sequence data is of low quality and is not suitable for analysis. New generation technologies for sequencing and bioinformatics genome assembly have provided new opportunities to generate less fragmented and more contiguous genomes, including that for fungal pathogens such as *A. psidii*. The Pacific Biosciences (Pacbio) sequencing technology is a powerful method as it produces long sequence reads that can span long stretches of repetitive DNA, which alternative sequencing technologies using shorter reads cannot achieve. This objective focused on developing a high quality genome assembly for *A. psidii* using a combination of PacBio sequencing and Hi-C technology.

4 The *A. psidii* sequencing consortium

An international consortium was established with the objective of creating a high quality genome sequence assembly of the pandemic strain of *A. psidii*. This consortium comprises the University of Sydney (Tobias, Dong, Park), Australian National University (Schwessinger), DPI Victoria (Tibbits), Queensland Department of Agriculture and Forestry (Shuey) and The New Zealand Institute for Plant and Food Research Limited (Chagné, Smith, Templeton, Deng, Wu). The consortium has allowed the combined trans-Tasman resources and expertise to freely interact to produce sequence data and undertake downstream analysis, resulting in a high quality genome assembly and annotation.

5 Genome sequencing and analysis

Based on flow cytometry and k-mer analysis, the *A. psidii* genome was estimated at a haploid size of around 1000 Mbp, which is the second largest fungal pathogen genome reported to date. The unexpected large size of the *A. psidii* genome posed a significant challenge as assembling a genome of this size requires significantly more raw sequence data, and substantially more computing power to process that data into assembled sequences.

A total of 157 giga-base pairs (Gbp) of raw sequencing PacBio data was produced at University of Sydney. We obtained raw read outputs from two generations of PacBio sequencers at the equivalent of 18x RSII (bax.h5) and 3x Sequel (bam) SMRT cells. Fasta and Arrow files were extracted, and the assembly process initiated with Canu (v1.6) long read assembly software (Figure 1). An initial assembly with this read coverage was completed within 2.5 months using the University of Sydney High Performance Computer (HPC) cluster. Contig numbers were high (22,474) and BUSCO analysis indicated that the contigs contained 80% complete and 6.6% fragmented gene models. A further eight PacBio SMRT cells (PacBio Sequel instrument) were therefore run for greater coverage. The additional sequencing data provided a total of 50–60x coverage of the genome. A second assembly took 4 months to process on the Sydney HPC. A third assembly was started using the Plant & Food Research HPC using different parameters

than the University of Sydney process; however, it was stopped because of the Sydney assembly finishing first and giving an acceptable output. The assembly spans 1.98 Gbp of diploid sequences and 1 Gbp of phased haploid sequences in a total of 3187 contigs with a N50 larger than 520 Kbp. It was tested for completeness using BUSCO (89% complete conserved BUSCOs) and the presence of bacterial DNA sequence contaminations was verified.

This latest assembly, scaffolded in 3187 contigs was frozen as the final version and used for building pseudo-chromosomes using the Hi-C technique. Hi-C libraries were constructed and sequenced for the same isolate. The Hi-C analysis enabled the clustering of the 3187 contigs into 14 pseudo-chromosomes (Figure 2). However, analysis of the telomeric regions indicated that some of the contigs may be mis-oriented within the pseudo-chromosomes and some manual curation is underway.

6 RNA sequencing and gene expression data generation

Gene expression analysis using high-throughput RNA sequencing can give useful information about the mechanisms of infection from plant pathogens. A trial was established with 24 mānuka plants selected from families that had sufficient numbers of resistant and susceptible plants. The plants were inoculated with an *A. psidii* spore suspension or with water for the infection control plants. Samples were taken at two time points (24 h and 48 h) after inoculation and RNA extracted and sequenced at the Australian Genome Research Facility (AGRF). A total of 353 Gbp of data was received in early June from approximately 1.2 billion sequencing reads (Table 1).

7 Glossary of terms

18x RSII (bax.h5)	An output file from the PacBio software
3x Sequel (bam)	An output file from the PacBio software
SMRT	Single-molecule real-time sequencing
Arrow	A variant caller tool that is part of the SMRT analysis software
Fasta	A DNA sequence alignment software package
Contigs	A consensus sequence assembled from overlapping DNA sequences
Hi-C	High-throughput chromosome conformation capture. Software that identifies chromosome structure and interactions in a genome
k-mer	All the possible sub-sequences from a DNA sequence
Telomeric regions	Telomeres are the sequences at the ends (tips) of chromosomes.

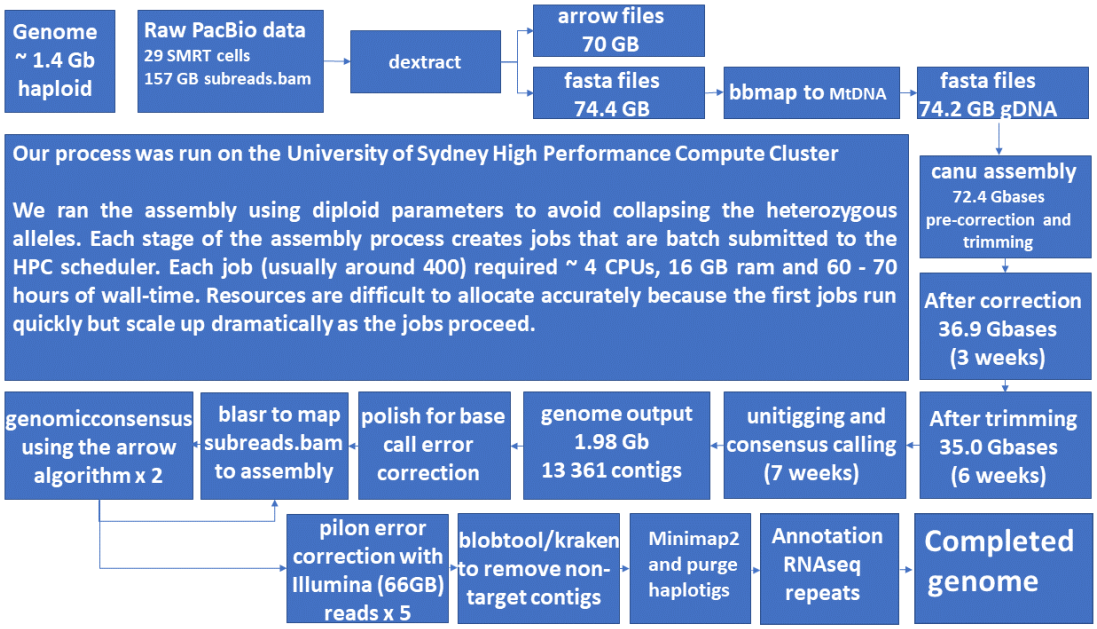

Figure 1. Bioinformatics pipeline used for the *Austropuccinia psidii* genome assembly using PacBio data.

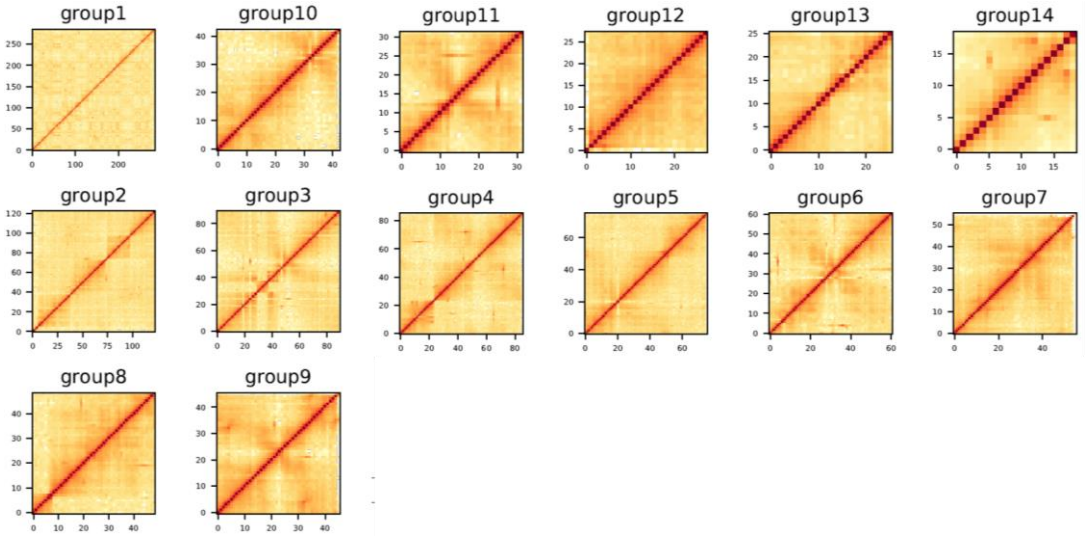

Figure 2. Hi-C analysis of the *Austropuccinia psidii* assembly contigs. The *A. psidii* genome assembly of 3187 contigs was organised into 14 pseudo-chromosomes ranging from 17 Mbp to 250 Mbp.

Table 1. Summary of raw sequencing data from RNA sequencing of mānuka seedlings inoculated with *Austropuccinia psidii*.

RNA library ID	Number of sequencing reads	Data yield (Gbp)
49	25,872,989	7.81
50	24,844,699	7.5
51	31,236,353	9.43
53	25,571,216	7.72
54	24,151,296	7.29
55	24,127,316	7.29
56	24,125,363	7.29
57	28,707,603	8.67
58	25,349,719	7.66
59	27,770,460	8.39
60	24,774,761	7.48
61	24,582,202	7.42
62	25,082,451	7.57
64	25,133,201	7.59
65	28,683,809	8.66
66	22,844,016	6.9
67	24,675,087	7.45
68	23,946,707	7.23
69	26,840,251	8.11
70	24,283,215	7.33
71	30,322,405	9.16
72	23,637,777	7.14
73	26,431,364	7.98
74	28,968,214	8.75
75	26,778,752	8.09
76	26,651,415	8.05
77	23,650,804	7.14
78	26,720,209	8.07
79	24,202,557	7.31
80	26,672,290	8.06
81	27,893,400	8.42
82	24,989,932	7.55
83	22,929,891	6.92
84	29,622,616	8.95
85	24,921,960	7.53
86	26,480,051	8
87	25,795,846	7.79
88	25,551,598	7.72
89	27,337,290	8.26
90	28,405,721	8.58
91	25,301,121	7.64
93	25,155,384	7.6
94	24,518,878	7.4
95	25,979,209	7.85
96	25,700,671	7.76
Total	**1,167,222,069**	**353**

Confidential report for:

Ministry for Primary Industries
Client ref: 18608

PUBLICATION DATA

Chagne D. Topic 2.1 – *Austropuccinia Psidii* De Novo sequencing. June 2019. A Plant & Food Research report prepared for: MPI. Milestone No. 77454. Contract No. 34575 & 35604. Job code: P/313056/01 & P/340203/01. SPTS No. 17807-2.1.

David Chagné
Science Group Leader, Molecular & Digital Breeding
June 2019

Zac Hanley
General Manager, Science — New Cultivar Innovation
June 2019

For further information please contact:

David Chagné
Plant & Food Research Palmerston North
Private Bag 11600
Palmerston North 4442
NEW ZEALAND
Tel: +64 6 953 7700
DDI: +64 6 953 7751
Fax: +64 6 351 7050
Email: David.Chagne@plantandfood.co.nz

This report has been prepared by The New Zealand Institute for Plant and Food Research Limited (Plant & Food Research).
Head Office: 120 Mt Albert Road, Sandringham, Auckland 1025, New Zealand, Tel: +64 9 925 7000, Fax: +64 9 925 7001.
www.plantandfood.co.nz