Draft genomes of two Australian strains of the plant pathogen, *Phytophthora cinnamomi* [version 1; referees: awaiting peer review]

Amy L. Longmuir¹, Peter L. Beech¹, Mark F. Richardson ²,³

¹Centre for Cellular and Molecular Biology, School of Life and Environmental Sciences, Deakin University, Burwood, VIC, 3125, Australia
²Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, VIC, 3220, Australia
³Bioinformatics Core Research Group, School of Life and Environmental Sciences, Deakin University, Geelong, VIC, 3220, Australia

Abstract

**Background:** The oomycete plant pathogen, *Phytophthora cinnamomi*, is responsible for the destruction of thousands of species of native Australian plants, as well as several crops, such as avocado and macadamia, and has one of the widest host-plant ranges of the *Phytophthora* genus. The currently available genome of *P. cinnamomi* is based on an atypical strain and has large gaps in its assembly. To further studies of the pathogenicity of this species, especially in Australia, more robust assemblies of the genomes of more typical strains are required. Here we report the genome sequencing, draft assembly, and preliminary annotation of two geographically separated Australian strains of *P. cinnamomi*.

**Findings:** Some 308 million raw reads were generated for the two strains. Independent genome assembly produced final genomes of 62.8 Mb (in 14,268 scaffolds) and 68.1 Mb (in 10,084 scaffolds), which are comparable in size and contiguity to other *Phytophthora* genomes. Gene prediction yielded > 22,000 predicted protein-encoding genes within each genome, while BUSCO assessment showed 82.5% and 81.8% of the eukaryote universal single-copy orthologs to be present in the assembled genomes, respectively.

**Conclusions:** The assembled genomes of two geographically distant isolates of *Phytophthora cinnamomi* will provide a valuable resource for further comparative analysis and evolutionary studies of this destructive pathogen, and further annotation of the presented genomes may yield possible targets for novel pathogen control methods.

This article is included in the Global Open Data for Agriculture and Nutrition gateway.
Corresponding author: Peter L. Beech (peter.beech@deakin.edu.au)

Author roles: Longmuir AL: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation; Beech PL: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; Richardson MF: Conceptualization, Data Curation, Formal Analysis, Methodology, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Longmuir AL, Beech PL and Richardson MF. Draft genomes of two Australian strains of the plant pathogen, *Phytophthora cinnamomi* [version 1; referees: awaiting peer review] F1000Research 2017, 6:1972 (doi: 10.12688/f1000research.12867.1)

Copyright: © 2017 Longmuir AL et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: The author(s) declared that no grants were involved in supporting this work.

**Introduction**

*Phytophthora cinnamomi* is a highly virulent plant pathogen that has a devastating impact on the Australian ecosystem, namely in the south-western areas of Western Australia and much of the south and east coasts of Victoria and New South Wales. In the south-west Botanical Province of Western Australia, alone, over 40% of the 5710 plant species present have been shown to be susceptible to *P. cinnamomi*. Significant genetic and phenotypic variation can occur within a signal clonal lineage of *P. cinnamomi* and susceptibility of a given host plant species has been shown to vary from site to site. Furthermore, despite the general lack of crossing during sexual reproduction, *P. cinnamomi* excels at adapting to new environments and developing virulence to new host species through asexual growth, making it a deadly and difficult-to-control pathogen. Unravelling how *P. cinnamomi* is able to adapt so quickly, and remain virulent to a wide range of hosts in Australia, is an important research goal.

The currently available genome of *P. cinnamomi* var. *cinnamomi* (Joint Genome Institute (JGI); NCBI Accession no. PRJNA682441) is based on the Rans isolate from Sumatra in 1922, which has been in culture for many decades and may not be representative of the current pathogenic strains present in Australia. Here we report and make available two Australian *P. cinnamomi* genomes, isolated from geographically very separate areas with different available host species. After analyses of genetic differences between these two *P. cinnamomi* genomes, it may be that key genes or gene families under high evolutionary pressure can be identified; this may aid further studies on more effective control of this pathogen.

**Sample collection and sequencing**

Two isolates of *P. cinnamomi* were selected from areas of infection on either side of the Australian continent: one from the Brisbane Ranges in southeastern Australia (DU054, A2 mating type) and the other from southwestern Western Australia (WA94.26, A2 mating type), both Deakin University culture collection. These isolates were selected to represent possible genetic diversity of *P. cinnamomi* in Australia arising from geographic isolation, and possible variation of selective pressures due to different host species. Isolates were maintained on V8 agar (V8A) (50 ml unclarified V8 ‘Original’ Juice (Campbells, Australia), 0.5 g CaCO$_3$ and 7.5 g biological agar per 500 mL of distilled water) at 25°C in darkness, as per 5. Genomic DNA was isolated from hyphae using a DNeasy Plant Mini Kit (Qiagen), following the manufacturers protocol. Illumina TruSeq Nano library preparation and sequencing on an Illumina HiSeq 2500 platform were performed by the Australian Genome Resources Facility (Walter and Eliza Hall Institute, Parkville, Australia) generating ~154 million (2x 150bp) raw reads per isolate. Raw reads are available in the NCBI Short Read Archive (SRA) under the Bioproject Accession: PRJNA413098

**Genome assembly**

Raw sequencing data for each isolate was first pre-processed using Trimmomatic v0.35 with the following parameters: *ILLUMINA-CLIP:TruSeq3-PE-2:fa:2:30:10:4 AVGQUAL:30 MINLEN:36*, to remove Illumina adapters and filter reads based on quality scores (Phred score). Only reads with average Phred $>30$ were retained. To ensure only the desired *P. cinnamomi* genomes were assembled, a second round of pre-processing was conducted to remove potential contaminants. MetaPhlAn v2, was run with default settings and identified the *Paenibacillus* genus as a likely contaminant. Using BBMap v0.35 (BBMap - Bushnell. B.), we mapped the Trimomatic-filtered reads to the closest species match (*Paenibacillus* sp., JDR-2, GenBank accession: GCA_000023585.1), with 2.7% and 2.0% of DU054 and WA94.26 reads mapping, respectively; these *Paenibacillus* reads were subsequently removed. The remaining reads were then mapped using BBMap to the human genome (GRCh38; NCBI accession: GCA_000001405.1), with $<0.5$% (~ 430,000 reads from DU054 and ~ 630,000 from WA94.26) being mapped and subsequently removed from the data set. Thus, the final set of reads (DU054, 149 million reads; WA94.26, 151 million reads) used for the assembly contained highly quality paired-end reads not belonging to either human or bacterial contaminants.

*De novo* contig assembly of the two genomes was conducted independently, using IDBA-UD v1.1.0. IDBA-UD was run using the following parameters: --min_k 20 --max_k 100 --step 20 --min_count 500 --min_support 2 --min_count 3. Briefly, these conducted a multiple K-mer assembly from k = 20 up to k = 100; only assembled contigs above 500 bp and those with a minimum depth coverage ≥3 were kept. As heterogeneous data can increase redundancy in genome assemblies, the IDBA-UD assembled contigs were run through the Redundans pipeline v0.12c with the following parameters: --threads 4 --min_length 500. Redundans uses paired-end mapping data to reduce assembled sequence redundancy and scaffold contigs into longer less fragmented sequences. The final assembled genome of DU054 was 62.80 Mb in 14,269 scaffolds with an N50 of 9,951; the longest scaffold was 1.54 Mb in length (Table 1). For WA94.26, the final genome was 68.07 Mb in length, in 10,085 scaffolds with the largest being 1.54 Mb and an N50 of 20,813. GC content remained consistent, at ~53%, between both isolate genomes across both assemblies and before and after processing with Redundans. The quality, as measured by the above metrics, of the presented genomes is comparable to the previously available *P. cinnamomi* var. *cinnamomi* Rans isolate genome (JGI). The final genome assemblies are available under the NCBI Bioproject Accession: PRJNA413098.

We used the BUSCO (benchmarking universal single-copy orthologs) pipeline v1.22 with the default e-value cutoff of 0.01, to assess the completeness of the assembled genomes and compared the results to the previously available Rans isolate. Utilizing the set of 429 conserved eukaryotic single-copy orthologs (hereafter BUSCOs), the analysis indicated 82.5% and 81.8% BUSCO completeness for DU054 and WA94.26 genomes, respectively. For DU054, 335 complete BUSCOs (including single-copy and duplicated BUSCOs) and 19 fragmented BUSCOs were identified, and 333 complete and 19 fragmented BUSCOs in WA94.26 (Table 2). Overall, we find comparable levels of BUSCO completeness with the Rans isolate, suggesting our two Australian isolate assemblies are as complete references as that currently available.

**Preliminary genome annotation**

In the absence of any available high quality ESTs (expressed sequence tags) or transcriptome (gene expression) data for
we conducted a preliminary protein-coding sequence prediction using GeneMark-ES v4.32, which utilises a self-training algorithm to identify exon, intron and intergenic regions as well as initiation and termination sites. GeneMark-ES was run using the default settings and a database of predicted gene models (i.e., predicted polypeptides) was constructed for DU054 and WA94.26 genomes. An initial 23,414 gene models were identified in DU054 and 22,573 in WA94.26. Of these, 14,735 pairs of predicted gene models appear to be orthologous between the two genomes (reciprocal best-hit Blastp, e value ≤ 1e-5). As a preliminary verification of these gene model builds, we identified orthologous counterparts to eight available Phytophthora genomes with more complete annotations (P. infestans, P. kernoviae, P. lateralis, P. nicotianae, P. parasitica (P1569_v1; Broad Institute), P. ramorum, P. sojae and P. cinnamomi var. cinnamomi). Accordingly, we used OrthoFinder v1.1.10 with default parameters, except we used DIAMOND as the alignment program with the diamond_more_sensitive flag. OrthoFinder first identifies ‘orthogroups’ (an extension of orthologues to include groups of genes descended from a single gene in the last common ancestor of a group of species) and then orthologues between each pair of species in the comparison. OrthoFinder assigned 88.5% (170,769) of the genes found in all the species to 19,089 orthogroups, and of these 50% of all the genes were contained in orthogroups, which had 10 or more genes within them. We found 2,931 orthogroups that contained genes for each of the species, and of these 1,309 orthogroups consisted entirely of single copy genes, see associated data repository. Using these single copy orthogroups gene trees were first constructed then the species tree was inferred using the distance-based implemented by fastme. The resultant species tree (see associated data repository) exhibits strong congruence to the Phytophthora phylogeny recently published by, providing more evidence that the genome assembly and preliminary annotation conducted here is valuable.

### Conclusions

In summary, we present the genome assembly of two geographically separated isolates of Phytophthora cinnamomi from Australia, representing the first genome assembly of an Australian-isolated strain. These high-quality genomes will act as a valuable resource, particularly for the further identification and analysis of protein-encoding genes, which are expressed during plant infection, such as members of the avirulence gene families. These gene families are of specific interest in the development of novel and effective pathogen control mechanisms.
Data availability

Raw reads are available in the NCBI SRA under the Bioproject Accession: PRJNA413098

The final assemblies are available at DDBJ/EMBL/GenBank under the accessions, PDCY00000000 and PDCZ00000000 and under the Bioproject Accession: PRJNA413098.

Supporting data, including OrthoFinder analysis and BUSCO assessment results can be found in the associated data repository: doi, 10.4225/16/59d15a6917a5e4e. Data are available under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0).

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com