Draft genomes of two Australian strains of the plant pathogen, *Phytophthora cinnamomi* [version 1; peer review: 2 approved, 1 approved with reservations]

Amy L. Longmuir¹, Peter L. Beech¹, Mark F. Richardson²,³

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.

Keywords

Phytophthora genome, plant pathogen, Phytophthora cinnamomi
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. PRJNA682411) 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 (VSA) [50 ml unclarified V8 ‘Original’ Juice (Campbells, Australia), 0.5 g CaCO3 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 v0.33 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. MetaPhAn v2, was run with default settings and identified the Paenibacillus genus as a likely contaminant. Using BBMap v0.35 (BBMap – Bashmmell. B), we mapped the Trimmomatic-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_000001405.15), 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 v1.10. IDBA-UD was run using the following parameters: --mink 20 --maxk 100 --step 20 --min_count 500 --min_support 2 --min_coverage 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 v1.2b 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 euarkyotic 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\textsuperscript{12}, P. kernoviae\textsuperscript{13}, P. lateralis\textsuperscript{14}, P. nicotianae\textsuperscript{15}, P. parasitica (P1569_v1; Broad Institute), P. ramorum\textsuperscript{16}, P. sojae\textsuperscript{16} and P. cinnamomi var. cinnamomi). Accordingly, we used OrthoFinder v1.1.10\textsuperscript{17} with default parameters, except we used DIAMOND\textsuperscript{18} as the alignment program with the diamond._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\textsuperscript{17}) and then orthologues between each pair of species in the comparison. OrthoFinder assigned 88.5% (170,769) of the genes found in all the genomes 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\textsuperscript{19}. Using these single copy orthogroups gene trees were first constructed then the species tree was inferred using the distance-based implemented by fastme\textsuperscript{20}. The resultant species tree (see associated data repository\textsuperscript{19}) exhibits strong congruence to the Phytophthora phylogeny recently published by\textsuperscript{21}, 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\textsuperscript{22}. These gene families are of specific interest in the development of novel and effective pathogen control mechanisms.

<table>
<thead>
<tr>
<th>Total BUSCOs</th>
<th>DU054</th>
<th>WA94.26</th>
<th>P. cinnamomi var. cinnamomi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete and single copy BUSCOs</td>
<td>262 (61.07%)</td>
<td>249 (58.04%)</td>
<td>281 (65.50%)</td>
</tr>
<tr>
<td>Complete and duplicate BUSCOs</td>
<td>73 (17.01%)</td>
<td>84 (19.58%)</td>
<td>49 (11.42%)</td>
</tr>
<tr>
<td>Fragmented BUSCOs</td>
<td>19 (4.42%)</td>
<td>18 (4.19%)</td>
<td>17 (3.96%)</td>
</tr>
<tr>
<td>Missing BUSCOs</td>
<td>75 (17.48%)</td>
<td>78 (18.18%)</td>
<td>82 (19.11%)</td>
</tr>
</tbody>
</table>

**Table 1. Summary of genomic features of assembled genomes comparing IDBA-UD output to scaffolded genome after Redundans processing and the P. cinnamomi Rans isolate genome.**

<table>
<thead>
<tr>
<th>Assembly size (Mb)</th>
<th>DU054</th>
<th>WA94.26</th>
<th>IDBA-UD</th>
<th>Redundans</th>
<th>IDBA-UD</th>
<th>Redundans</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. scaffolds</td>
<td>33,475</td>
<td>14,268</td>
<td>36,333</td>
<td>10,084</td>
<td>4,085</td>
<td>9,951</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>4,085</td>
<td>9,951</td>
<td>4,075</td>
<td>20,813</td>
<td>23,414</td>
<td>22,573</td>
</tr>
<tr>
<td>No. predicted genes</td>
<td>NA</td>
<td>23,414</td>
<td>NA</td>
<td>22,573</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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, PDCY0000000 and PDCZ0000000 and under the Bioproject Accession: PRJNA413098.

Supporting data, including OrthoFinder analysis and BUSCO assessment results can be found in the associated data repository: doi, 10.4225/16/59d15a6917a5e16. 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


David J. Studholme
Biosciences, University of Exeter, Exeter, UK

This manuscript announces the availability of genomic sequence data from Phytophthora cinnamomi strains DU054 and WA94.26. This is a useful resource for researchers interested in this important pathogen. The authors have deposited and made available the raw sequence data in SRA and their assemblies in GenBank, which is commendable. The genome annotation and protein sequences do not appear to be deposited in GenBank, however. This does not preclude publication, but when discussing predicted genes in the manuscript, the authors should be up-front about this or provide full details of the annotations in supplementary data or deposit them in some public repository.

One oversight that the authors should be aware of is the previous publication of two genome sequences of this species, one of which (MP94-48) is from Australia. See Studholme et al. (2015)1. So, the authors' assertion (in their Conclusions section) that this is the first genome assembly from an Australian strain should be revised. The authors should also include those two assemblies in their comparisons of assembly quality metrics. And also please revise the several other mentions of previously sequenced genome throughout the text in the light of the additional two previously sequenced genomes. Also, it would be interesting to assess how similar or different all these four available P.c. genome sequences are to each other, e.g. by calculating pairwise ANIs.

Some specific points that should be addressed around the methodology:

1. Why were reads mapped against the human genome? Why should contamination from human DNA be more prevalent or likely than from other organisms?

2. The authors make good efforts to remove contaminating Paenibacillus sequence reads. Interestingly, we also observed contamination of Phytophthora genomic DNA with this bacterial genus. However, the authors go on to claim that the data contained "highly quality reads not belonging to ... bacterial contaminants". Their approach does not remove non-Paenibacillus bacterial contaminants.

3. Please cite a reference to support the claim that "heterogeneous data can increase
redundancy in genome assemblies". It is not entirely clear what this statement means, precisely, and in any case it is not self-evident and needs to be supported by peer-reviewed publication.

4. The use of BUSCO version 1.22 is questionable, given that versions 2 and 3 are now available. Furthermore, rather than using the general Eukaryote set of BUSCOs, the authors should use the Stramenopile set.

5. The completeness of the genome assemblies is rather poor (only < 65% of expected genes are present intact in a single copy). It would be useful to compare/benchmark this against other available Phytophthora genome sequences. For example, our recent sequencing of P. ramorum genomes, we found around 81-85% of Stramenopile BUSCOs were intact and single-copy in each genome (See PubMed ID 28243575).

6. Towards the end of page 4, the authors claim that the "preliminary annotation ... is valuable". I agree and would go further to say that not just the annotation but the genome sequencing per se is valuable. I would also suggest including a brief explanation of how/why the presented data is valuable.

7. The authors say that their annotation is valuable, but the annotation has not apparently been deposited in a public repository. Therefore, please either make this valuable resource available, or remove the claim that it is valuable.

Some very minor points:
1. In the Introduction, it was not obvious to me what is meant by a "Botanical Province". Please consider explaining this term.

2. Please add an apostrophe after "manufacturers".

3. At several places in the text, the authors write "parameters" when they really mean "parameter values" or "options" or "switches". Please check and revise.

4. Please write "high-quality" not "highly quality".

5. On page 3, the authors say that no gene expression data are available for this species. This is untrue, since EST data (i.e. expressed sequence tags) are available. Furthermore, in the SRA, there are several RNAseq datasets available:

   **Illumina HiSeq 2000 sequencing:**
   - RNAseq analysis of germinating cysts of Phytophthora cinnamomi 1 ILLUMINA (Illumina HiSeq 2000) run: 46,420 spots, 4.2M bases, 3.5Mb downloads Accession: ERX709652 Select item 14623972.

   **Illumina HiSeq 2000 paired end sequencing**
   - 1 ILLUMINA (Illumina HiSeq 2000) run: 9.9M spots, 1.8G bases, 1.1Gb downloads Accession: ERX943317 Select item 1426113.

   **Phytophthora cinnamomi library**
   - 1 ILLUMINA (Illumina HiSeq 2000) run: 88.1M spots, 17.6G bases, 10.3Gb downloads Accession: SRX124562 Select item 1426104.
Phytophthora cinnamomi library 1 ILLUMINA (Illumina HiSeq 2000) run: 30,453 spots, 6.1M bases, 2.6Mb downloads Accession: SRX124561 Select item 1426095.

Phytophthora cinnamomi library 1 ILLUMINA (Illumina HiSeq 2000) run: 50.6M spots, 10.1G bases, 5.9Gb downloads Accession: S RX124560 Select item 1426086.

Phytophthora cinnamomi library 2 ILLUMINA (Illumina HiSeq 2000) runs: 38.5M spots, 7.7G bases, 4.5Gb downloads Accession: SRX124559

6. When quoting N50 values, please include the units. For example, the N50 for DU054 was 9,951 bp or nt.

7. The authors refer to (on page 4) "more complete annotations" of several species. Among these examples is P. lateralis and a citation of our paper (PubMed 23678994) about the sequencing of this species' genome; however, I would not agree that its annotation is "more complete".

8. On page 3, second paragraph, the authors write "the available genome". It is not the "genome" that is available; rather it is the "genome sequence". Once the authors have addressed all these issues, I would be very pleased to see this indexed.

References

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of
expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Mark Richardson, Deakin University, Geelong, Australia

Thank you very much for providing a thorough review and pointing out several oversights we made. We have endeavoured to rectify these as you will see from our responses below. Importantly, we have included the preliminary gene prediction results in the associated public repository with the current supplementary materials. We have also revised the manuscript to include the additional genomes from Studholme et al. 2015 and included them in a comparative BUSCO completeness assessment. We feel a more comprehensive comparative analysis (including ANIs) is beyond the scope of this research note, but this will be part of a future paper.

For clarity, the below responses are separated into Major and Minor subheadings and numbered as per points in the review in order to avoid duplication of text.

Major:

1. Contamination for human DNA should not be more prevalent than any other. As this was one of the first times we cultured this species we carried out this pre-filtering to ascertain whether or not we had any inadvertent contamination. The results show this was not the case.

2. We find it interesting that the reviewer has also detected Paenibacillus contamination during their work. While removing contamination through mapping to the Paenibacillus genome alone would not warrant our statement, this is not what we did. We used MetaPHIAn to first screen our raw reads to identify which, if any, bacterial species might be present. Only Paenibacillus could be detected. Thereby, once removed, we are confident that no other bacterial contamination exists. If others had been identified with MetaPHIA then they could be removed in the same way.

3. We have added a citation to this extent and clarified what we mean in the text.

4. We have repeated this analysis with version 3.02 and used the suggested ortholog set.

5. With respect, we feel that the reviewer’s statement that the completeness was poor is unfounded, especially if we consider that they are not making a ‘like for like’ comparison by comparing results from the eukaryotic set to those from the stramenopile set. Nevertheless, the updated BUSCO analysis using the stramenopile set reveals the genome assemblies presented here have BUSCO completeness of ~91 to 94 %, which falls within the range for the previous P. cinnamomi assemblies (86 -97% completeness, see Table 2 for full comparison).

6. Thank you for this suggestion, we have done so.
7. This is a very valid point. We have now included the preliminary gene predictions with the supplementary data.

Minor:
1. We have changed this to the more commonly understood ‘ecoregion’
2. Done
3. Done
4. Done
5. We have removed this statement.
6. Done
7. We have removed this statement.
8. Addressed

**Competing Interests:** No competing interests were disclosed.
Yes

Are sufficient details of methods and analysis provided to allow replication by others?  
Partly

If applicable, is the statistical analysis and its interpretation appropriate?  
Yes

Are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions drawn adequately supported by the results?  
Yes

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 28 Feb 2018

Mark Richardson, Deakin University, Geelong, Australia

Thank you for the positive review, we have added the additional information you have requested.

Competing Interests: No competing interests were disclosed.

Reviewer Report 17 November 2017

https://doi.org/10.5256/f1000research.13945.r27739

© 2017 Ayub N. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nicolás Daniel Ayub

National Scientific and Technical Research Council, Buenos Aires, Argentina

The work was carried out professionally and resulted in good draft genomes of two pathogen strains belonging to Phytophthora genus. In my opinion, this article is an important contribution to future studies about the molecular mechanism involved in Phytophthora-plant interaction. Particularly, in the first steps of pathogen adhesion, where the virulence factors related to this are little known.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 28 Feb 2018
Mark Richardson, Deakin University, Geelong, Australia

Thank you for the positive review.

Competing Interests: No competing interests were disclosed.
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