DATA NOTE

Data on whole genome sequencing of the oomycete Pythium insidiosum strain CBS 101555 from a horse with pythiosis in Brazil

Theerapong Krajaejun1*, Weerayuth Kittichotirat2*, Preecha Patumcharoenpol1, Thidarat Rujirawat3, Tassanee Lohnoo3 and Wanta Yingyong3

Abstract

Objectives: The oomycete Pythium insidiosum infects humans and animals worldwide, and causes the life-threatening condition, called pythiosis. Most patients lose infected organs or die from the disease. Comparative genomic analyses of different P. insidiosum strains could provide new insights into its pathobiology, and can lead to discovery of an effective treatment method. Several draft genomes of P. insidiosum are publicly available: three from Asia (Thailand), and one each from North (the United States) and Central (Costa Rica) Americas. We report another draft genome of P. insidiosum isolated from South America (Brazil), to serve as a resource for comprehensive genomic studies.

Data description: In this study, we report genome sequence of the P. insidiosum strain CBS 101555, isolated from a horse with pythiosis in Brazil. One paired-end (180-bp insert) library of processed genomic DNA was prepared for Illumina HiSeq 2500-based sequencing. Assembly of raw reads provided genome size of 48.9 Mb, comprising 60,602 contigs. A total of 23,254 genes were predicted and classified into 18,305 homologous gene clusters. Compared with the reference genome (the P. insidiosum strain Pi-S), 1,475,337 sequence variants (SNPs and INDELs) were identified in the organism. The genome sequence data has been deposited in DDBJ under the accession numbers BCFP01000001–BCFP01060602.

Keywords: Pythium insidiosum, Pythiosis, Oomycete, Genome, Gene cluster, Sequence variant

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
*P. insidiosum* using the next generation sequencing technologies. Comparative genomic analyses of different *P. insidiosum* strains could provide new insights into its biological processes and pathogenesis, which can lead to discovery of a novel method for pathogen control. Five draft genomes of *P. insidiosum* are deposited in the public repositories: three from Asia (Thailand; Clade-II and -III strains), and one each from North (the United States; Clade-I strain) and Central (Costa Rica; Clade-I strain) Americas [8–11]. Here, we report another draft genome data of *P. insidiosum* (Clade-I) isolated from South America (Brazil), as opposed to the other 5 strains (with published genome sequences) isolated from other regions of the world, to serve as a resource for comprehensive genomic studies in the future.

**Data description**

The *P. insidiosum* strain CBS 101555, isolated from a granulomatous lesion at the abdomen of a horse with pythiosis living in the southern region of Brazil, was cultured in Sabouraud dextrose broth at 37 °C for 1 week. Hyphal mat was harvested from the culture, cultured in Sabouraud dextrose broth at 37 °C for 1 week. Hyphal mat was harvested from the culture, subjected to genomic deoxyribonucleic acid (DNA) extraction, using the conventional extraction method, optimized for *P. insidiosum* [12]. The identity of the strain was checked by single nucleotide polymorphism-based multiplex PCR and sequence homology analysis of the rDNA sequence (Accession number: AB971181) [13, 14]. The obtained genomic DNA was sequenced, using the Illumina next generation sequencing platform, as previously-described [8–10]. Briefly, the genomic DNA was processed to prepare a paired-end (180-bp insert) library for Illumina HiSeq 2500-based sequencing (Yourgene Bioscience, Taiwan). To guarantee read lengths of at least 35 bases, obtained raw reads underwent quality trims including single-nucleotide polymorphisms (SNPs) and insertion/deletion of bases (INDELs), in the genome of *P. insidiosum* [12]. The identity of the strain was checked by single nucleotide polymorphism-based multiplex PCR and sequence homology analysis of the rDNA sequence (Accession number: AB971181) [13, 14].

The obtained genomic DNA was sequenced, using the Illumina next generation sequencing platform, as previously-described [8–10]. Briefly, the genomic DNA was processed to prepare a paired-end (180-bp insert) library for Illumina HiSeq 2500-based sequencing (Yourgene Bioscience, Taiwan). To guarantee read lengths of at least 35 bases, obtained raw reads underwent quality trims by CLC Genomics Workbench (Qiagen). The Cutadapt 1.8.1 [15] was used to remove the adaptor sequences. The resulting genome data contained 34,617,696 raw reads with an average length of 122 bases, providing an average contig length of 806 bases (range 300–30,744), N50 of 953 bases, and ‘N’ composition of 0.9%. The draft assembled genome size of the organism was 48,855,945 bases. MAKER2 [17] predicted 23,254 genes in the draft genome. Basic Local Alignment Search Tool (BLAST) was used to annotate predicted genes by comparing to the NCBI non-redundant protein database using E-value cut off 10\(^{-6}\). Product description of the best blast hit was used as the product description of the query gene. The genome sequence data has been deposited in the DNA Data Bank of Japan (DDBJ) under the Accession numbers BCFP01000001–BCFP01060602 (Data file 1; Table 1).

The 23,254 predicted genes can be classified into 18,305 homologous gene clusters (Data file 2; Table 1), using the method described by Kittichotirat et al. [18] and Rujirawat et al. [19], and the following setting: BLAST E-value of 10\(^{-6}\), pairwise sequence identity of at least 30%, and pairwise alignment coverage for both query and subject sequences of at least 50%. Based on the BLAST search with E-value cut-off of 10\(^{-6}\) against the Clusters of Orthologous Groups of Proteins (COGs) database [20, 21], 3288 gene clusters (18%) were assigned to 24 COGs groups, while the rest (15,017 gene clusters [82%]; designated as uncharacterized cluster) did not match any COGs. Details on percentages and frequency of each assigned COGs group were shown in Data file 3 (Table 1).

The obtained draft genome was analysed for sequence variants, by using the Burrows–Wheeler Alignment tool [22]. Approximately, 44% of the processed reads (n=15,084,792) of the *P. insidiosum* strain CBS 101555 can map the reference genome of the *P. insidiosum* strain Pi-S (the genome size of 53,239,050 bases, comprising 1192 contigs; Accession number BBXB00000000.1) [10]. FreeBayes [23] can identify 1,475,337 sequence variants, including single-nucleotide polymorphisms (SNPs) and insertion/deletion of bases (INDELs), in the genome of the organism (Data file 4; Table 1).

In conclusion, *P. insidiosum* is an understudied pathogen that causes the life-threatening condition, called

**Table 1 Overview of data files/data sets**

| Label | Name of data file/data set | File types (file extension) | Data repository and identifier (DOI or accession number) |
|-------|-----------------------------|-----------------------------|--------------------------------------------------------|
| Data file 1 | Whole genome sequence | FASTA | DDBJ (Accessions numbers: BCFP01000001–BCFP01060602) (http://getentry.ddbj.nig.ac.jp/) |
| Data file 2 | Gene clusters | MS Excel file (.xlsx) | Mendeley database (https://doi.org/10.17632/jyjyx5gk7s1.1) (https://data.mendeley.com/datasets/jyjyx5gk7s1/1) |
| Data file 3 | Clusters of Orthologous Groups of Proteins (COGs) | MS Excel file (.xlsx) | Mendeley database (https://doi.org/10.17632/5rhfd4n37k1.1) (https://data.mendeley.com/datasets/5rhfd4n37k1/1) |
| Data file 4 | Sequence variants | MS Excel file (.xlsx) | Mendeley database (https://doi.org/10.17632/4y8hdw7tb71.1) (https://data.mendeley.com/datasets/4y8hdw7tb71/1) |
pythiosis, in humans and animals worldwide. We sequenced the draft genome of the *P. insidiosum* strain CBS 101555, isolated from a pythiosis horse living in the southern region of Brazil. The obtained genome will be a fundamental resource for exploring biology and pathogenesis of this invasive microorganism.

**Limitations**
The draft genome was obtained from short-read assembly of one Illumina-based paired-end (180-bp insert) library, without any mate pair library, resulting in as many as 60,602 contigs. The estimated genomic coverage is limited to ~ 87-fold. The mitochondrial genome sequences were not excluded from the nuclear genome assembly.

**Abbreviations**
DNA: deoxyribonucleic acid; DDBJ: DNA Data Bank of Japan; BLAST: Basic Local Alignment Search Tool; COGs: Clusters of Orthologous Groups of Proteins; SNP: single-nucleotide polymorphism; INDEL: insertion or deletion of bases.

**Authors' contributions**
WK and TK conceived the project. WK, PP, TR, TL, and WY performed the experiments. WK, PP, TR, and TK analyzed the data. WK and TK wrote the manuscript. WK and TK acquired the research funds. All authors read and approved the final manuscript.

**Author details**
1 Department of Pathology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. 2 Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkhunten, Bangkok, Thailand. 3 Research Center, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand.

**Acknowledgements**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Availability of data and materials**
The genome data described in this study can be accessed online at DDBJ (the draft genome sequence; Accession numbers BCFP01000001–BCFP01060602) and Mendeley database (i.e., gene clusters [https://doi.org/10.17632/yjyzx5gk7s.1], COGs [https://doi.org/10.17632/5rhfd4n37k.1], and sequence variants [https://doi.org/10.17632/4y8hdw7tb7.1]).

**Consent for publication**
Not applicable.

**Ethics approval and consent to participate**
This study is a part of the research project that has been approved by the Committee on Human Rights Related to Research Involving Human Subjects, at the Faculty of Medicine, Ramathibodi Hospital, Mahidol University (Protocol Number: ID 05-60-77).

**Funding**
This work was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University [Grant numbers CF_60001 and CF_61007] and the Thailand Research Fund [Grant Number BRG5980009]. The authors acknowledge the financial support provided by King Mongkut’s University of Technology Thonburi through the “KMITT 55th Anniversary Commemorative Fund”. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Publisher's Note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**References**
1. Gastastra W, Lipman LJ, De Cock AWAM, Exel TK, Pegge RBG, Scheurwater J, et al. *Pythium insidiosum*: an overview. Vet Microbiol. 2010;146:1–16.
2. Mendoza L, Hernandez F, Ajello L. Life cycle of the human and animal oomycete pathogen *Pythium insidiosum*. J Clin Microbiol. 1993;31:2967–73.
3. Supabandhu J, Fisher MC, Mendoza L, Vanitanakom N. Isolation and identification of the human pathogen *Pythium insidiosum* from environmental samples collected in Thai agricultural areas. Med Mycol. 2008;46:41–52.
4. Presser JW, Goss EM. Environmental sampling reveals that *Pythium insidiosum* is ubiquitous and genetically diverse in North Central Florida. Med Mycol. 2015;53:674–83.
5. Miller RI. Investigations into the biology of three “phycomycotic” agents pathogenic for horses in Australia. Mycopathologia. 1983;81:23–8.
6. Zambrano CG, Fonseca ACS, Valente JS5, Braga GC, Saafs EV, Azevedo MI, et al. Isolation and characterization of *Pythium* species from swampy areas in the Rio Grande do Sul, Brazil, and evaluation of pathogenicity in an experimental model. Pesquisa Veterinária Brasileira. 2017;37:459–64.
7. Krajaejun T, Sathapatayavongs B, Pracharktam R, Nitiyanant P, Leelachaikul P, Wanachiwawanin W, et al. Clinical and epidemiological analyses of human pythiosis in Thailand. Clin Infect Dis. 2006;43:569–76.
8. Kittichotirat W, Patumcharoenpol P, Rujirawat T, Lohnoo T, Yingyong W, Krajaejun T. Draft genome and sequence variant data of the oomycete *Pythium insidiosum* strain Pi45 from the phylogenetically-distinct Clade-III. Data in Brief. 2017;15:896–900.
9. Patumcharoenpol P, Rujirawat T, Lohnoo T, Yingyong W, Vanitanakom N, Kittichotirat W, et al. Draft genome sequences of the oomycete *Pythium insidiosum* strain CBS 573.85 from a horse with pythiosis and strain CR02 from the environment. Data in Brief. 2018;16:47–50.
10. Rujirawat T, Patumcharoenpol P, Rujirawat T, Yingyong W, Lerkuthirit T, Tangphatsorrungroj S, et al. Draft genome sequence of the pathogenic oomycete *pythium insidiosum* strain PI–S, isolated from a patient with human pythiosis. Genome Announc. 2015;3:e00574.
11. Ascunce MS, Huguet-Tapia JC, Braun EL, Ortiz-Urquiza A, Keyhani NO, Goss EM. Whole genome sequence of the emerging oomycete pathogen *Pythium insidiosum* strain CDC-B5653 isolated from an infected human in the USA. Genom Data. 2016;7:60–1.
12. Lohnoo T, Jongrui N, Rujirawat T, Yingyong W, Lerkuthirit T, Nampon U, et al. Efficiency comparison of three methods for extracting genomic DNA of the pathogenic oomycete *Pythium insidiosum*. J Med Assoc Thai. 2014;97:342–8.
13. Rujirawat T, Sridapan T, Lohnoo T, Yingyong W, Kumsang Y, Sae-Chew P, et al. Single nucleotide polymorphism-based multiplex PCR for identification and genotyping of the oomycete *Pythium insidiosum* from humans, animals and the environment. Infect Genet Evol. 2017;54:429–36.
14. Chasprasert A, Krajaejun T, Pannamunosorn S, Pranyachatigul C, Wanachiwawanin W, Sathapatayavongs B, et al. *Pythium insidiosum* Thai isolates: molecular phylogenetic analysis. Asian Biomed. 2009;3:623–33.
15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBO Rep. 2011;12:1491.
16. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
17. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-data-base management tool for second-generation genome projects. BMC Bioinform. 2011;12:491.
18. Kittichotirat W, Bumgarner RE, Asikainen S, Chen C. Identification of the pangenome and its components in 14 distinct *Aggregatibacter actino-mycetemcomitans* strains by comparative genomic analysis. PLoS ONE. 2011;6:e22420.
19. Rujirawat T, Patumcharoenpol P, Lohnoo T, Yingyong W, Kumsang Y, Payattikul P, et al. Probing the phylogenomics and putative pathogenicity genes of *Pythium insidiosum* by oomycete genome analyses. Sci Rep. 2018;8:4135.

20. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinform. 2003;4:41.

21. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43(Database issue):D261–9.

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

23. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint. 2012. arXiv:1207.3907[q-bio.GN].