THAPBI PICT – a metabarcoding analysis pipeline developed as a *Phytophthora* ITS1 Classification Tool

Peter Cock[a], David Cooke[b], Leighton Pritchard[c]

Bioinformatics Community Conference 2020

Repository: https://github.com/peterjc/thapbi-pict/
Documentation: https://thapbi-pict.readthedocs.io/
License: MIT

Molecular barcodes are often used for environmental monitoring to identify species present in a sample, using PCR primers to amplify a diagnostic genome-region of the organisms of interest. We are interested in metabarcoding where multiple samples are multiplexed for high-throughput sequencing on the Illumina platform using overlapping paired end reads, which imposes an upper limit on the length of the expected PCR amplification product. Each sample should yield a collection of marker sequences reflecting the community composition, and matching these to a database of known species can give a taxonomic breakdown.

THAPBI PICT is a metabarcoding tool developed during the UK funded Tree Health and Plant Biosecurity Initiative (THAPBI) Phyto-Threats project, which focused on identifying *Phytophthora* species in commercial tree nurseries. *Phytophthora* (from Greek meaning plant-destroyer) are economically important plant pathogens, important in both agriculture and forestry. This project used an ITS1 (Internal Transcribed Spacer one) marker, a region of eukaryotes genomes between the 18S and 5.8S rRNA genes, with nested primers to target *Phytophthora*. With appropriate primer settings and a custom database, THAPBI PICT can be applied to other organisms and/or barcode marker sequences - making it more than just a *Phytophthora* ITS1 Classification Tool (PICT).

The analysis pipeline starts from demultiplexed paired FASTQ files, as produced by the Illumina MiSeq platform. These are quality trimmed, overlapping reads merged and primer trimmed (calling external tools) and then deduplicated giving a much smaller list of unique sequences and associated read counts passing a minimum threshold. These are matched to a curated database using a choice of methods, producing both plain text and formatted Excel report. An edit graph in XGMML format is also produced for display in Cytoscape.

THAPBI PICT is released as open source software under the MIT licence. It is written in Python, a free open source language available on all major operating systems. Version control using git hosted publicly on GitHub is used for the source code, documentation, and database builds including tracking the hand curated reference set of *Phytophthora* ITS1 sequences. Continuous integration of the test suite is currently run on both TravisCI and CircleCI. Software releases are to the Python Packaging Index (PyPI) as standard for the Python ecosystem, and additionally packaged for Conda via the BioConda channel. This offers simple installation the tool itself and all the command line dependencies on Linux or macOS. The documentation is currently hosted on Read The Docs, updated automatically from the GitHub repository.

[a]Information and Computational Sciences, James Hutton Institute, Invergowrie, Dundee, UK
[b]Cell and Molecular Sciences, James Hutton Institute, Invergowrie, Dundee, UK
[c]Strathclyde Institute of Pharmacy & Biomedical Sciences, Glasgow, UK