DATA NOTE

Updated mtCOI reference dataset for the *Bemisia tabaci* species complex [version 1; referees: 2 approved]

Laura M. Boykin¹, Anders Savill¹, Paul De Barro²

¹School of Molecular Sciences and Australian Research Council Centre of Excellence in Plant Energy Biology, University of Western Australia, Crawley, Perth, WA, Australia
²CSIRO, Ecosciences Precinct, Brisbane, QLD, Australia

Abstract

Members of the whitefly *Bemisia tabaci* species complex cause millions of dollars of damage globally and are considered one of the world’s most invasive species. They are capable of causing extensive damage to major vegetable, grain legume and fiber crops. All member of the species complex are morphologically identical therefore, data from the partial mitochondrial cytochrome oxidase subunit I (mtCOI) gene sequence has been used to identify the various species. The current reference dataset that is widely used is found on the CSIRO data portal. However, the reference set stored on the CSIRO data does not include newly added sequences (2013-2017), therefore an updated reference dataset is needed. All mtCOI data for the *Bemisia tabaci* species complex were downloaded on 22 May 2017 from GenBank and after quality checking, a dataset of 1,071 unique sequences and 696 base pairs was generated (https://doi.org/10.6084/m9.figshare.5437420.v1).

This article is included in the Global Open Data for Agriculture and Nutrition gateway.
Corresponding author: Laura M. Boykin (laura.boykin@uwa.edu.au)

Author roles: Boykin LM: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Savill A: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; De Barro P: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: The authors have no competing interests.

How to cite this article: Boykin LM, Savill A and De Barro P. Updated mtCOI reference dataset for the Bemisia tabaci species complex [version 1; referees: 2 approved] F1000Research 2017, 6:1835 (doi: 10.12688/f1000research.12858.1)

Copyright: © 2017 Boykin LM 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
Members of the *Bemisia tabaci* (whiteflies) species complex are among the world’s most devastating insect pests and cause billions of dollars (US) of damage each year, leaving farmers in the developing world food insecure (De Barro *et al.*, 2011). As a species complex with at least 34 members, identification is based on the use of the 657 bp portion of the 3’ end of the mitochondrial COI (mtCOI) (Boykin *et al.*, 2012, Boykin *et al.*, 2013). In order to identify members of the complex correctly, a curated reference dataset is a useful resource. In 2012, a reference mtCOI dataset was made available on the CSIRO data portal (De Barro & Boykin, 2012). Errors in the dataset were subsequently identified and so the dataset was updated on 15 May 2017 (http://doi.org/10.4225/08/591a4018dfca8) (De Barro & Boykin, 2017), but did not include new additions from GenBank (post 2012). Therefore, the dataset described herein represents the most up-to-date reference resource for members of the complex.

Methods
The CSIRO dataset (http://doi.org/10.4225/08/591a4018dfca8), updated 15 May 2017 was used as the starting point. The existing records were updated to include host plant data. New records post-2012 were then downloaded on 22 May 2017 directly from GenBank. All downloaded data was treated as follows:

1) Data was classified with BLAST using the new CSIRO reference data set
2) Sequences that caused gaps in the alignment were removed
3) Sequences that had stop codons present were removed
4) Clustal Omega (Sievers & Higgins, 2014) was used for preliminary alignment and fine tuning of the alignment was carried out with MAFFT (Katoh & Standley, 2013)
5) Duplicate sequences were then removed using BBMAP Dedupe (Bushnell, 2017)

In addition, all MEAM2 sequences were removed as they have now been confirmed to be pseudogenes (Tay *et al.*, 2017).

Data availability

Competing interests
The authors have no competing interests.

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

References

Whitefly (*Bemisia tabaci*) is becoming a global hazard for crop and ornamental plants. Identification of correct species is always better for the implication of best control strategy. In this regard, the effort made by Boykin et al for speedy and accurate identification of *B. tabaci* species complex is very significant. Additionally, the dataset developed with enriched quality is also very useful for whitefly biologist working in the area of evolution, mitochondrial genomics and crop management. All the bioinformatics tools used to generate this refined dataset are ideal to make such analysis.

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

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

**Are the datasets clearly presented in a useable and accessible format?**
Yes

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

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The updated mtCOI reference dataset for the *Bemisia tabaci* species complex will add a valuable contribution to researches for a fast and accurate identification of members from the *B. tabaci* species complex based on the partial mitochondrial COI gene. The high quality data is easily accessible for download and gathers whiteflies collected globally. The availability of a reliable and updated reference dataset is an essential tool that will aid the scientific community to identify and classify correctly this pest, the first step in the crop management against whiteflies.

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

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

**Are the datasets clearly presented in a useable and accessible format?**
Yes

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

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