Turning DNA barcodes into an alternative tool for identification: African fruit flies as a model

Massimiliano Virgilio[1,2], Kurt Jordaens[3], Floris Breman[3], Norman Barr[4], Thierry Backeljau[2] & Marc De Meyer[1]

[1]Royal Museum for Central Africa, Tervuren Belgium
[2]Royal Belgian Institute of Natural Sciences, Brussels, Belgium
[3]Joint Experimental Molecular Unit, Brussels & Tervuren, Belgium
[4]United States Department of Agriculture (APHIS), Edinburg, TX USA

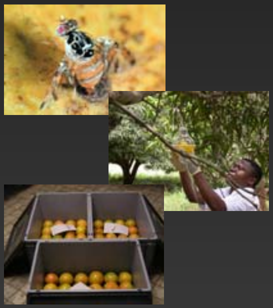
Objective

To verify whether morphological identification of intercepted fruit fly specimens can be corroborated by molecular identification through DNA barcodes.

Introduction

Several fruit flies (Diptera: Tephritidae) are pests of subtropical and tropical horticultural crops and pose a major threat to production and international trade of crops. For NPPO's there is a need for a fast, accurate, and unambiguous identification tool. Traditional morphological identification has limitations: it is only applicable to adults), is time consuming, and requires a local expert. Molecular diagnostics through DNA barcoding could provide a valid alternative.

The Consortium for the Barcode of Life (CBOL) initiated and supports the Tephritid Barcoding Initiative (TBI) as a demonstrator project. TBI's aim is to develop a DNA barcoding system for fruit flies. The Royal Museum for Central Africa (Tervuren) and the Royal Belgian Institute of Natural Sciences (Brussels), through their Joint Experimental Molecular Unit (JEMU) and a BELSPO Action 1 project, have been developing and testing a barcode library for African fruit flies.

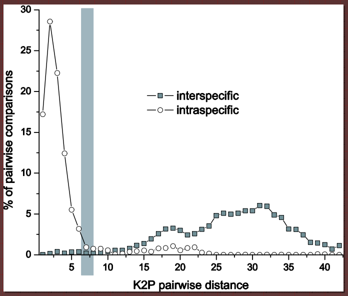


Materials and Methods

A DNA barcode library was generated including 602 sequences from 153 indigenous African and alien invasive fruit fly taxa, originating from 30 different African countries. Emphasis was on those African fruit fly genera of economic importance (EI): *Ceratitis*, *Dacus* and *Bactrocera*. The library comprises all species of economic significance with an average 9.8 sequences (SE=1.7) per species from different specimens, to consider intraspecific geographic variation. The library includes representatives of 85% of all taxa regularly encountered in para-pheromone traps during surveys in different parts of the African continent.

The reliability of the library was then tested on 235 'unknown' specimens intercepted by three European NPPO's or collected during recent monitoring surveys in three different countries (Togo, DR Congo, Mozambique). All material was first identified using morphological characters. They were then blind-tested by generating a barcode sequence and comparing this with the reference library. Using the Best Match Criterion (Meier, 2006), each was assigned the species name of the DNA barcode with the smallest genetic distance (K2P model) and considered as 'correctly identified' when morphological and molecular IDs matched.

Figure 1:
Distance analysis.
Distributions of interspecific divergence (grey squares) and intraspecific variation (white circles) based on pairwise K2P distances of 602 DNA barcodes used for the identification of African tephritid pests. Grey bar markes the overlap between the 95% percentiles of intra and interspecific-congeneric distributions (6.23%<K2P<7.98%)



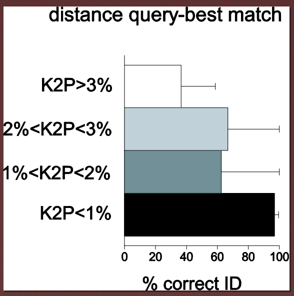


Figure 2.
Percentages of correct species identifications based on DNA barcodes for four levels of genetic distance between sequences of an unknown and a record in the reference library using Best Match criteria. Distances are shown for five taxonomic groups (*Bactrocera*, *Ceratitis*, *Dacus*, other genera). Bars show average percent correctly identified species; lines show maximum percent.

Results

The distribution of pairwise K2P distances (Fig. 1) showed that 95% of all intraspecific distances were in the interval 0.00 - 7.98% and that 95% of mean interspecific, congeneric distances were in the interval 6.23 – 13.55%. We did not observe a true barcoding gap as 6.31% of all pairwise comparisons were shared between the 95% percentiles of intra- and interspecific distributions (6.23%<K2P<7.98%).

Of the 235 unknown specimens (intercepted or collected), 188 could be identified to species level based on morphological characters. Overall, the Best Match based on DNA barcodes agreed with morphological identification for 82.3 % (SE=8.7%) with 100.0% of success in *Bactrocera,* 80.6% in *Ceratitis*, 89.7% in *Dacus* and 59.1% in the other genera. Most of the misidentified queries showed relatively high genetic distances with their best matches. Only 36.5% (SE=22.3%) of queries with distances >3% were correctly identified, 66.7% of queries with 2%>K2P>3% and 62.5% of queries with 1%>K2P>2%. Yet, 96.8% of queries with K2P <1% were successfully identified (Fig. 2).

Mismatches between morphological and molecular ID could be related to a) lack of a representative barcodes in the reference database (n= 15, 10 taxa), b) queries belonging to species complexes or to species groups with limited intraspecific divergence (*Ceratitis anonae, C. capitata,* n=2), c) taxa that demonstrate high intraspecific divergence and are in need of taxonomic revision, (*Dacus humeralis*, n=1), d) laboratory contamination or mislabeling (n=1), e) errors of unknown source (n=5). Forty-one out of the 47 unknowns that could not be identified morphologically to species or genus level were larvae. All of these were represented with an adult from the same rearing series. If the morphological ID of the corresponding adult is used as reference for the larval ID, 100% match was found between morphological and molecular ID for this particular subset.

Discussion

This study shows that DNA barcodes can provide a reliable alternative for accurate species identification if the following issues are taken into account:

1.Representativity of the library: accurate identification requires the taxon to be represented in the reference library. If not, the Best Match Criterion will produce a false positive. In our simulations, 15 out of the 24 mismatches (62.5%) were due to lack of representatives of a query species in the library.

2.Distance query-best match: in 70.8% of misidentified unknowns, K2P distances to a library record were higher than 2%. Yet, the performance of DNA barcoding was remarkably good (with a proportion of correct IDs = 96.8% (SE=2.7%) when distances were <1%). Hence, a divergence threshold might be used as cut-off mark to define when the best match to a library record should not be believed because there is no correct match in the reference library.

3.Customized distance thresholds: a barcode reference library can, a priori, be adjusted based on taxonomic knowledge of a group and the intraspecific divergence observed. Thresholds for identification will vary between the separating power in species complexes with little interspecific divergence (i.e. *Ceratitis* FAR complex) and taxa that show a high intraspecific divergence and may need taxonomic revision (i.e. *Dacus humeralis* and related taxa). The identification success, based on DNA barcodes, can be further increased this way.

4.In order to build a workable identification tool based on DNA barcodes, there is the need for close collaboration with taxonomists specialized on that particular target group.

Acknowledgements

We would like to thank Valérie Balmès (Laboratoire National de la Protection des Végétaux, France), Sharon Reid (Food and Environment Research Agency, UK), Eddy Dijkstra (Plant Protection Service, the Netherlands), Komina Amevoin (Université de Lomé, Togo) and Mervyn Mansell (USDA/APHIS Pretoria Branch, South Africa) for providing samples. The financial support of Belgian Science Policy (through JEMU and Action 1 of M. Virgilio) is acknowledged.

What is DNA Barcoding?

Since a proposal by Hebert et al. (2003), researchers have explored the idea that all biological species can be identified using a short gene sequence from a standardized position in the genome – a 'DNA barcode' – analogous to the black stripes of the Universal Product Code used to distinguish commercial products. In study after study, DNA barcoding is proving effective in:

- Assigning specimens to known species using only a tiny piece of tissue,
- Discovering new variation within what were presumed to be single species,
- Documenting the biodiversity of poorly known taxonomic groups and poorly sampled geographic areas.

Building the Global Reference Barcode Library

From Voucher Specimens in Museums…

Over the past 300 years, taxonomists have collected and described more than 1.7 million species of plants, animals and microbes. They have built collections of hundreds of millions of specimens from these species. These specimens have been studied, identified, cataloged, and now reside in museums, herbaria, botanical gardens, zoos and other repositories.

… To DNA sequences…

Voucher specimens in museums provide tissue samples that will produce a reference barcode for that species. Using the standard and widely available tools of molecular biology, DNA is extracted from the tissue of these specimens, the barcode region is isolated, replicated by PCR amplification, and sequenced.

… To a Public Global Barcode Database

BOLD, the Barcode of Life Data Systems is a public workbench that researchers are using to assemble and analyze their barcode records. Information on the voucher specimen, species name and barcode sequence are assembled on BOLD and then submitted to one of the three global databases of gene sequences: GenBank, EMBL and DDBJ, where they are available without charge.

How does it work?

The DNA barcode of an unidentified specimen can be read using standard gene sequencing techniques. DNA barcoding includes three types of activities:

- Working with organisms: Collecting, identifying, and preserving voucher specimens in secure repositories
- Laboratory procedures: Sampling and processing tissue from specimens to obtain DNA barcode gene sequences
- Managing data: Sharing the DNA barcode sequence and data about its voucher specimen in a public database

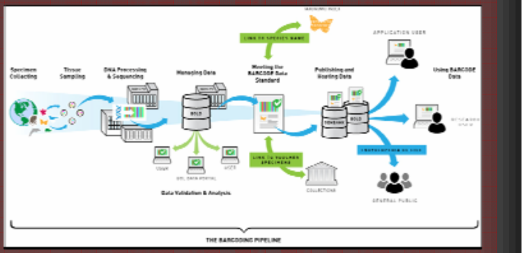


Figure 3. Schematic of the "barcoding pipeline" from specimens to data.

Advantages of DNA Barcoding

1.Works with fragments. Barcoding can identify a species from bits and pieces. Barcodes from gut contents and feces are being used to reconstruct foodwebs and predator-prey relations.

2.Works for all stages of life. Barcoding can identify a species in its many forms, from eggs and seed, through larvae and seedlings, to adults and flowers.

3.Unmasks look-alikes. Barcoding can distinguish among species that look alike, uncovering dangerous organisms masquerading as harmless ones and enabling a more accurate view of biodiversity.

4.Enables digital comparisons. Written as a sequence of four discrete nucleotides - CATG – along a uniform locality on genomes, DNA barcodes of life are a system for comparing specimens and species in quantitative, objective terms.

5.Democratizes access. A standardized library of barcodes will empower many more people - border inspectors, amateur naturalists, schoolchildren and others - to identify the species around them.

6.Accelerates species discovery and description. Taxonomists are still writing the encyclopedia of life. The public library of DNA barcodes linked to vouchered specimens and their binomial names is a workbench that scientists can use to document and study species. DNA barcode records are part of the growing network of biological knowledge that is emerging as an online Encyclopedia of Life on Earth that will include a web page for every species of plant and animal.

DNA barcoding is currently being tested or used routinely for:

- Improving food security by identifying agricultural pests
- Protecting public health by identifying disease vectors
- Protecting endangered species by identifying products and derivatives
- Improving food safety and protecting consumers by verifying food labeling
- Protecting water quality through biological assessment