DNA barcoding and mini-barcoding as molecular tools for identification of fruit flies *(Diptera: Tephritidae)*

(Received: 20. 01. 2016; Accepted: 04.02.2016)

Gessmallah A. E. *, Virgilio M. ‘*** , De Meyer M.** , Nevado B.*** ,**
Backeljau T. * ‘**** and Yousif M.T. ***

* The National Institute for Promotion of Horticultural Exports, University of Gezira, Wad Medani, Sudan.
** Royal Museum for Central Africa, Tervuren, Belgium.
*** Royal Belgian Institute of Natural Sciences, Brussels, Belgium.
**** University of Antwerp, Antwerp, Belgium.
Corresponding e-mail : azizatniphe@gmail.com

ABSTRACT

The objectives of this study were to assess the proportions of correct identifications (ID) in tested genera and to elucidate the efficacy of three fragments (min-barcode) of COI barcode region in providing comparable information. The DNA barcoding in three tephritid genera was tested by considering 493 DNA barcodes involving 86 species belonging to the genera Bactrocera (33 species), Ceratitis (21 species) and Dacus (32 species). Barcoding simulations were performed by using a reference dataset of 15,948 insect DNA barcodes. It was performed under a “best case scenario” viz. by providing one or more potential con-specific matches in the reference dataset for each query. Results showed that the Best Match (BM) criterion (i.e. the criterion currently adopted by BOLD) yielded different proportions of correctly identified taxa in Bactrocera (BM=0.839), Ceratitis (BM=0.868) and Dacus (BM=0.962). The proportions of correctly identified queries by using three mini-barcode fragments (MB1, MB2, MB3) of 220bp (corresponding to the first, second and last third of the COI barcode region) ranged from 0.71 ± 0.17 in MB1 to 0.83 ± 0.10 in MB2. Currently, however, the application of DNA barcoding in tephritid species is limited by the low number of barcoded taxa in the reference databases. This situation increases the probability of making Type II errors (i.e. incorrect ID for queries without conspecifics in the reference database). On the other hand, the probability of making Type I errors (incorrect ID for queries with conspecifics in the database) is relatively limited (4-16% in our simulations). These considerations suggest that DNA barcoding may not be a fool proof method for the molecular ID of tephritid fruit flies. Still, DNA barcoding could be effective under well-defined conditions, where only a limited number of well-known tephritid taxa, with well characterized intraspecific variation, are to be distinguished.

Key words: *Fruit-fly, Barcoding gap, Mini-barcoding, Conspecifics molecular identification.*

INTRODUCTION

Species identification based on morphological characters has significant limitations, which sometimes leads to incorrect identification. Thus, there is a need for a new approach for taxon recognition based on DNA construction. In this regards, DNA barcoding is defined as a technique for identifying organisms based on a short standardized fragment of genomic DNA (Utsugi, 2011). Herbert *et al.* (2003 a) proposed the 650 *bp* long mitochondrial cytochrome oxidase (CO1*)* for the identification of all living organisms. In their study for identifying *Sargassum* species, Lydiane and Claude (2010) found the *CO1* to have the best results of barcoding compared to other markers (*COX3*, Mitochondrial spacer *mtsp* and nuclear sequences). In general, the mitochondrial genome of animals is a better target for analysis than the nuclear genome (Herbert *et al.*, 2003a). Analysis based on cytochrome oxidase 1 genome (*CO1*) has merits of being very efficient for species identification since it has universal primers, which are very robust, and of a great range of phylogenetic signal (Folmer *et al*., 1994). Moreover, it has a highly constrained amino acid sequence that allows a broad application of primers and limits its information content at deeper phylogenetic levels. It reliably discriminates a diverse range of taxa at the species level) and that the universal primer originally designed for marine invertebrates can be applied to all animal phyla (Hebert *et al.*, 2003b). In contrast, limitations of using mitochondrial DNA (*mt* DNA) to infer species boundaries include retention of ancestral polymorphism, male -biased gene flow and selection following hybridization and paralogy i.e. homology that arises *via* gene duplication (Herbert *et al*., 2004).

In general, the extreme diversity of insects and their economical epidemiological and agricultural importance have made this group a major target of DNA barcoding (Utsugi 2011). The reliability of DNA barcoding in Diptera was evaluated by considering an alignment of 4,272 DNA barcodes involving 345 species from 75 genera and 24 families (Herbert *et al.*, 2003 a,b).
The objectives of this study were to:

Evaluate the performance of different identification criteria, quantify the identification success provided by different fragments of the COI barcode region and investigate relationships between barcode length and identification success in Diptera.

MATERIALS AND METHODS

This study was conducted at the Royal Belgian Institute of Natural Sciences, Brussels, Belgium. The NucleoSpin® tissue method (Alex *et al.*, 2005) was used in this study to extract DNA from specimens. Specimens tissue were added to the micro centrifuge tubes (1.5 ml), 180 µl buffer T1 and 25 µl proteinase K solution were added to the samples and vortex to mix. Then samples were incubated in Thermomixture at 56C^0 until complete lysis was obtained at 48 hours. DNA was extracted following the standard protocol (Macherey-Nagel, 2007) for animal tissues .Three fragments with sizes of 220, 280 and 340 *bp* were recognized and amplified from the 5' region of the cox1 gene from the mitochondrial DNA to give a full barcode of 660 *bp*, using different combinations of six newly designed primers. Table (1) shows primers, their sequence and the length of the fragments. Primers were developed by the Joint Experimental Units (JEMU) of Royal Belgium Institute for Natural Sciences (RBINS) and Royal Museum for Central

Africa (RMCA). The process of amplification was conducted in a thermocycler (4 Biometra Tpersonal thermocycler from Biometra GmbH, Germany), the parameters of amplification were given in Table 2. The amplification products were separated electrophoretically in a 1.2 % agarose gel. The running procedure was conducted according to the manufacturer's instructions for 15 min. The gel was visualized and photographed on a UV trans-illuminator equipped with a digital camera.

Table (1): Primers and their sequence, and length of the fragments.

Fragment	Length	Primers	Sequence of the primers
1	340bp	Fa CO1L1440_1464dt	TGTAAAACGACGGCCAGTTYTCAACAATCATAARGATATTGG
		Re Teph_H343_362t (Rc)	ATAGTAGAAAACGGAGCTGGGTCATAGCTGTTTCCTG-
2	220 bp	Fa Teph_L280_306t	ATGTAAAACGACGGCCAGTCGAATAAATAATATAAGATTTTGATTA
		Re Teph_H526_548t (Rc)	TTTGACCGAATACCTTTATTTGTGTCATAGCTGTTTCCTG
3	280bp	Fa Teph_L499_521t	TGTAAAACGACGGCCAGTATTAATATACGATCAACAGGAAT
		Re CO1H2123_2148dt(Rc)	CAGGAAACAGCTATGACTAWACTTCGGRTGWCCAAARAATCA

Table (2): Parameters of PCR.

a. Mixture

Product	µl/sample
dNTP (2mM)	2.5
10 x Taq buffer	2.5
Primer 1 (2um)	0.5
Primer 2 (2um)	0.5
MgCl2 (25 mM)	0.5
dd H_2O	15.40
Taq enzyme (5Ux ul)	0.10
DNA	3
Total	25

b. Conditions

Stage	Step	Temperature(^{0}C)	Time (min.)	Cycles
1	1	94	3.00	
2	1	94	0.5	
	2	50	0.5	45
	3	72	0.5	
3	1	72	7.00	

The positive PCR products were purified using a vacuum pressure device. A sample of 20 µl of each PCR product were added to 30µl Nano water (Nano pure +++/Deionized+) and then the plate was put in the vacuum (400-600 pressure) for 15min.Purified PCR products were subjected to sequencing reactions. The process of sequencing was conducted in Biometra TP professional thermocycler using the Big Dye Cycle Sequencing Kit. A forward and reverse reaction was performed for each sample using M13 Forward and M13 reverse primers. The sequenced products were purified following the clean-up protocol and sequenced in both directions with an ABI prism 3130XL Genetic Analyzer (16-caplilary sequencer) – from Applied Biosystem, Germany, following manufacturer's instructions. Sequencing mixture and conditions were given in Table 3.

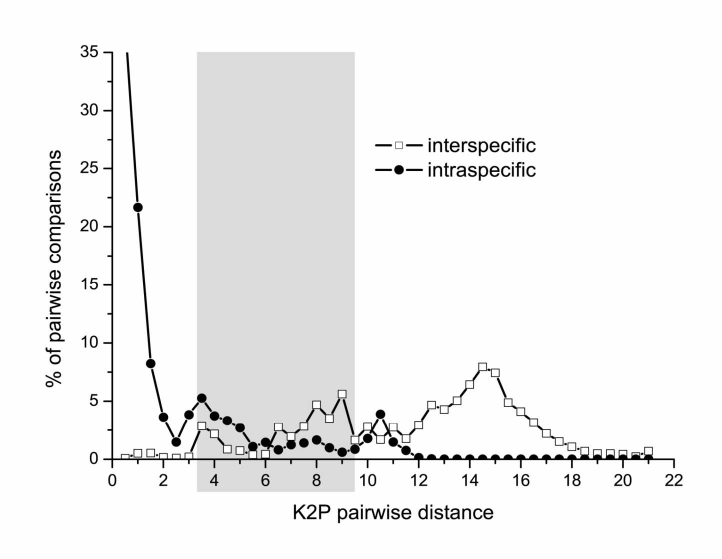

Fig. (1): Distributions of inter-specific (white squares) and intra-specific (black circles) pairwise K2P distances. In grey: overlap between the 95% percentiles of intra-and inter-specific distributions (3.14 %< K2P<9.62%).

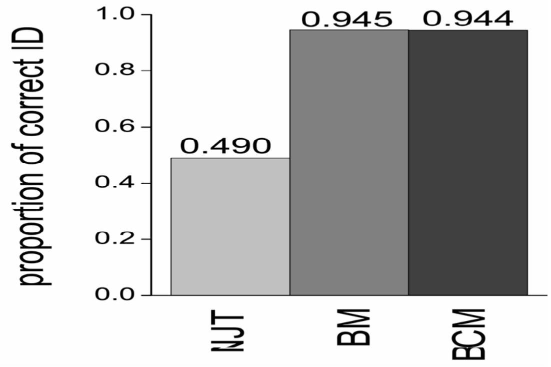

Fig. (2): Proportion of correctly identified queries through Neighbor-Joining Tree (NJT), Best Match (BM), and Best Close Match (BCM).

DNA barcodes available in June 2009 in BOLD (http://www.barcodinglife.org) was used in this study. All the publicly available sequences belonging to the order *Diptera* were downloaded and aligned. DNA barcodes were trimmed in order to include only the barcode region, *viz.* the 658 bp *COI* fragment amplified by the "universal primers" of Folmer *et al.* (1994). Sequences shorter than 550 bp and DNA barcodes with incomplete species information (e.g. sequence names including sp., cf., nr… *etc.*) were discarded. Pairwise Kimura's two parameter (K2P) distances were calculated and their frequency distributions for intra-specific and congeneric inter-specific distances were plotted as described by Kimura (1980).

Table (3): Parameters of sequencing reactions.

a. Mixture

Product	µl/sample
Ready reaction mix	2.0
5x seq buffer	1.0
Primer M 13(2uM)	2.0
dd H_2O	2.0
DNA	3
Total	10

b. Conditions

Stage	Step	Temperature (^{0}C)	Time (min.)	Cycles
1	1	96	1.00	
2	1	96	0.10	25
	2	50	0.05	
	3	60	4.00	

The proportion of correct matches were estimated by three identification criteria (*viz.* Best Match: BM; Best Close Match: BCM; tree-identification: NJT) as described by Meier *et al.* (2006). Relationships between barcode length and identification success were analyzed through non-linear regression. The DNA barcodes were divided in three non-overlapping "mini-barcodes" of 220, 219 and 219 bp corresponding to the first, second and last third of the barcode region (hereafter MB1, MB2, MB3). The number of base pairs of each mini-barcode was further reduced at both 5' and 3' ends in order to obtain fragments of approximately 75%, 50%, 25% and 10% of the initial mini-barcode length (164, 110, 55 and 22bp, respectively). Non-linear regression fitting was implemented for each combination of identification criterion and barcode fragment following the first order exponential decay model y = y0 + ae(-x/t), where y0 = Y offset, a = amplitude, t = exponential time constant following Sokal and Rohlf (1995).

RESULTS AND DISCUSSION

Intra- and inter-specific genetic distances (Fig. 1) were largely overlapping with 26.43% of pair-wise comparisons shared between the 95% percentiles of distributions. Moritz and Cicerozoon (2014) stated that when intra- and inter-specific distances are widely overlapped the DNA barcoding-based identification is not effective. Moreover, a mean intra-specific divergence of 10 times was proposed as the standard threshold for differentiating species (Herbert *et al.*, 2004 a, b).

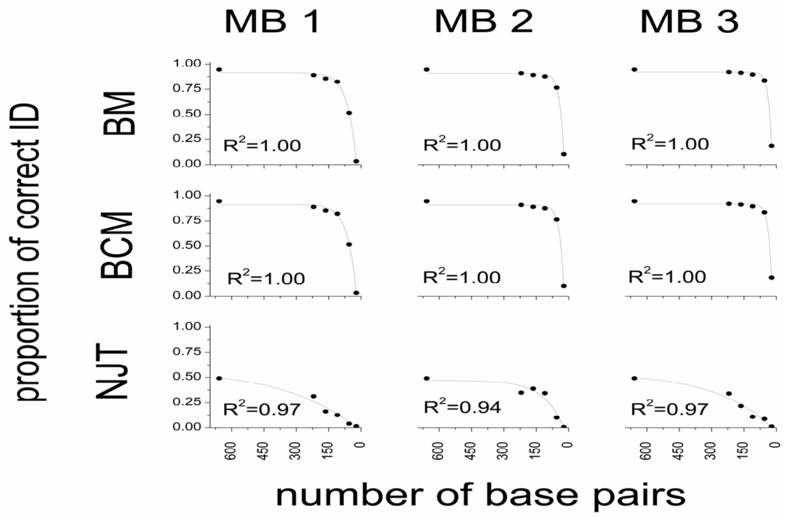

Fig. (3): Relationships between barcode length and identification success of Diptera.

Although there was no clear “barcoding gap” (Hebert *et al.*, 2004), the Best Match BM and Best Close Match BCM criteria yielded relatively high proportions of correct identification of BM=0.945, BCM=0.944, respectively. Results were in a line with Lydiane and Cluade (2010) who found results obtained by BM and BCM appeared un-related to the overlap results. They performed better than Neighbor-Joining Tree NJT of a proportion of correctly identified queries (ID)=0.490 (Fig. 2). Different regions of the barcode fragment provided comparable information (Fig. 3) and mini-barcodes of 220bp still yielded substantial proportions of correct IDs (BM = 0.906 ± 0.016; BCM = 0.905 ± 0.016). These results support the BM and BCM methods *per se*. Lydiane and Cluade (2010) found that identification success using BM and BCM was the highest (91.3%) and ambiguous compared to other criteria, mis-identification and no match scores were among the least in identifying *Sargassum* species. Nevertheless, the application of DNA barcoding in *Diptera* is currently limited by the low number of barcoded taxa. It was demonstrated that identification success of a barcode marker my vary in regard to the size, geographical span and relatedness of species of/in the dataset considered (Cao *et al.*, 2001; Meyer and Paulay, 2005; Ledford, 2008; Liu *et al.*, 2010). For instance, lack of DNA barcodes for ~96% of the described *Diptera* species implies a high probability of making Type II errors (i.e. incorrect identification for queries without conspecifics in the reference database). Conversely, the probability of Type I error (misidentification of queries with conspecifics in the database) is relatively low (approximately 5% in our simulations). In conclusion these considerations suggest that DNA barcoding may not be a foolproof method for the molecular ID of Diptera though it could be effective under well-defined conditions, where only a limited number of well characterized taxa are to be distinguished. Effort should be directed toward exploring the utility of the DNA based diagnostic tools at the level of immature stages (the available stage inside the fruits during inspection, to facilitate identification of these pests at ports of entry.

REFERENCES

Alex, M.S.; Brian, I.F. and Paul, D.N.H. (2005). DNA bracoding for effective biodiversity assessment of a hyper diverse arthropod group: the ants of Madagascar. Philosophical Transact Royal Soc. 360: 1825-1834.

Cao, H.; Ca, J.N.; Liu, Y.P.; Wang, Z.T. and Xu, L.S. (2001). Correlative analysis between geographical distribution and nucleotide sequence o chloroplast *matK* gene o *Cndium monnien* fruit in China. Chin Pharm J. 36:373-376.

Folmer, O. B. M., Hoeh. W., Lutz. R. and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome C oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotech 3:294-299.

Hebert, P.D.N.; Cywinska, A.; Ball, S.L. and deWaard, J. R. (2003a). Biological identifications through DNA barcodes. Proceedings of the Royal Society Lond. Ser. B. 270: 313–322.

Hebert, P.D.N.; Ratnasingham, S. and deWaard, J.R. (2003b). Barcoding animal life: Cytochrome oxidase subunit 1divergences among closely related species. Proceedings of the Royal. Society Lond. Ser. B. 270: 96–99.

Hebert, P. D. N., Stoeckle, M.Y., Zemlak, T. S. and Francis, C. M. (2004). Identification of birds through DNA barcodes. PLoS Biol., 2:e312.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol., 16:111-120.

Ledford, H. (2008). Botanical identities: DNA barcoding or plants comes a step closer. Nature 451: 616.

Liu. Y.N.; Xu, L. and Don, D.Q. (2010). PCR amplification, cloning and sequence analysis o rDNA ITS of *Arctium lappa* from different geographical origin in China. Trad. Herbal Drugs, 33:26-28.

Lydiane, M. and Claude, D. (2010). Assessment of five markers as potential barcodes for identifying Sargassum subgenus *Sargassum* species (Phaeophyceae, Fucales).Cryptogamie, Algologie, 31(4): 467-485.

Macherey-Nagel, Gmb H. (2007). Genomic DNA from tissue user manual Nucleo Spin. NucleoMag 96 preps MN Cat. No. 744 400 1. www.mn-net.com.

Meier, R., Shiyang, K., Vaidya, G. and Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst Biol 55:715-728.

Meyer, C.P. and Paulay. G. (2005). DNA barcoding error rates based on comprehensive sampling .PLoS Biol. 3 (12): 422.

Moritz, C. and Cicero, C. (2004). DNA barcoding : Promise and pitfalls. www.lposbiology.org. vol. 2, Issue 10. 279-354.

Sokal, R. R. and Rohlf, F. J. (1995). Biometry: the principles and practice of statistics in biological research: Freeman Press.

Utsugi, J. (2011). Current progress in DNA barcoding and future implications or entomology. Entomological Science 14 (2): 107-124.

الملخص العربي

إستخدام تقنية تسلسل الحمض النووي في تصنيف ذبابة الفاكهة (Diptera: Tepheridae)

عبد العزيز الأمين قسم الله*، ماسمليانو فيرقيليو ،*** مارك ديمير**، نيفادو***، تيري باكلجو *** ،****
و محمد طه يوسف***

*المعهد القومي لتنمية الصادرات البستانية ، جامعة الجزيرة ، واد مدني، السودان.
**المتحف الملكي لوسط أفريقيا ، ترفورين، بلجيكا.
***المعهد الملكي البلجيكي للعلوم الطبيعية ، بروكسل، بلجيكا.
****جامعة أنتويرب، إنتويرب، بلجيكا.

هدفت هذه الدراسة الى تحديد نسب صحة تصنيف النوع لعدد من انواع ذبابة الفاكهة مع توضيح مدى كفاءة استخدام ثلاث تسلسلات قصيرة مأخوذة من الحمض النووي Cytochrome Oxidase (CO1) في توفير معلومة يمكن مطابقتها مع قاعدة بيانات مرجعية. تم إستخدام ٤٩٣ تسلسل حمض نووي والتي تتبع الى ٨٦ نوع منتمية الي ثلاثة اجناس من ذبابة الفاكهة منها ٣٣ نوع منتمية الي الجنس *Bactracera* و٢١ نوع ينتمي الى الجنس *Ceratitis* و٣٢ تنتمي الى الجنس *Dacus*. تم اجراء هذه الدراسة بإستخدام تسلسل الحمض النووي لعدد ١٥٩٤٨ من افراد ذبابة الفاكهة كقاعدة بيانات مرجعية. ولضمان دقة النتائج فقد تم وضع كل من التسلسلات المراد معرفة تصنيفها مع واحد او اكثر من تسلسلات مماثلة متواجدة في قاعدة البيانات المرجعية. تم الحصول على نسب دقة مقارنة للنوع (Best Match (BM مختلفة باستخدام نظام Barcode of Life Data Systems في الجنس *Bactracera* (بدقة مقارنة=٠.٨٣٩)، والجنس *Ceratitis* (بدقة مقارنة =٠.٨٦٨) والجنس *Dacus* (بدقة مقارنة=٠.٩٦٢). وقد تراوحت نسب صحة تصنيف النوع باستخدام تسلسلات قصيرة بطول ٢٢٠ *bp* من الجين CO1 (MB1 ، MB2 ، MB3) بين ٠.٧١±٠.١٧ في التسلسل MB1 و ٠.٨٣±٠.١٠ للقطعة MB2 . يعتبر استخدام تقنية تحديد النوع باستخدام تسلسلات من الحمض النووي محدودا في ذبابة الفاكهة Tephritid مما يزيد من احتمالية وجود خطأ من النوع الثاني type IIالمرتبط باجراء تصنيف النوع باستخدام هذه التقنية بدون وجود تسلسلات مشابهة في نظم المعلومات المرجعية. على الجانب الآخر، فان احتمالية وجود النوع الاول من الخطأ Type I يعتبر محدودا ١٦.4%) وهو الخطأ المرتبط بوجود مراجع في انظمة المعلومات المرجعية. حاليا لايعتبر استخدام التصنيف الجزيئي بمفرده كافياً لتصنيف ذبابة الفاكهة حيث تتطلب كفاءة هذه التقنية توفر ظروف معملية تتمثل في استخدام عدد محدود من ذبابة الفاكهة مع الدراسة الدقيقة المسبقة للتباين الوراثي بين هذه الافراد.