

## **Identifying IAS based on DNA barcoding using currently available sequence data: details on applied material and methods.**

N. Smitz, S. Gombeer, K. Meganck, A. Vanderheyden, Y.R. Van Bourgonie, T. Backeljau, and M. De Meyer

Please use the following citation:

N. Smitz, S. Gombeer, K. Meganck, A. Vanderheyden, Y.R. Van Bourgonie, T. Backeljau, and M. De Meyer, "Identifying IAS based on DNA barcoding using currently available sequence data: details on applied material and methods." 2019. [Online]. Available from: <https://bopco.be/output/iasfactsheets>.

### *Screening the online repositories for available sequences*

The online repositories BOLD and GenBank were screened for sequences of all members of the genera to which the invasive alien species (IAS) listed on the EU Regulation 2016/1141 (and its update 2017/1263) belong. All available data (including full mitochondrial and chloroplast genomes) were downloaded and imported into Geneious Prime® 2019 (Biomatters Ltd., Auckland, New Zealand). This dataset was cleaned by removing sequences with incomplete species identifications, sequences which were labelled as unverified in the sequence description, shotgun and mRNA sequences, microsatellite sequences, and sequences of parasites, viruses and bacteria. The leftover sequences, including those extract from the genomes were divided by DNA marker. Note that if the genus of the IAS was monospecific, the same procedure was followed at a higher taxonomic level. The investigated DNA markers and the taxa included in the final Neighbour-Joining (NJ) tree are reported in Table 2 of each factsheet.

### *Preparing the datasets for analyses*

In case the geographical origin of the IAS sequences was not provided directly by GenBank or BOLD, the literature was consulted. This information was used to evaluate the completeness of the sequence datasets with regard to the geographical range of the IAS in order to assess the amount of intraspecific genetic variation represented in the dataset of available sequences. Once the geographic data of the IAS was complemented, identical sequences (i.e. duplicates) were deleted as follows: (i) for the congeneric species of the IAS, identical sequences were searched and deleted per species, in order to retain knowledge on the overall sequence sharing among species within the genus; (ii) for the IAS, identical sequences were searched and deleted per country in order to retain information on the geographical representation in the final NJ-tree. Because Geneious Prime® defines duplicates as identical sequences which also have the same length, deleting identical sequences was repeated at a later stage after trimming (see below). For each DNA marker under investigation outgroup sequences belonging to species of a related taxon were added to the marker datasets.

### *Cluster analyses to evaluate the performance of the markers for species identification*

DNA sequences were aligned using Geneious Prime® Alignment since this method determines sequence read directions and detects sequences that are entered in reverse format. This alignment was trimmed to the length of the longest IAS sequences after which the dataset was again screened for identical sequences (see above). The leftover sequences were re-aligned using ClustalW (as implemented in Geneious Prime®: cost matrix = IUB, Gap open cost = 15, Gap extend cost = 6,66) and sequences that were less than half the trimmed alignment length were discarded. To finalise the dataset for NJ-tree building, sequences which were causing gaps in the alignment of coding marker regions, which contained a large number of ambiguity codes or which did not align with any of the sequences in the dataset (e.g. due to mislabelling of species or DNA marker region) were deleted if multiple other sequences were available for that species (or country in case of IAS). The final dataset was realigned using ClustalW, after which the alignment was used to

construct a NJ-tree with Geneious Prime® (genetic distance model: Tamurei-Nei). The bootstrapping re-sampling method was used to estimate the confidence levels of the clusters obtained in the NJ-trees (number of bootstrap replicates: 1000; bootstrap support threshold: 70 %). The tree was rooted using the outgroup sequences. Upon request, trees can be made available. Please contact the BopCo team via [bopco@naturalsciences.be](mailto:bopco@naturalsciences.be) or through the BopCo website (<https://bopco.be>).

#### *Interpretation and evaluation of the Neighbour-Joining trees*

The ability of the different DNA markers to provide reliable identifications of the IAS, was evaluated by scoring five issues which are presented in Table 1 of each factsheet:

- (1) Are there sufficient DNA sequences available of the IAS to represent its intraspecific variation?  
There is no set numerical cut-off for the assessment of this issue. Each case (species and marker) was evaluated depending of the amount of intra- and inter-species sequence variation and the quality of the sequences (e.g. length);
- (2) Is the geographic coverage of the IAS sequences representative?  
In case at least one sequence of both the native and the invasive region is present, the geographic coverage is considered sufficient;
- (3) Do the conspecific IAS sequences form a well-supported cluster?  
Here both the inclusiveness for all sequences of the IAS cluster(s) as well as the bootstrap value are assessed;
- (4) Are there DNA sequences that might come from misidentified reference specimens and which therefore may confound the analyses?  
In the case of a suspected misidentification, the sequence reliability is evaluated based on its place of publication (scientific journal or direct submission to repository), the correctness of other sequences from that study, the potential for confusion due to similar looking species or a species which occurs in the same geographic range, and the clustering of the other conspecific sequences in the genus. These issues are considered in the discussion of the factsheet if applicable;
- (5) Are DNA sequences available for all species in the genus of the IAS (or at higher taxonomic level for monospecific genera)?  
At least one sequence per species is considered sufficient.

The downloaded data, the dataset preparation process and the outcome of the NJ-tree analyses are discussed for each IAS individually, resulting in one factsheet per IAS, taking into account the current knowledge on the taxonomy, the geographical distribution, and the available literature on phylogeny and sequence divergence. Based on this information each DNA marker is evaluated with respect to its ability to provide a reliable identification of the IAS.

In each factsheet a final conclusion is formulated specifying which DNA marker is currently the most reliable option to identify the IAS using the online reference repositories. When applicable the need for additional data is described.

