USER MANUAL

Citation: Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T, Virgilio M (2013) Adhoc: an R package to calculate *ad hoc* distance thresholds for DNA barcoding identification. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 329–336. doi: 10.3897/zookeys.365.6034

Table of content

1 Preliminary note 1

2 Install R and the required packages 1

3 Prepare the input file 2

4 Function "checkDNAbcd" 2

5 Function "adhocTHR" 4

6 References 7

1 Preliminary note

This manual is conceived for non-expert R users and provides a step-by-step guide to apply two R functions and calculate *ad hoc* distance thresholds following Virgilio et al. (2012). An *ad hoc* distance threshold relies on an estimated probability of relative identification error as calculated for a particular reference library of DNA barcodes. *Ad hoc* distance thresholds are obtained through two consecutive functions. A first function (checkDNAbcd) imports the reference library and provides descriptive information on the imported dataset in order to allow some quality control (sequence labeling, sequence lengths, etc.). The output of the first function is used by the second function (adhocTHR), to calculate an *ad hoc* distance threshold. This second function considers each sequence of the reference library as a query and finds its best match(es) among the DNA barcodes of the library. The function then quantifies the relative identification errors (see below) obtained for a set of arbitrary distance thresholds (30 distance values by default), performs linear (or polynomial) regression and calculates the *ad hoc* distance threshold corresponding to an expected relative identification error (5% by default).

2 Install R and the required packages

Both functions have been developed under R version 2.15.1. and tested under version 3.0.1

Install R: http://cran.r-project.org/doc/manuals/R-admin.html

Download and install the packages spider (Brown et al. 2011), ape (Paradis et al. 2004), pegas (Paradis 2010) and polynom (Venables et al. 2013), which depend on the following other packages:

gee, nlme, lattice, Hmisc, igraph, network, MASS, adegenet and ade4 (http://www.freestatistics.org/cran/).

Execute R and load the package spider by clicking on "Load package…" in the menu "packages".

3 Prepare the input file

A single input file (named "input.fas") of aligned non-interleaved DNA sequences in FASTA format is needed (the reference library of DNA barcodes). Sequence labels should have the following structure: ">species_name_any_additional_information" as in the following example (note that character strings have to be separated by underscores):

>Bactrocera_amplexa_Kenya_1052_JEMU

CCCTTTATTTTATTTTCGG

>Bactrocera_cucurbitae_Benin_AB33598852_JEMU

AATTATATTTTATTTTCGG

4 Function "checkDNAbcd"

4.1. Run the function and check the imported data

This function imports the reference library of DNA barcodes, provides an overview of its content, calculates all pairwise distances and delivers an output that will be used by the next function.

Change the location of the R workspace by clicking on "Change dir…" in the menu "File" and select the directory where the file "input.fas" is located (working directory).

Copy the script of checkDNAbcd (available on http://jemu.myspecies.info/computer-programs) and paste it on your R console to run the script.

checkDNAbcd will:

- create a function called checkDNAbcd in your R workspace.
- run the function through the command checkDNAbcd("input.fas")->out1.
- export two spreadsheets, mylabels.csv and listsp.csv, in the working directory:
    - mylabels.csv provides both parts of the species names and the complete label of each sequence (as extracted from the label).
    - listsp.csv lists the number of sequences (Nseq) and haplotypes (Nhap) for each species of the reference library.
- create a list (out1) containing the tables mentioned above (type out1$mylabels and out1$listsp to visualize them) and four matrices containing sequence lengths (out1$DNAlength), all pairwise distances (out1$dist), intraspecific pairwise distances (out1$spdist$intra) and interspecific pairwise distances (out1$spdist$inter).
- plot histograms (Figure 1) for the distributions of:
    - sequence lengths,

- pairwise interspecific distances,
- pairwise intraspecific distances,
- combined intra- and interspecific distances.

These histograms and output files allow a rapid quality check of the reference sequence library and the format of the sequence labels.

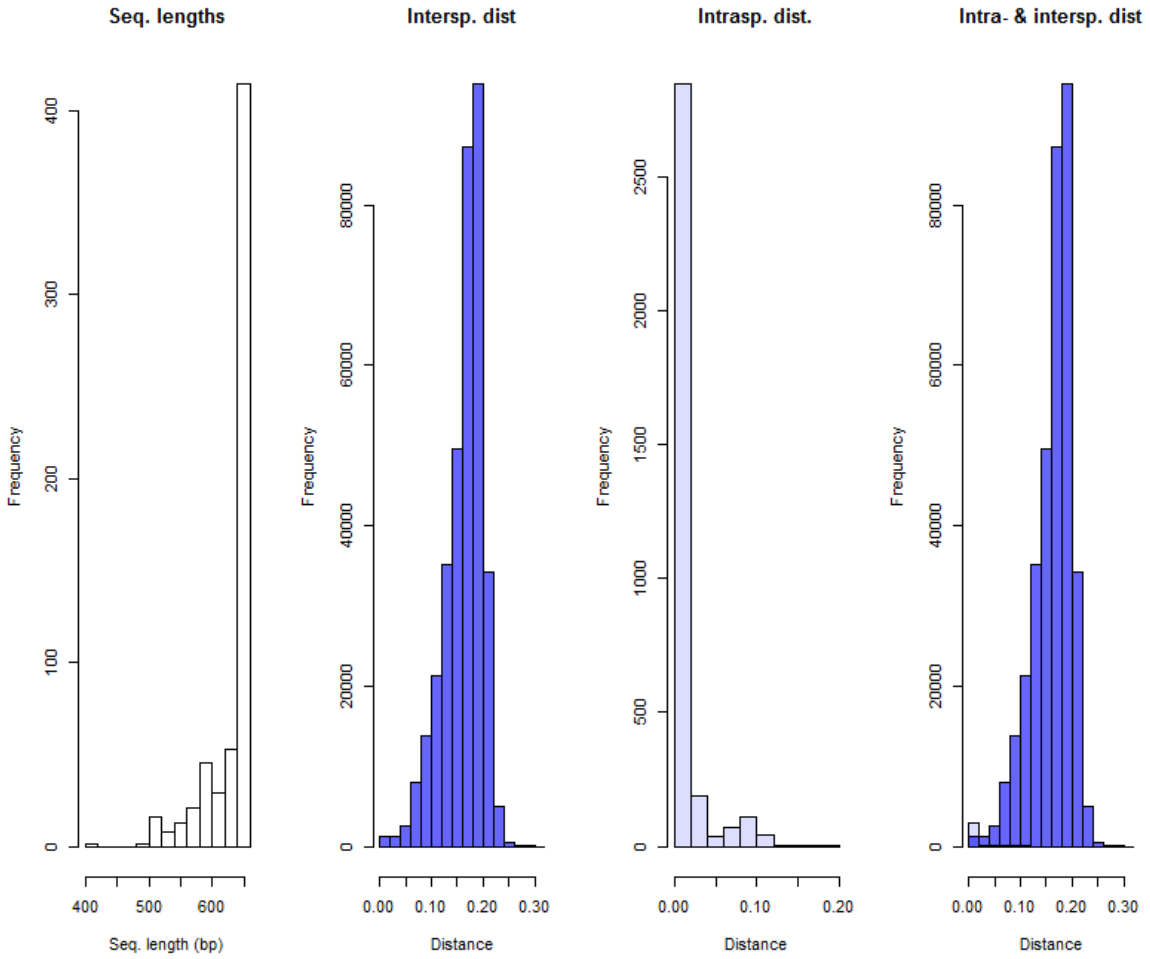

Figure 1 Sequence lengths, intra- and interspecific distances measured in the imported library: All four histograms were obtained with the function checkDNAbcd using the example file (input.fas). From left to right: distribution of sequence lengths, distribution of pairwise interspecific, intraspecific and both intra- and interspecific distances.

4.2. Options for the checkDNAbcd function

By default, the checkDNAbcd function calculates K2P (Kimura 1980) distances ("K80") between all pairs of DNA sequences. Other methods can be set by the user (a complete list is available at: http://127.0.0.1:10099/library/ape/html/dist.dna.html). For example, changing the command line

```
checkDNAbcd("input.fas")
```

to

```
checkDNAbcd("input.fas", DistModel="raw")
```

will allow using uncorrected p-distances ("raw"):

5 Function "adhocTHR"

5.1. Run the function and get an *ad hoc* threshold

Copy the script of the function adhocTHR (available on http://jemu.myspecies.info/computer-programs), paste it on your R console and press "enter".

adhocTHR will:

- create a function called adhocTHR in your R workspace.
- run it through the command: adhocTHR("out1")->out2.
- export four spreadsheets (in the same directory as the input file): Bestmatches.csv, ID.csv, redflagged.csv and redflaggedSP.csv. Each of them is described below.
- create a list (out2) with the above-mentioned tables, the coefficients of the regression (out2$reg) and the inferred *ad hoc* threshold (out2$THR).
- plot the distribution of relative identification errors (RE) observed at each distance threshold, the linear fitting and the graphical representation of the inferred *ad hoc* distance threshold (Figure 2).

Description of the four spreadsheets exported by adhocTHR:

- Bestmatches.csv reports, for each query (sequence identified by its label in the table):
  - distBM: the distance to its best match(es),
  - idBM: the assignment of a species name according to the best match criterion (*sensu* Meier, 2006),
  - IDeval: the evaluation of the identification according to the best match criterion *i.e.*:
    - TP = true positive, (the query and its best match(es) belong to the same species, correct identification)
    - FP = false positives (the query and its best match(es) belong to different species, erroneous identification)
    - FPambiguous = at least one correct and one erroneous match for the same query, ambiguous identification.

    N.B. When more than one species name is found among the best matches, only one of the allospecific assignments is given in idBM.
- ID.csv summarizes the results of best close match identification (*sensu* Meier, 2006) and calculates at each arbitrary distance threshold:
  - the number of TP, FP, TN and FN with

- TP = true positives: correctly accepted identifications, the query and its best match(es) belong to the same species, the distance query-best match is below the threshold,
- FP = false positives: erroneously accepted identifications, the query and its best match(es) belong to different species, the distance query-best match is below the threshold,
- TN = true negatives: correctly discarded identifications, the query and its best match(es) belong to different species, the distance query-best match is above the threshold,
- FN = false negatives: erroneously discarded identifications, the query and its best match(es) belong to the same species, the distance query-best match is above the threshold,.

  - OE = overall identification error (the overall proportion of erroneous identifications): (FP+FN)/total number of reference DNA barcodes,
  - RE = relative identification error (the proportion of erroneous identifications among queries that where not discarded): FP/(TP+FP),
  - accuracy: (TP + TN)/total number of reference DNA barcodes,
  - precision: TP/(TP+FP).

- redflaggedSP.csv lists the species responsible for ambiguous identifications (grouped per row).
- redflagged.csv details the matches obtained for each ambiguous identification with:
  - Nb_species: number of species responsible for the ambiguous identification,
  - Nb_conspecific_seq: number of conspecific reference sequences,
  - Nb_allospecific_seq: the number of allospecific reference sequences and
  - List_of_best-matches: the labels of all best matches responsible for the ambiguous identification.

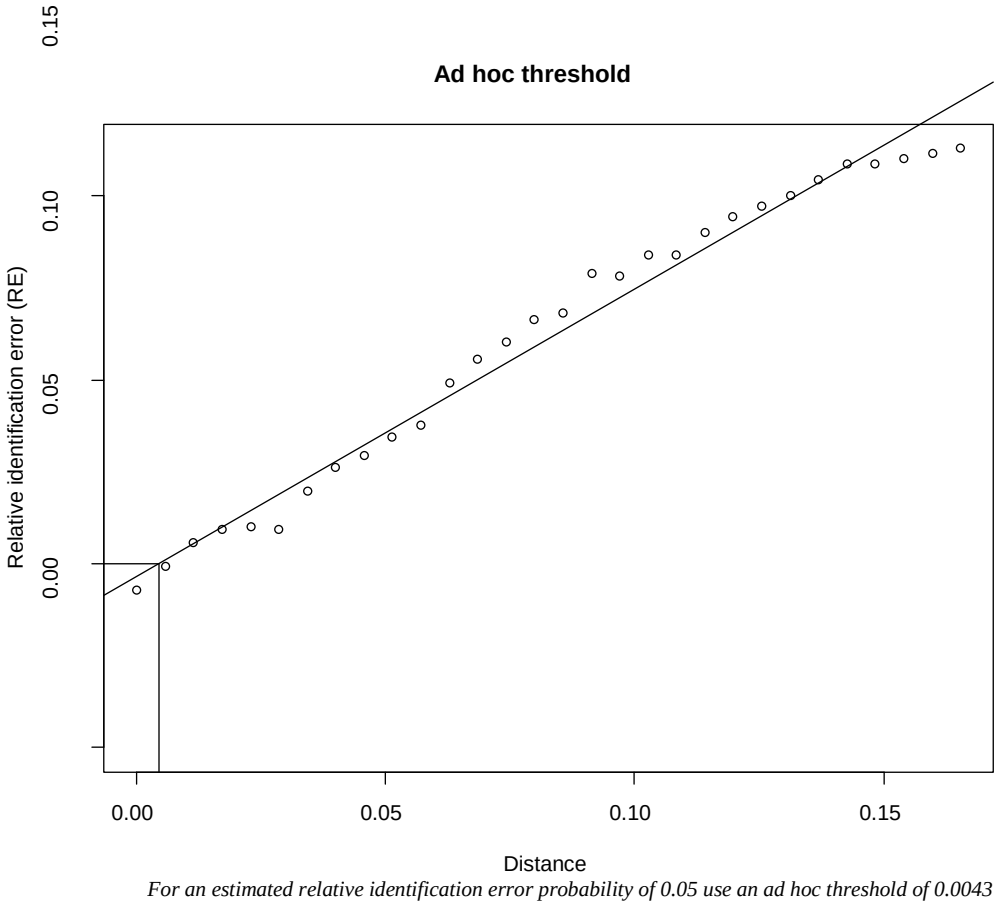

Figure 2 Estimation of the *ad hoc* distance threshold: This graph was obtained with the function adhocTHR for the example file (input.fas) using default settings (K2P genetic distance, 30 arbitrary distance thresholds, estimated relative identification error of 5%).

5.2. Options for the adhocTHR function

By default, the function calculates an estimated relative error probability of 5% and samples 30 distance thresholds equidistantly distributed between K2P = 0.0 and the largest distance between all pairs of query - best match(es) obtained for the particular imported dataset. These parameters can be modified by the user, for example the relative error probability can be set to 10% and the number of arbitrary distance thresholds to 40 by replacing:

```
adhocTHR(out1)->out2
```

with

```
adhocTHR(out1, NbrTh=40, ErrProb=0.1)->out2
```

By default, ambiguous identifications are treated as incorrect (FP or FN depending on the distance threshold). However, it is possible to consider ambiguously identified queries as being correctly

identified (TP or TN). This option (Ambig="correct") can be useful when all ambiguous identifications are due to synonymies or if a complex of species is treated as a single taxonomical entity. Another option allows to exclude all ambiguous matches from the calculations (Ambig="ignore"). For this, the command line at the end of the script:

```
adhocTHR(out1)->out2
```

should be modified as follows:

```
adhocTHR(out1, Ambig="correct")->out2
```

or:

```
adhocTHR(out1, Ambig="ignore")->out2
```

If the user wants to evaluate the effects of ambiguous identifications on the relative identification errors and hence on the estimated *ad hoc* distance threshold, it is also possible to exclude all ambiguous matches from calculations. In this case, the command line at the end of the script

```
adhocTHR(out1)->out2
```

should be changed to:

```
adhocTHR(out1, Ambig="ignore")->out2
```

The user has the possibility to estimate the *ad hoc* distance threshold on the basis of polynomial regression (rather than linear). For this, check that the package polynom (Venables et al. 2013) is installed and modify the command line at the end of the script

```
adhocTHR(out1)->out2
```

to:

```
adhocTHR(out1, Reg="polynomial")->out2
```

5.3. Comments about the output

In particular cases, (*e.g.* reference libraries with low taxon coverages) all best matches might result in correct identifications, with RE = 0.0 at all distance thresholds and the regression line being parallel to the x axis. In this situation, adhocTHR will give the following message: "All identifications are correct when using the best match method (no distance threshold considered). An *ad hoc* threshold for best close match identification cannot be calculated".

In other cases, reaching an estimated RE of 5% might not be possible, even at the most restrictive distance threshold (distance threshold = 0.00) and regression fitting will intercept the y axis above the RE value. In this case, adhocTHR will give the following message: "The estimated RE cannot be reached using this reference library". The user should then either increase the relative error probability (RE) or try and increase the taxon coverage of the library (see Virgilio et al. 2012).

6 References

Brown SDJ, Collins RA, Boyer S, Lefort MC, Curtis N, Malumbres-Olarte J, Vink CJ, Cruickshank RH (2012) Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources **12**: 562-565. doi: 10.1111/j.1755-0998.2011.03108.x.

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution **16**: 111-120. doi: 10.1007/BF01731581.

Meier R, Shiyang K, Vaidya G, Ng, PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology **55**: 715-728. doi:10.1080/10635150600969864.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**: 289-290. doi: 10.1093/bioinformatics/btg412.

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419-420. doi: 10.1093/bioinformatics/btp696

Venables B, Hornik K, Maechler M (2013) polynom: a collection of functions to implement a class for univariate polynomial manipulations. R package version 1.3-7. http://cran.r-project.org/web/packages/polynom/index.html

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. (2012) Identifying insects with incomplete DNA barcode libraries, African Fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE **7**: e31581. doi: 10.1371/journal.pone.0031581.