Reviewer’s report

Title: Supervised DNA Barcodes species classification: analysis, comparisons and results

Version: 1  Date: 18 December 2013

Reviewer: Damon Little

Reviewer’s report:

Major Compulsory Revisions

page 4: There are many more similarly and character methods. Either list them all or be clear that you are only listing a subset and explain how the subset was selected.

page 6: Please explain how ambiguous bases (e.g. K) and missing data (e.g. -) are treated.

page 7: Please indicate which gene(s) are used as barcodes or each dataset.

page 8: Please explain how the sequences for each reference set were selected.

page 9: Please explain how the averages for each algorithm were calculated. Were they weighted by the number samples?

page 9: There needs to be a statistical test of differences among algorithms. Is the better performance observed in some algorithms statistically significant?

Minor Essential Revisions

page 2: Not all barcodes are mitochondrial: plants use plastid sequences, fungi use nuclear sequences.

page 2: Names of protein coding genes should be set in italics.

page 2: The double plural of 'Classifiers families' is not correct.

page 2: Because 'kingdoms' is plural, 'animals, fungi and plants' should not be.

page 2: 'function based', 'rule based', and 'trees based' need to be hyphenated.

page 3: Although the publication of Hebert et al. did, in many sense, begin DNA barcoding the technique nor the name were novel.

page 3: Not all barcodes are mitochondrial: plants use plastid sequences, fungi use nuclear sequences.

page 3: Alignment of barcode sequences is not trivial. The only one that can be aligned across the kingdom that it is used is rbcL.
page 3: What is a 'strictly related species'?

page 3: 'has promoted' not 'promote'.

page 3: 'species by' not 'species starting by'.

page 3: 'composed of' not 'composed with'.

page 3: Isn't the number of specimens required for reliable classification a function of the amount of intraspecific variation? Please be more specific in your statement.

page 3: Please rephrase 'only one data set could be even provided'.


page 4: 'framework and software are not' not 'framework and software is not'.

page 4: The double plural of 'Classifiers families' is not correct.

page 4: Because 'kingdoms' is plural, 'animals, fungi and plants' should not be.

page 5: 'and then changing some values in order to compare the different performances' is not proper English.

page 5: 'its own file' not 'an own file'.

page 5: 'program, available at' not 'software, available on'.

page 5: 'adapting different input formats to permit the execution of these and new experiments' is not proper English.

page 5: Names of protein coding genes should be set in italics.

page 6: Omit 'Indeed,' in the first sentence.

page 6: 'that need to be' not 'which need to be'.

page 6: 'FASTA format shows' is not proper English.

page 6: 'divided' (or 'separated') not 'dived'.

page 6: Omit ',', respectively'.

page 6: 'when an insufficient' not 'when not a sufficient'.

page 6: 'no support ' not 'actually no support '.

page 7: 'based on' not 'basing on'.

page 7: I do not understand what 'which allow to keep a general applicability' means.
I do not understand what 'due to the incomplete clustering in Barcode tree' means.

'data from' not 'data coming from'.

'from' not 'and coming from'.

'. 112 species' not ', whose 112'.

'Drosophila' must be set in italics.

I do not understand what 'was usually selected' means.

'without a Barcode gap' not 'without Barcode gap'

I do not understand what 'which show a huge number of effective' means.

'and there exist 15 species with a number of representing sequence is higher than five.' is not proper English.

'Inga' must be set in italics.

At this point I stopped correcting English errors, there are too many. Please have a native speaker carefully edit the revised manuscript.

Discretionary Revisions

I think the dataset descriptions would be better as a table.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Not suitable for publication unless extensively edited

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

'I declare that I have no competing interests