Journal of Computational Biology: http://mc.manuscriptcentral.com/liebert/jcb

Coalescent-based DNA barcoding: multilocus analysis and robustness

Journal:	*Journal of Computational Biology*
Manuscript ID:	JCB-2011-0122
Manuscript Type:	Original Paper
Keyword:	STATISTICS, SEQUENCE ANALYSIS, coalescence, genetics
Abstract:	DNA barcoding is the assignment of individuals to species using standardized mitochondrial sequences. Nuclear data are sometimes added to the mitochondrial data to increase power. A barcoding method for analysing mitochondrial and nuclear data is developed. It is a Bayesian method based on the coalescent model. Then this method is assessed using simulated and real data. It is found that adding nuclear data can reduce the number of ambiguous assignments. Finally, the robustness of coalescent-based barcoding to departures from model assumptions is studied using simulations. This method is found to be robust to past population size variations, to within-species population structures and to designs that poorly sample populations within species.

Coalescent-based DNA barcoding: multilocus analysis and robustness

Olivier David[1,*] Catherine Larédo[1,3] Raphaël Leblois[2,4]

Brigitte Schaeffer[1] Nicolas Vergne[1,5]

[1]UR341, Mathématiques et informatique appliquées, INRA, F-78350 Jouy-en-Josas, France

[2]Muséum National d'Histoire Naturelle, UMR 5202 MNHN/CNRS, Laboratoire Origine Structure Evolution de la Biodiversité (OSEB), Paris, France

[3]Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 et 7, UMR CNRS 7599, Paris, France

[4]Centre de Biologie et de Gestion des Populations (CBGP), UMR INRA-IRD-CIRAD 1062, Montferrier-sur-Lez, France

[5]Laboratoire de Mathématiques Raphaël Salem, UMR 6085 CNRS-Université de Rouen, 76801 Saint-Etienne-du-Rouvray, France

*Corresponding author

Abstract

DNA barcoding is the assignment of individuals to species using standardized mitochondrial sequences. Nuclear data are sometimes added to the mitochondrial data to increase power. A barcoding method for analysing mitochondrial and nuclear data is developed. It is a Bayesian method based on the coalescent model. Then this method is assessed using simulated and real data. It is found that adding nuclear data can reduce the number of ambiguous assignments. Finally, the robustness of coalescent-based barcoding to departures from model assumptions is studied using simulations. This method is found to be robust to past population size variations, to within-species population structures and to designs that poorly sample populations within species.

Key words: Bayesian inference, classification, coalescent, DNA barcoding, species assignment.

1 Introduction

DNA barcoding is the assignment of individuals to species or higher taxonomic levels using standardized genetic data observed on the target individuals and samples from each species (Frézal and Leblois, 2008; Valentini et al., 2009). The DNA barcode project is conceived as a standard system for fast and accurate identification of all eukaryotic species (Hebert et al., 2003; Miller, 2007). The DNA barcode itself consists of a 648 bp region of the cytochrome c oxidase 1 (COI) gene. Additionally to the mitochondrial COI gene, nuclear loci are sometimes also considered to improve assignment performance (Austerlitz et al., 2009; Elias et al., 2007).

DNA barcoding is a classification problem rather than a clustering one since the classes (species) are predefined and do not have to be inferred from the data (but see Pons et al. (2006) for an application of clustering to barcoding). Barcoding assignment methods can be divided into similarity methods based on the match between the query sequence and the reference sequences such as BLAST search, phylogenetic approaches (Hebert et al., 2003; Elias et al., 2007), classification algorithms with no underlying biological models such as the nearest-neighbour method and methods based on population genetics (Matz and Nielsen, 2005). Two Bayesian methods based on models have recently been developed. In the method of Munch et al. (2008a), species are assumed to evolve according to a phylogenetic model while the within-species variation is not modelled. Conversely, in TheAssigner, the method of Abdo and Golding (2007), species are assumed to evolve independently and the dependence between sequences within species is modelled using a classical population genetics model called the coalescent. The latter is a model for the genealogical tree of a random sample of genes drawn from a large panmictic population (Chapter 10, Ewens, 2004; Kingman, 1982a,b; Tajima, 1983). Model-based barcoding methods raise various issues. Current methods assume that the data are mitochondrial and cannot cope with nuclear data. Moreover, their robustness to departures from model assumptions has been

little studied.

The main objective of the present paper is to study how to take account of nuclear data in coalescent-based classification and to study the robustness of this type of classification to departures from model assumptions. First a coalescent-based classification for assigning individuals to species using mitochondrial data is developed (Section 3). Then this method is extended to take account of nuclear data (Section 4). Finally the performance and robustness of coalescent-based classification are studied using simulated and real data sets (Sections 5 and 6).

Mary Ann Liebert, Inc., 140 Huguenot Street, New Rochelle, NY 10801

2 Bayesian classification

First we briefly review some basic material on Bayesian classification. In this method, individuals are assumed to belong to c classes. A data set y is available that includes measurements observed on reference individuals whose class is known. The objective is to predict the class $z \in \{1, \ldots, c\}$ of a test individual given its data x and the reference data y.

In Bayesian classification, a test individual is assigned to the class with the largest posterior probability of membership (Abdo and Golding, 2007; Munch et al., 2008a; Ripley, 1996). The assignment may be considered as ambiguous if the latter probability does not exceed some specified threshold. According to Bayes theorem, the posterior probability that a test individual belongs to class i is equal to $P(z = i|y, x) = P(z = i, x|y)/P(x|y) = r_i / \sum_k r_k$, where:

$$r_i = P(z = i|y)P(x|y, z = i). \quad (1)$$

In this equation, $P(z = i|y)$ is the probability that the test individual belongs to class i given the reference data y prior to the knowledge of x and plays the role of a prior probability of membership. The probability $P(x|y, z = i)$ is the conditional probability that an individual sampled in class i has data x. Bayesian classification is optimal for the 0–1 loss function (Chapter 2, Ripley, 1996) and provides a measure of assignment confidence.

3 Species assignment with mitochondrial data

We now apply Bayesian classification to DNA barcoding. In this section, the data consist of mitochondrial DNA sequences. The assumed demographic model is a set of isolated and panmictic species with a common ancestry at a given time in the past (i.e., the divergence time). This demographic model is the same as the one of Abdo and Golding (2007). The mitochondrial locus is assumed to evolve according to the coalescent model within each species independently. Following the standard coalescent, it is assumed that species sizes do not vary over time, that there is no migration between species and that all alleles are neutral. All individuals are assumed to be sampled at the same time and the species of any test individual is assumed to be represented in the reference data y. In this model, mutations occur on each ancestral lineage of species i according to a Poisson process with parameter $\theta_i/2$. The assumed mutation model is the infinitely many-sites model (ISM), in which a gene is considered as an infinitely long DNA sequence and each new mutant site is sampled uniformly and independently along the sequence (Chapter 9, Ewens, 2004). Finally, it is assumed that at each site it is known which base is the mutant base or the ancestral base (Section 7) and that there are no missing data or errors in the data.

The mutation parameters θ_i are first assumed to be known. Then, under the assumption that species evolve independently, the probability $P(x|y, z = i)$ in (1) is equal to:

$$P(x|y, z = i) = P(x|y_i, z = i),$$

where y_i denotes the data of species i in the reference data base. This probability will be written for simplicity as $P(x|y_i)$ in what follows. Generally it cannot be calculated explicitly under the ISM but it can be estimated as follows. It is equal to (p. 420, De Iorio

and Griffiths, 2004, supplementary materials A):

$$P(x|y_i) = \frac{n_i(x)+1}{n_i+1}\frac{P^0(x,y_i)}{P^0(y_i)}, \tag{2}$$

where P^0 is the probability of an unordered sample, n_i is the number of genes in the sample of species i and $n_i(x)$ is the number of genes with sequence x in the sample of species i. The probabilities $P^0(x,y_i)$ and $P^0(y_i)$ can be estimated using importance sampling (IS) (De Iorio and Griffiths, 2004, supplementary materials A). Note that the probability $P^0(y_i)$ needs to be estimated only once if there are several individuals to assign.

Mutation processes are generally unknown for most species and the vector θ of mutation parameters is thus usually not known. In this case, the posterior probabilities of membership can be estimated by plug-in, that is by assuming that θ is known and equal to an estimate $\widehat{\theta}$, computed from the reference data set as in Abdo and Golding (2007). The vector θ may be estimated, for example, using the method of Watterson (1975), by coalescent-based maximum likelihood or Bayesian methods (Bahlo and Griffiths, 2000; Kuhner et al., 1995). Alternatively, a predictive approach can be used in which the dependence of probabilities on θ is removed by integration (Chapter 2, Ripley, 1996, supplementary materials A).

4 Species assignment with mitochondrial and nuclear data

Individuals are now assumed to be genotyped at l diploid nuclear loci in addition to the mitochondrial locus. The two sequences of an individual at a nuclear locus are assumed to be known (Section 7). The genetic data of a test individual are denoted by $x = (x_0, \ldots, x_l)$, where x_0 is the mitochondrial sequence and x_j $(j \geq 1)$ is the pair of sequences at nuclear locus j. Each locus is assumed, as in the previous section, to evolve according to the coalescent model within each species independently. All the loci are assumed to evolve independently (Hudson, 1991; Nordborg, 2001) and there is no recombination within a locus. Mutations are assumed to occur according to the ISM with parameter $\theta_{ij}/2$ for species i and locus j. Finally, for simplicity, all parameters θ_{ij} are assumed to be known in this section.

With independent loci, the quantity r_i in (1) is equal to (Chapter 8, Ripley, 1996):

$$r_i = P(z = i|y) \prod_{j=0}^{l} P(x_j|y_{ij}, z = i), \tag{3}$$

where y_{ij} is the reference data for species i and locus j. This equation allows us to easily combine the mitochondrial and the nuclear informations. For a nuclear locus in a diploid species, (2) becomes:

$$P(x_j|y_{ij}, z = i) = \frac{(n_{ij}(x_{j1}) + 1)(n_{ij}(x_{j2}) + 1 + \delta_j)}{(n_{ij} + 2)(n_{ij} + 1)} \frac{P^0(x_j, y_{ij})}{P^0(y_{ij})},$$

where $\delta_j = 0$ if the test individual is heterozygote at locus j and $\delta_j = 1$ if the test individual is homozygote at locus j. In this equation, x_{j1} and x_{j2} denote the two test sequences at locus j, n_{ij} denotes the number of genes sampled for species i at locus j, $n_{ij}(x_{j1})$ denotes the multiplicity of allele x_{j1} in the sample of species i and locus j. The probabilities $P^0(x_j, y_{ij})$

and $P^0(y_{ij})$ can then be estimated using IS on coalescent histories as before.

Mary Ann Liebert, Inc., 140 Huguenot Street, New Rochelle, NY 10801

5 Simulation study

Simulations were carried out to assess the methods described above. In these simulations, one ancestral species split T generations ago into two new species with effective size N_e and mutation parameter θ. There were n reference individuals in each species. First sequences were simulated for a mitochondrial locus and a diploid nuclear locus to study the effect of adding nuclear data. Then to test the robustness of the methods developed, mitochondrial sequences were simulated assuming that species size varied over time or that each species was divided into several populations exchanging migrants. To mimic extreme sampling strategies that can be done in structured populations, we considered an "extended" sampling, in which the reference individuals were sampled in all populations for each new species, and a "clustered" sampling, in which all reference individuals were sampled from a single population in each new species. Details on these simulations are presented in supplementary materials B.

The simulated data were analysed with the nearest-neighbour classification (1NN) and the developed Bayesian assigner (BA) (supplementary materials B). The 1NN method was used because it had been found to be efficient compared with other barcoding methods (Austerlitz et al., 2009) and it was expected to be robust since it was not based on a specific biological model. This method was implemented with bagging in order to obtain a measure of confidence for an assignment (Hastie et al., 2001, supplementary materials B). Assignment performance was quantified using sensitivity and specificity (Munch et al., 2008a,b). Specificity is the fraction of non-ambiguous assignments (Section 2) that are correct. Sensitivity is the fraction of all the assignments that are correct.

The simulations with nuclear data first showed that performances were the best for the combination of the mitochondrial and the nuclear data, intermediate for the mitochondrial data (Fig. 1), and the least good for the nuclear data alone (Fig. S2 and S3). The poor results for the nuclear data alone were probably due to the larger effective size we used for

the nuclear locus, leading to smaller scaled divergence times T/N_e and thus lower levels of differentiation between the two new species. Nevertheless adding nuclear data clearly increased sensitivity (Fig. 1). This was mainly due to a reduction of the number of ambiguous assignments since specificity did not increase much (Fig. 1). Our simulations also showed that 1NN and BA had similar performances, except for the nuclear data alone for which 1NN had a low sensitivity (Fig. S3). However, we can note that BA had more ambiguous assignments than 1NN but made fewer errors among the non-ambiguous assignments (Fig. 1). Another important result was that the estimation of mutation parameters did not change the BA performance much (Fig. S1). Finally and as it was expected, increasing the values of θ, T or n improved the performance of both methods as in Austerlitz et al. (2009).

For past population size variations, the main results were that past expansions strongly increased specificity, sensitivity and the rate of non-ambiguous assignments, whereas past contractions had the opposite effect of decreasing specificity and sensitivity (Fig. 2). Our simulations also showed that past expansions affected both methods similarly, but 1NN always showed a slightly better performance than BA. On the contrary, it is interesting to note that the effect of past contractions was more pronounced for 1NN than for BA, resulting in much better performances for BA. Finally, the effect of past population size variations was found to be important for all the growth rate values we used and to be stronger for expansions than for contractions.

The effect of population structure was more complex because it depended on the sampling strategy. Compared with the unstructured species results, a population structure with a weak migration mainly affected sensitivity and the rate of non-ambiguous assignments, that both increased for the “clustered” samples and decreased for the ”extended” samples (Fig. 2). This result was unexpected as the population of origin of a test individual was represented by two individuals in the reference samples for the “extended” samples but not for the “clustered” samples. Finally, we note that population structure affected both

methods similarly and that the effect of population structure became noticeable only when migration was weak enough.

Mary Ann Liebert, Inc., 140 Huguenot Street, New Rochelle, NY 10801

Figure 1: *Effect of adding nuclear data on the performance of coalescent-based barcoding.* Specificity is the fraction of non-ambiguous assignments that are correct. Sensitivity is the fraction of all the assignments that are correct. The probability threshold is the threshold used to decide if an assignment is ambiguous. 1NN and BA are the nearest-neighbour classification and the developed Bayesian assigner with a known value of θ. The subscripts m and mn denote the mitochondrial data and the combination of mitochondrial and nuclear data, respectively. Adding nuclear data increases sensitivity and reduces the ambiguity of assignments.

Figure 2: *Robustness of coalescent-based barcoding to past population size changes and population structures.* 1NN and BA are the nearest-neighbour classification and the developed Bayesian assigner with estimated mutation parameters. Results for past population size changes are presented on the first two lines, with Gf being the growth factor. A growth factor larger than one indicates a population expansion from divergence to present, whereas a growth factor smaller than one indicates a population decline. Results for population structures are presented on the last two lines, with Nm being the number of migrants exchanged between adjacent populations in one generation. BA appears robust since its performance is similar to the one of 1NN that is model-free.

6 Analysis of real data sets

We chose to test our method on two different data sets that contained both differentiated and undifferentiated species. The first data set used came from the study of Hebert et al. (2004) on Astraptes species and consisted of mitochondrial sequences (CO1 locus). The second data set used came from the study of Elias et al. (2007) on Ithomiinae species and consisted of mitochondrial (CO1 locus) and nuclear data (EF1α locus). The data were analysed with 1NN, BA and TheAssigner (Abdo and Golding, 2007). The performance of each method was quantified using a leave-one-out analysis in which each haplotype was used as a test sequence after reducing its multiplicity by one in the reference data. Details on these data sets and their analyses are given in supplementary materials C.

The results first showed that adding nuclear data reduced the ambiguity of the BA assignments (Fig. 3). The analyses also showed that no method had the highest specificity in all cases (Fig. 3). Moreover BA had a lower sensitivity than the other methods and thus assigned fewer individuals (Fig. 3), except for the nuclear Ithomiinae data alone (Fig. S4). Another result of our analyses was that some posterior probabilities of membership were sensitive to the choice of the ancestral bases (supplementary materials C). Finally a few conditional probabilities were estimated with the predictive method (supplementary materials A) and the corresponding estimates were close to the plug-in estimates.

Figure 3: *Performance of coalescent-based barcoding with real data.* 1NN, TA and BA are the nearest-neighbour classification, TheAssigner and the developed Bayesian assigner. The subscripts m and mn denote the mitochondrial data and the combination of mitochondrial and nuclear data. Adding nuclear data increases sensitivity and reduces the ambiguity of BA assignments. No method has the best specificity for both data sets.

Mary Ann Liebert, Inc., 140 Huguenot Street, New Rochelle, NY 10801

7 Discussion

Classification inputs. Bayesian classification requires prior probabilities of membership. When these probabilities are not known, they may be estimated from the reference data provided that these data can be considered as a random sample among all the species considered (page 53, Ripley, 1996) or they may be fixed to $1/c$.

The developed methods require the ancestral sequence of each locus. If this sequence is not known, it can be inferred from the data (Bahlo and Griffiths, 2000; Gascuel and Steel, 2010) or posterior probabilities of membership may be estimated using unrooted trees (Section 5, Tavaré and Zeitouni, 2004; Bahlo and Griffiths, 2000). Moreover, many sequences from the barcoding reference database could be used as outgroups and thus greatly facilitate the inference of the ancestral sequence.

Finally, both alleles of an individual at a nuclear locus were assumed to be known. Current genotyping technologies are able to determine which two bases are present at each site of a nuclear locus but not the two sequences of the locus. It is a general problem for most nuclear sequence analysis methods, and statistical methods, known as phasing methods, can infer these two sequences from unphased data together with missing data (Scheet and Stephens, 2006).

Classification assumptions. The mutation model considered in this paper was the ISM, a model that requires fewer computations than models with a finite number of sites. However it assumes that a particular mutation can only occur once so that in particular there is no homoplasy. It is more adapted to situations where species are closely related since the assumption of absence of homoplasy is more likely to be satisfied in this case. This does not seem to be a problem for DNA barcoding since species that are distantly related to a test individual can be discarded using simpler methods (Austerlitz et al., 2009; Munch et al., 2008b). In our study, classification methods were compared using data sets that were compatible with the ISM so that all the methods had the same amount of information.

Species classification based on the ISM could be extended to account for different mutation rates for transitions and transversions.

The species of a test individual was assumed to be represented in the reference data. The conditional probabilities of an allele $P(x|y_i)$ can be used to check if this assumption is satisfied: low probabilities are an indication that this assumption may not be satisfied.

The developed methods are based on various simplifying assumptions. It would be interesting to relax some of these assumptions to improve classification performance. The program genetree can perform likelihood estimations with varying population size and population structures under the ISM (Bahlo and Griffiths, 2000). Divergence models and models that combine phylogenetic and population genetics models do not assume that species are independent (Matz and Nielsen, 2005; Pons et al., 2006).

Performance of the developed methods. The method developed to combine mitochondrial and nuclear informations appeared satisfactory. Adding nuclear data reduced the ambiguity of assignments in our analyses.

We showed that coalescent-based classification was robust to departures from demographic stability and panmixia and to designs that did not sample the within-species variation efficiently. It performed similarly to a model-free method (1NN) in the robustness study. Demographic expansion was found to increase the power of barcoding. This is an expected result, however, considering that speciation events are probably often associated with founder events followed by demographic expansions or selective sweeps on the mitochondria, it may highlight the reasons why DNA barcoding works so well with a limited sequence information.

Finally, no assignment method was found to be always the best in our analyses. Similar results were obtained by Austerlitz et al. (2009) when comparing phylogenetic and statistical methods. However the developed Bayesian assigner generally appeared more cautious than the other methods in the sense that it assigned fewer individuals but made fewer errors

among the assigned individuals.

The supplementary materials referenced in Sections 3, 5 and 6 are available at arxiv.org.

Mary Ann Liebert, Inc., 140 Huguenot Street, New Rochelle, NY 10801

Acknowledgements

This study was funded by the Agence Nationale de la Recherche (IFORA ANR-06-BDIV-014 and EMILE NT09-611697 projects). We thank F. Austerlitz for helpful comments.

Disclosure statement

No competing financial interests exist.

References

Abdo, Z., Golding, G.B., 2007. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. Systematic Biology 56, 44–56.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., Larédo, C., 2009. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics, Special Issue Biodiversity Informatics .

Bahlo, M., Griffiths, R.C., 2000. Inference from gene trees in a subdivided population. Theor. Popul. Biol. 57, 79–95.

De Iorio, M., Griffiths, R.C., 2004. Importance sampling on coalescent histories. I. Adv. Appl. Prob. 36, 417–433.

Elias, M., Hill, R.I., Willmott, K.R., Dasmahapatra, K.K., Brower, A.V., Mallet, J., Jiggins, C.D., 2007. Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc R. Soc. B. 274, 2881–9.

Ewens, W.J., 2004. Mathematical population genetics. volume 27 of *Interdisciplinary Applied Mathematics*. Springer. second edition.

Frézal, L., Leblois, R., 2008. Four years of DNA barcoding: Current advances and prospects. Infection, Genetics and Evolution 8, 727 – 736.

Gascuel, O., Steel, M., 2010. Inferring ancestral sequences in taxon-rich phylogenies. Mathematical Biosciences 227, 125 – 135.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics, Springer.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004. Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America 101, 14812–14817.

Hebert, P.D.N., Ratnasingham, S., deWaard, J.R., 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. B 270, S96–S99.

Hudson, R.R., 1991. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7, 1–44.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Processes and their Applications 13, 235 – 248.

Kingman, J.F.C., 1982b. On the genealogy of large populations. Journal of Applied Probability 19, 27–43.

Kuhner, M.K., Yamato, J., Felsenstein, J., 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140, 1421–1430.

Matz, M.V., Nielsen, R., 2005. A likelihood ratio test for species membership based on DNA sequence data. Philosophical Transactions of the Royal Society B - Biological Sciences 360, 1969–1974.

Miller, S.E., 2007. DNA barcoding and the renaissance of taxonomy. Proceedings of the National Academy of Sciences 104, 4775–4776.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E., Nielsen, R., 2008a. Statistical

assignment of DNA sequences using bayesian phylogenetics. Systematic Biology 57, 750–757.

Munch, K., Boomsma, W., Willerslev, E., Nielsen, R., 2008b. Fast phylogenetic DNA barcoding. Philosophical Transactions of the Royal Society B 363, 3997 – 4002.

Nordborg, M., 2001. Coalescent theory, in: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), Handbook of Statistical Genetics, John Wiley & Sons, Inc., Chichester, U.K.. pp. 179–212.

Pons, J., Barraclough, T., Gomez-Zurita, J., Cardoso, A., Duran, D., Hazell, S., Kamoun, S., Sumlin, W., Vogler, A., 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55, 595–609.

Ripley, B.D., 1996. Pattern recognition and neural networks. Cambridge University Press, Cambridge, UK.

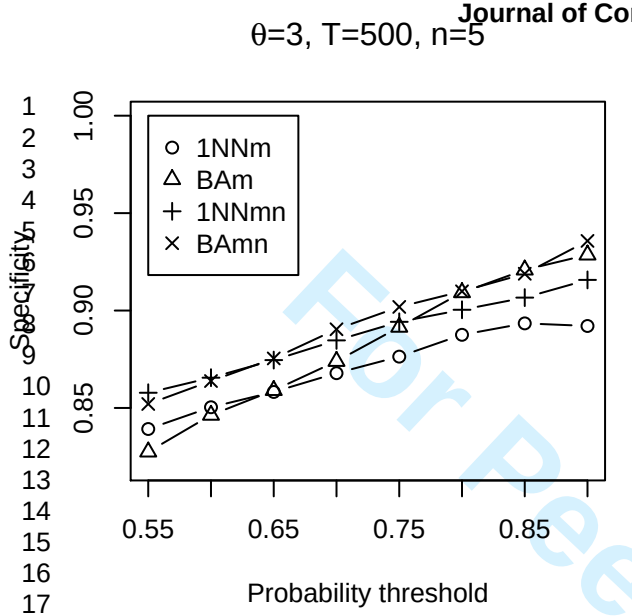
Scheet, P., Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78, 629–44.

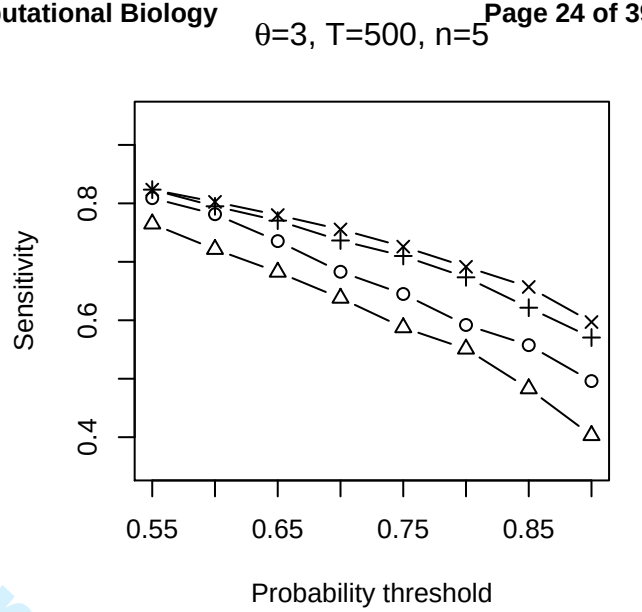
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.

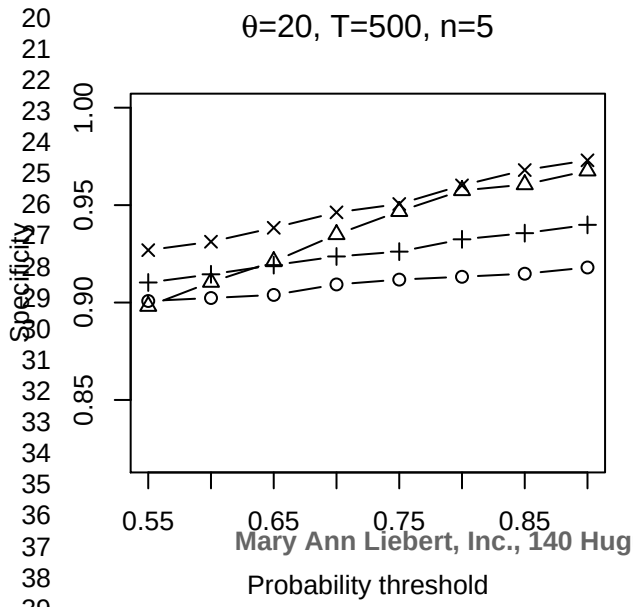
Tavaré, S., Zeitouni, O., 2004. Lectures on probability theory and statistics : Ecole d'été de probabilités de Saint-Flour XXXI - 2001. Lecture notes in mathematics, Springer.

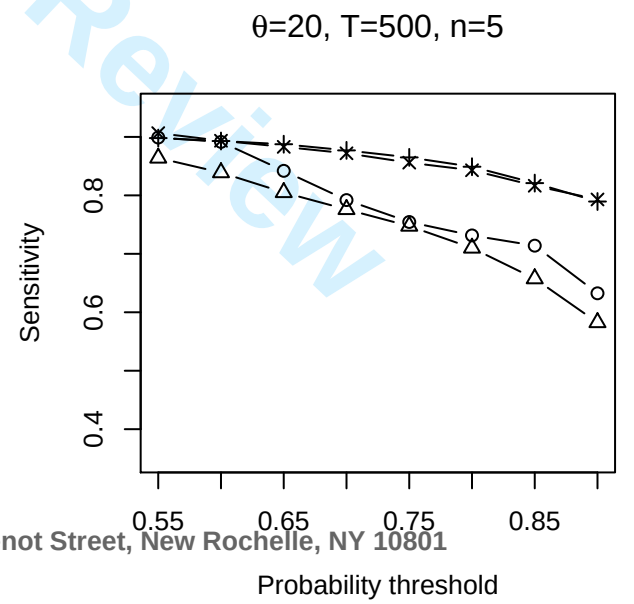
Valentini, A., Pompanon, F., Taberlet, P., 2009. DNA barcoding for ecologists. Trends in ecology & evolution 24, 110–7.

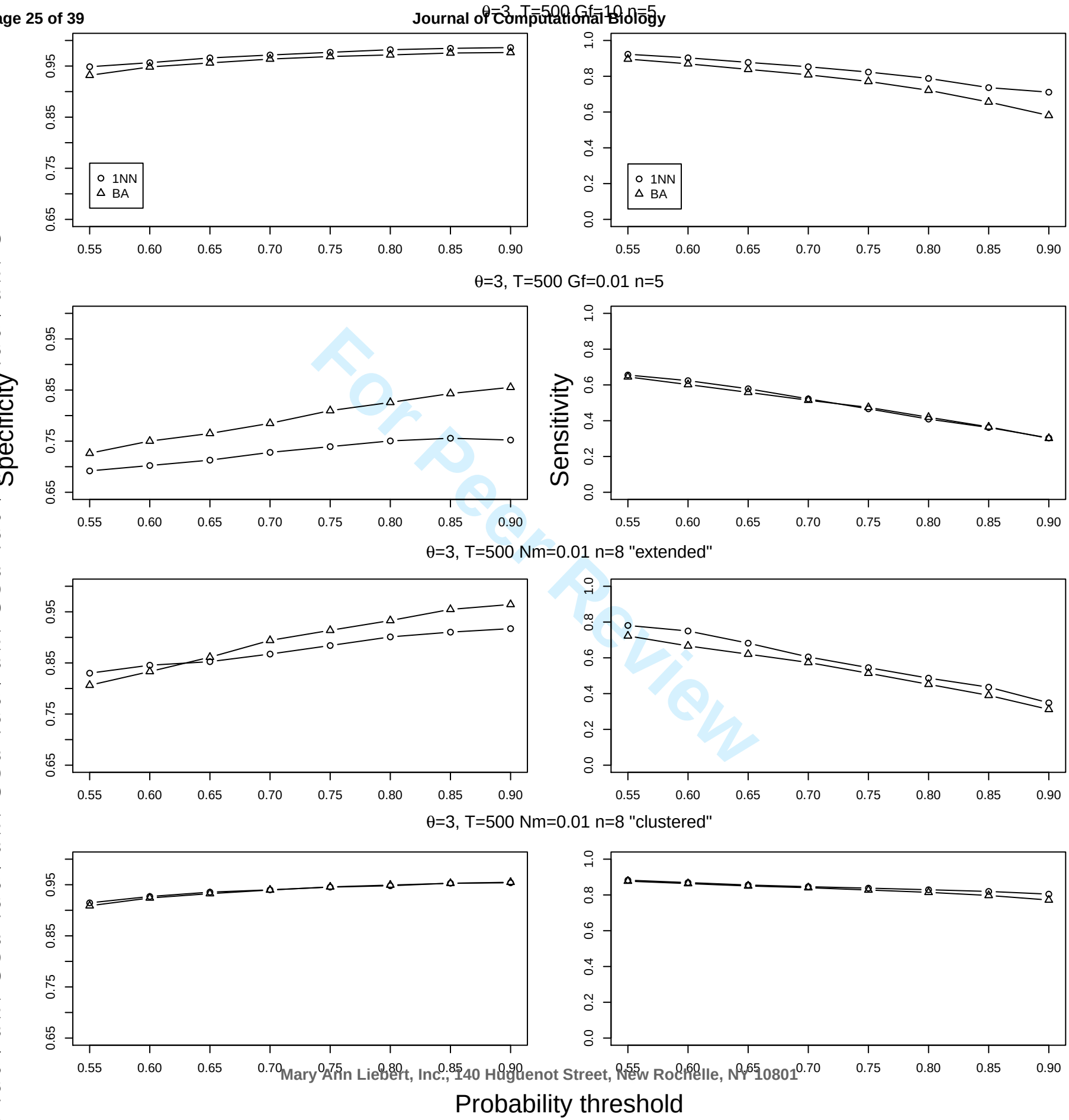
Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7, 256–276.

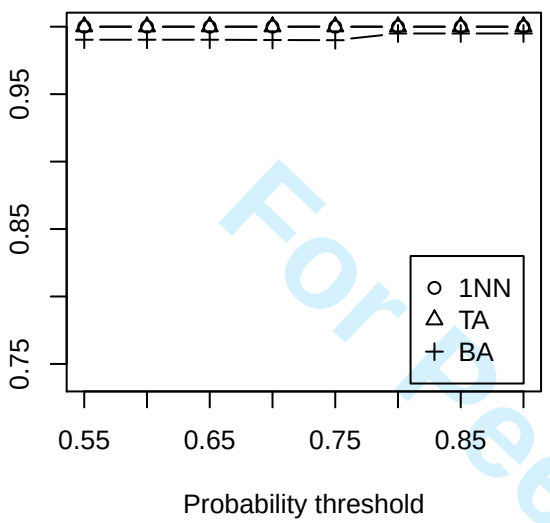


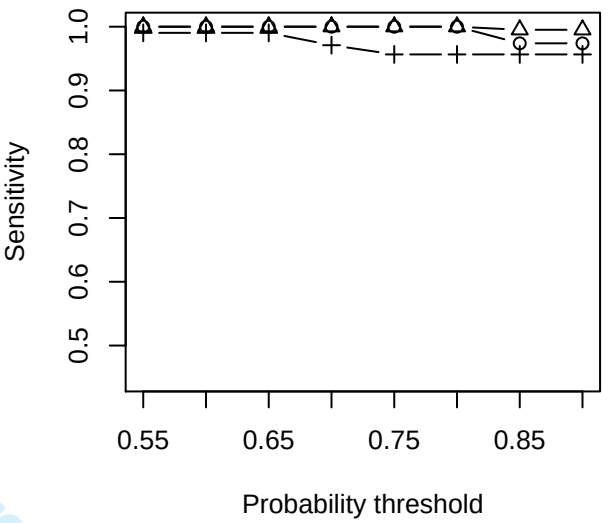


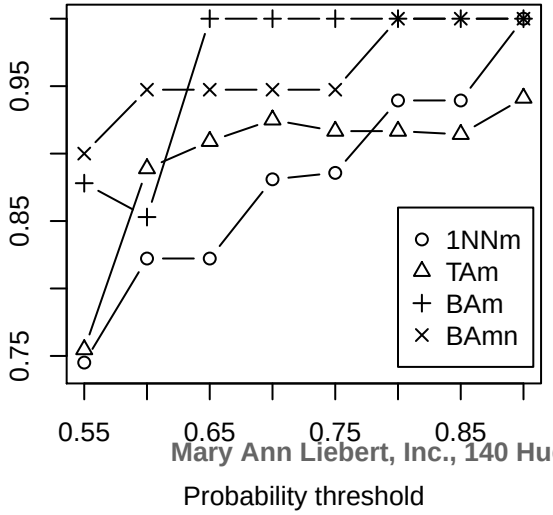


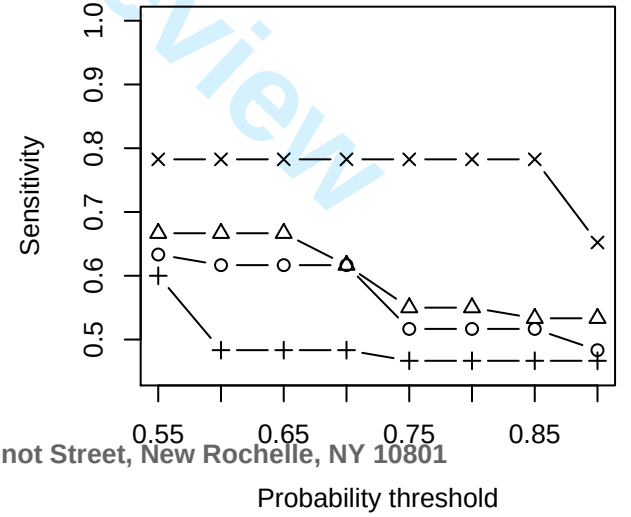












Supplementary materials for Coalescent-based DNA barcoding: multilocus analysis and robustness by O. David, C. Larédo, R. Leblois, B. Schaeffer and N. Vergne

A Mathematical appendix

Equation (2). According to Bayes theorem, the probability $P(x|y_i)$ can be written as:

$$P(x|y_i) = P(x, y_i)/P(y_i).$$

As we need to distinguish the gene of the test individual, samples of genes such as $\{x, y_i\}$ or y_i are here considered as ordered. However we wish to relate $P(x|y_i)$ to probabilities of unordered samples as efficient methods have been developed to estimate such probabilities. Indeed samples are usually considered as unordered for statistical inference in population genetics since allelic multiplicities are usually the only useful information in a sample and the fact that an allele is carried by a particular individual is usually not informative (Section 5, Tavaré and Zeitouni, 2004; Stephens and Donnelly, 2000). The probability $P(y)$ of an ordered sample y is related to the probability $P^0(y)$ of the corresponding unordered sample by the following equation (Section 5, Tavaré and Zeitouni, 2004):

$$P(y) = \frac{n_1! \dots n_h!}{n!} P^0(y),$$

where the sample is assumed to include n sequences and h haplotypes with multiplicities $n_1, \dots, n_h$. These equations lead to (2).

Importance sampling. The probabilities $P^0(x, y_i)$ and $P^0(y_i)$ in (2) can be estimated using importance sampling (IS) (De Iorio and Griffiths, 2004). In this method, $P^0(y)$ is written as (Stephens and Donnelly, 2000):

$$P^0(y) = \sum_H P(y|H) \frac{P(H)}{Q(H)} Q(H).$$

In this formula, H is a coalescent history, that is a series $H = (H_{-k}, k = 1, \dots, m)$, where H_{-k} is the vector of the multiplicities of the genetic types of the sample after the kth event affecting the genealogy backward in time (i.e. coalescence or mutation)

and H_{-m} corresponds to the genetic type of the most recent common ancestor of the sample. The distribution Q is a proposal distribution on such coalescent histories. An efficient proposal distribution is given in De Iorio and Griffiths (2004) and is uniform on possible history changes for genes back in time. The probabilities $P(y|H)$ and $P(H)$ have explicit expressions that depend on mutation parameters that are first assumed to be known (De Iorio and Griffiths, 2004). The probability $P^0(y)$ is thus estimated by:

$$\widehat{P}^0(y) = \frac{1}{M}\sum_{k=1}^{M} P(y|H^{(k)})\frac{P(H^{(k)})}{Q(H^{(k)})},$$

where $H^{(1)}, \ldots, H^{(M)}$ are independent samples from Q. The proposal distribution Q improves the estimator of $P^0(y)$ because it allows to only simulate histories that are compatible with the data.

Predictive approach. When mutation parameters are not known, a predictive approach can be used in which the dependence of probabilities on θ is removed by integration. The IS method can then be adapted as follows. The probability $P^0(y_i)$ is written as:

$$P^0(y_i) = \sum_H \int P(y_i|H,\theta_i)\frac{P(H|\theta_i)}{Q(H|\theta_i)}Q(H|\theta_i)\pi(\theta_i)d\theta_i,$$

where π is the prior distribution of θ_i. Note that the proposal distribution of De Iorio and Griffiths (2004) does not depend on θ_i and can be written as $Q(H)$ instead of $Q(H|\theta_i)$. The probability $P^0(y_i)$ is then estimated by:

$$\widehat{P}^0(y_i) = \frac{1}{M}\sum_{k=1}^{M} P(y_i|H^{(k)},\theta_i^{(k)})\frac{P(H^{(k)}|\theta_i^{(k)})}{Q(H^{(k)}|\theta_i^{(k)})},$$

where $\theta_i^{(1)}, \ldots, \theta_i^{(M)}$ are independent samples from the prior π. This estimator could be further improved by sampling the values of θ_i using a proposal distribution rather than the prior π. This predictive approach better takes account of the uncertainty on θ than the plug-in method.

Prior distribution of mutation parameters. As for the prior π, the parameters θ_i are assumed to be independent a priori and to follow a gamma prior $G(a,b)$ with density function $G(a,b,x) = b^a x^{a-1}e^{-bx}/\Gamma(a)$. An empirical Bayes approach is used and the parameters a and b are estimated from the reference data (Chapter 5, Carlin and Louis, 2008). The number of polymorphic sites of species i S_i satisfies:

$$\begin{aligned} E(S_i|a,b) =& a/b\, w_{1i}, \\ var(S_i|a,b) =& a/b\, w_{1i} + a/b^2\, w_{1i}^2 + a/b(a/b+1/b)\, w_{2i}, \end{aligned}$$

where $w_{1i} = \sum_{j=1}^{n_i-1} 1/j$ and $w_{2i} = \sum_{j=1}^{n_i-1} 1/j^2$. The quantities S_i are assumed to be independent. The estimating equations estimators of $(a/b, 1/b)$ are the solution to

the following system of equations (p. 67 and 68, Huet et al., 2004):

$$\sum_{i=1}^{c} \frac{\partial E(S_i)}{\partial a/b} \times \frac{S_i - E(S_i)}{var(S_i)} = 0,$$

$$\sum_{i=1}^{c} \frac{\partial var(S_i)}{\partial 1/b} \times \frac{(S_i - E(S_i))^2 - var(S_i)}{var(S_i)^2} = 0.$$

This system of equations can be solved numerically using the nls2 package of R (Huet et al., 2004; R Development Core Team, 2005).

B Simulation study

B.1 Data set simulation

Simulations were carried out to quantify the precision of DNA barcoding methods. In these simulations, one ancestral species split into two new species T generations ago. Sequences were simulated for a mitochondrial locus and a nuclear diploid locus that evolved in the ancestral and the new species according to a coalescent model independently. There was no recombination within the nuclear locus. The mutation model was the ISM with a mutation parameter θ that was the same for both loci and for the ancestral and the two new species. The ancestral and the new species had the same effective size N_e. The effective size was equal to $N_e = 1000$ for the mitochondrial locus and $N_e = 4000$ for the nuclear locus. N_e was four times smaller for the mitochondrial locus because mitochondrial genes are in a single copy and transmitted by females only. We simulated 2000 data sets with n reference individuals in each species and one test individual. The data set simulations were performed with the program ms (Hudson, 2002), considering four combinations of parameter values that led to imperfect assignments: $(\theta = 3, T = 500, n = 5)$, $(\theta = 3, T = 1000, n = 5)$, $(\theta = 3, T = 500, n = 10)$ and $(\theta = 20, T = 500, n = 5)$. Considering demographic situations leading to imperfect assignments allowed us to better compare the performance of the different methods.

To test the robustness of the methods developed, we also considered variable past population sizes and population structures, two factors that often occur in natural populations. We designed the simulations so that the results were easily comparable with the first set of simulations described above and we simulated mitochondrial sequences only. Thus for variable population size, we simulated an exponential change in one of the new species (i.e. from present to T going backward in time), so that the new species at present and the ancestral species had the same size as in the previous simulations (i.e., $N_e = 1000$). The population size change was characterized by its growth factor Gf and its duration (T generations). The population size was given by $N_e(t) = N_e \exp(-\alpha t)$, where t was the time before the present, measured in units of $2N_e$ generations, and $\alpha = \ln(Gf)\, 2N_e/T$. The population size change effect was thus a founder effect at divergence time T when the size of

the new species was then N_e/Gf. A growth rate larger than one thus indicated a population expansion from divergence to present, whereas a growth rate smaller than one indicated a population decline. The growth rate values we considered were $Gf = \{0.001, 0.01, 0.1, 10, 100\}$ and the test individual was taken from the species with a variable size. The other parameters were set to the baseline values used above, that is $(\theta = 3, T = 500, n = 5)$.

For the population structure simulations, we considered stepping stone models, with 4 populations within each new species, exchanging migrants, with their neighbouring populations only, at rate m. The size of these populations was set to $N_e = 1000/4$ so that the total size of each species was 1000 as in the baseline simulations. The $N_e m$ values, the number of migrants exchanged at each generation between two adjacent populations, were set to $\{0.001, 0.01, 0.1, 1, 10, 100\}$. There was no migrant exchange between populations of the different species, which thus remained completely isolated. Then, from divergence to the most recent common ancestor, (i) the ancestral species was composed of the 8 subpopulations present in the new species, with a size of $N_e = 1000/8$ for each subpopulation so that the total size was 1000; and (ii) all migration rates were set to $N_e m = 100$ to limit the effect of structure in the ancestral species. To mimic extreme sampling strategies that can be observed in the barcoding data base, we considered an "extended" sampling, in which two reference individuals were sampled in each population of the two new species, and a "clustered" sampling, in which all reference individuals were sampled in the most adjacent populations of the two new species (i.e., populations 4 and 5 if populations are linearly labelled from 1 to 8). The test individual was always taken from the most distant population in species two (i.e., population 8). The other parameters values were set to $(\theta = 3, T = 500, n = 8)$.

B.2 Analysis

The simulated data were analysed with the nearest-neighbour classification (1NN) and the developed Bayesian assigner (BA). The nearest-neighbour method assigned a test individual to the species of its nearest neighbour. It was used because it required little computing time, it was known to be efficient (Austerlitz et al., 2009) and it was expected to be robust to changes in the evolutionary scenario since it was not based on a specific biological model. The method of Abdo and Golding (2007) was not used in these simulations to limit computing time. The program of Munch et al. (2008) was not used because we were not able to make it work on our computers. The prior probabilities of membership were equal in all analyses. For each simulation, the test individual was assigned to the species with the largest probability of membership, if the latter probability exceeded some specified probability threshold. Otherwise, the assignment was considered as ambiguous.

The 1NN classification was implemented with bagging in order to obtain a measure of confidence for an assignment (Hastie et al., 2001). It was applied to 200 bootstrap samples of the reference data. Bootstrapping within species rather than globally gave similar results (results not shown). The probability that a test in-

Data set	Species	Individuals	Haplotypes	Polym. sites
Astraptes group A	4	115	17	18
Astraptes group B	3	92	8	23
Ithomiinae Forbestra	2	11 (8)	4 (3)	3 (2)
Ithomiinae Hypothyris	3	20 (8)	10 (3)	34 (6)
Ithomiinae Melinaea	3	29 (7)	13 (5)	15 (2)

Table S1. Descriptive statistics of the five data sets analyzed. For the Ithomiinae data, the numbers without/with brackets relate to the mitochondrial/nuclear locus.

dividual belonged to a species was estimated for a given bootstrap sample by the proportion of individuals belonging to that species among the nearest neighbours. Then the global probability of membership to that species was estimated by the average probability of membership over the bootstrap samples. The genetic distance between two mitochondrial sequences was the number sites that differed between the sequences. As no alternative method to BA existed to analyse the nuclear sequences, a version of the 1NN method that could analyse such data was developed. The genetic distance between a pair of test nuclear alleles and a reference nuclear sequence was the sum of the genetic distances between each test allele and the reference sequence. The mitochondrial and nuclear assignments were combined using (3) in which the probabilities $P(x_j|y_{ij}, z = i)$ were estimated by bootstrap as explained above.

For BA, the probabilities $P^0(x, y_i)$ and $P^0(y_i)$ in (2) were estimated using the program genetree (Bahlo and Griffiths, 2000) with $500,000$ IS simulations. The ancestral sequence was assumed to be known. The value of θ was either known or estimated for each species using the method of Watterson (1975).

As expected, specificity increased in our results as the probability threshold increased because assignments with a low assignment probability were less reliable. On the other hand, sensitivity decreased as the probability threshold increased because the number of ambiguous assignments increased with the probability threshold. Figure S1 shows the effect of estimating mutation parameters on the BA perfomance. Figures S2 and S3 show the results for the nuclear data.

C Analysis of Astraptes and Ithomiinae data

The first data set used came from the study of Hebert et al. (2004) on Astraptes species. The second data set used came from the study of Elias et al. (2007) on Ithomiinae species. As the ISM was not compatible with the whole data sets, only five subsets of the data containing sequences from closely related species were analyzed (Table S1). The sites that had missing data or that were not polymorphic were removed. In addition, 12 sites were removed from the mitochondrial Ithomiinae Hypothyris data, and 1 site was removed from the nuclear Ithomiinae Hypothyris data because these sites were not compatible with the ISM.

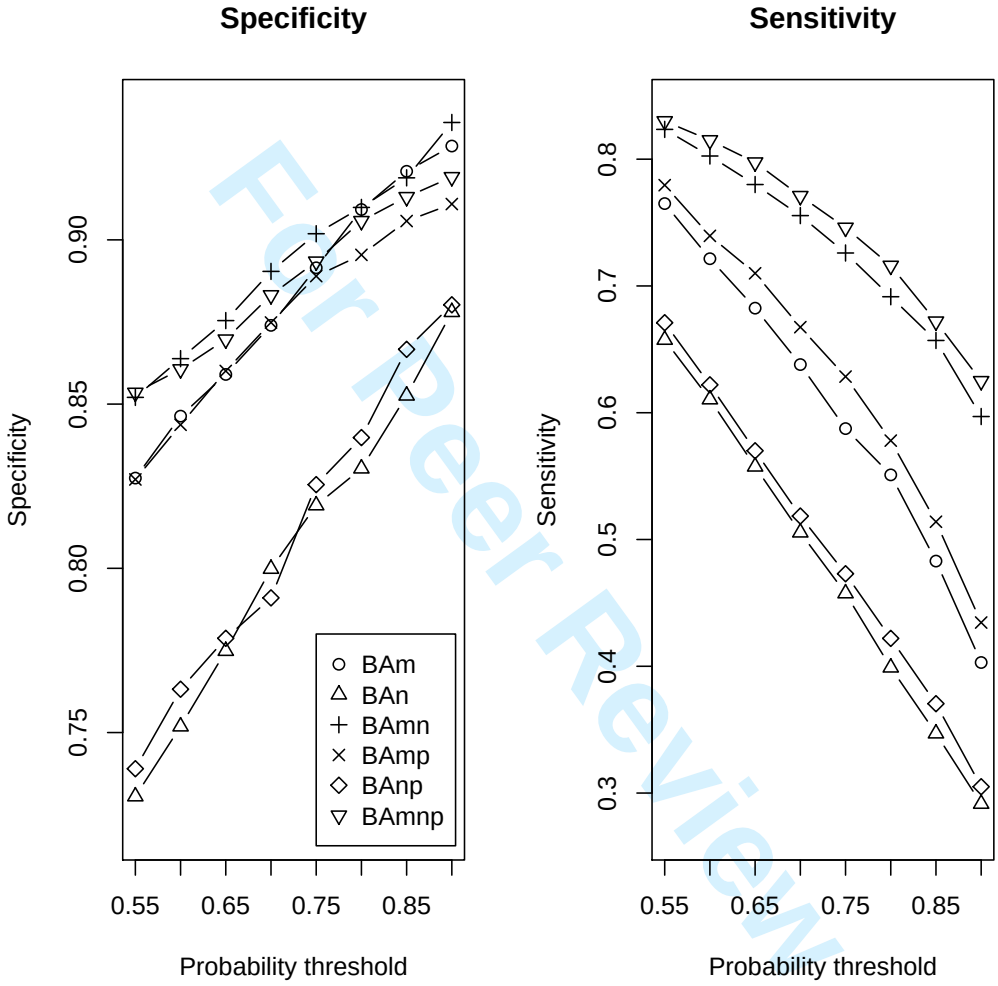


Figure S1. *Effect of estimating mutation parameters on the performance of coalescent-based barcoding.* The data were simulated with $\theta = 3$, $T = 500$ and $n = 5$. BA is the developed Bayesian assigner. The subscripts m, n and mn denote the mitochondrial data, the nuclear data and the combination of these data. The subscript p indicates a plug-in method with θ estimated using Watterson's estimator. Coalescent-based barcoding appears not very sensitive to the uncertainty on mutation parameters.

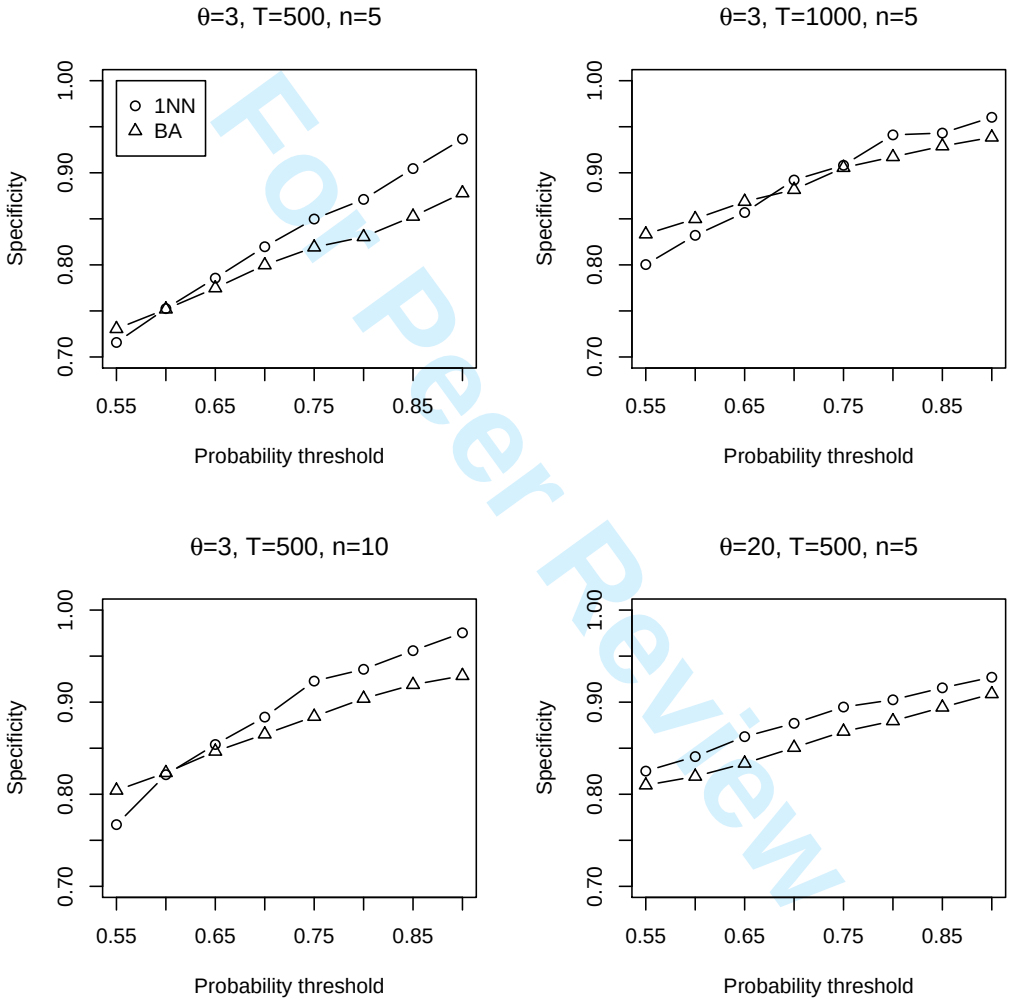


Figure S2. *Results for specificity with the nuclear data.* 1NN and BA are the nearest-neighbour classification and the developed Bayesian assigner with a known value of θ.

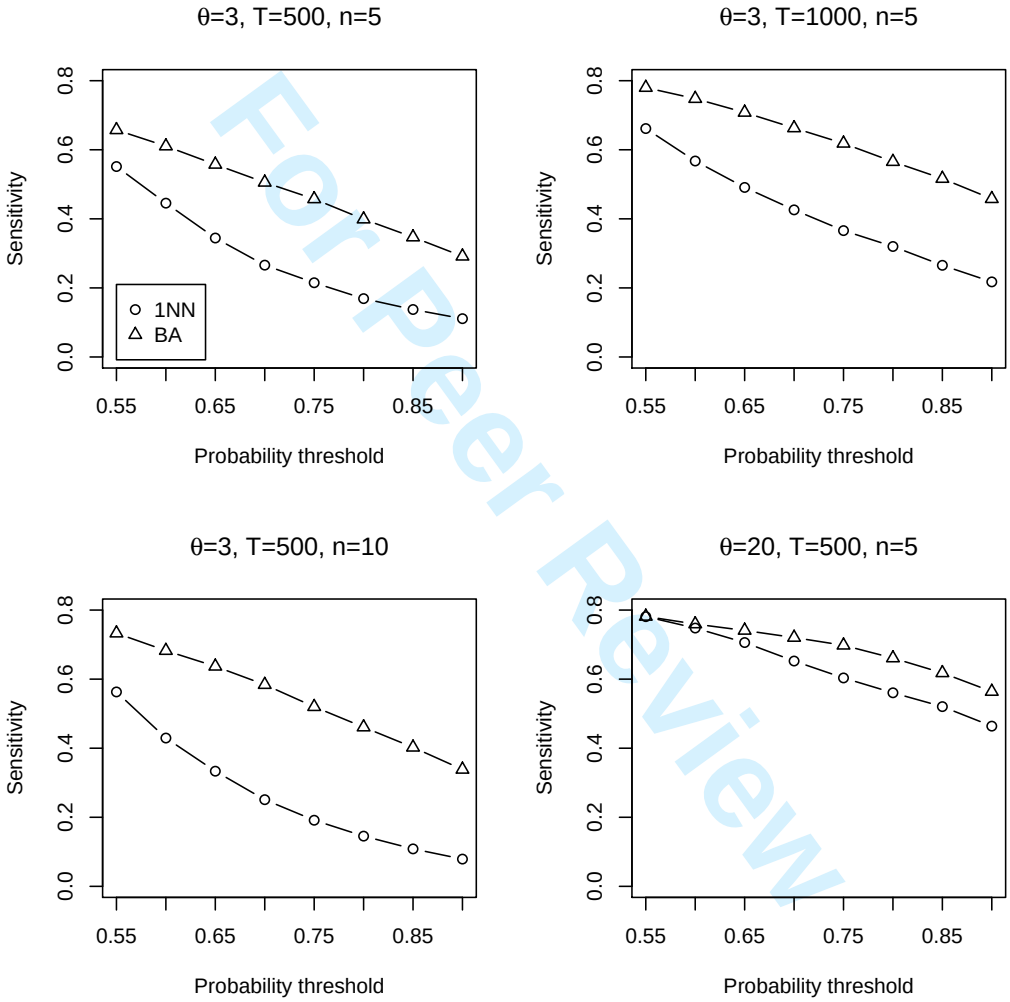


Figure S3. *Results for sensitivity with the nuclear data.* 1NN and BA are the nearest-neighbour classification and the developed Bayesian assigner with a known value of θ.

The data were analysed with TheAssigner (Abdo and Golding, 2007), 1NN and BA. The program of Munch et al. (2008) was not used because we were not able to make it work on our computers. The nuclear data were modeled with a haploid model because they contained one sequence per individual only. The prior probabilities of membership were equal in all analyses.

For BA, the probabilities $P^0(x, y_i)$ and $P^0(y_i)$ in (2) were estimated using the program genetree (Bahlo and Griffiths, 2000) with $150,000$ IS simulations. The parameter θ was estimated for each species using the method of Watterson (1975). The ancestral base at a polymorphic site was chosen equal to the most frequent base.

The 1NN method was implemented with bagging as in the simulation study. The number of bootstrap samples was equal to 200. The genetic distance between two sequences was the number sites that differed between the sequences.

TheAssigner was used with the following parameter values: sample_mcmc = 10000, thin_mcmc = 20 and burn_in_mcmc = 3000. The mutation model was F81. The per-site mutation parameter was the estimate of θ divided by sequence length.

The sensitivity results for the nuclear Ithomiinae data are shown in Figure S4. All the methods had a specificity of one for these data and a probability threshold greater than or equal to 0.55.

References

Abdo, Z., Golding, G.B., 2007. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. Systematic Biology 56, 44–56.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., Larédo, C., 2009. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics, Special Issue Biodiversity Informatics .

Bahlo, M., Griffiths, R.C., 2000. Inference from gene trees in a subdivided population. Theor. Popul. Biol. 57, 79–95.

Carlin, B.P., Louis, T.A., 2008. Bayesian Methods for Data Analysis. Texts in Statistical Science, Chapman & Hall/CRC. third edition.

De Iorio, M., Griffiths, R.C., 2004. Importance sampling on coalescent histories. I. Adv. Appl. Prob. 36, 417–433.

Elias, M., Hill, R.I., Willmott, K.R., Dasmahapatra, K.K., Brower, A.V., Mallet, J., Jiggins, C.D., 2007. Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc R. Soc. B. 274, 2881–9.

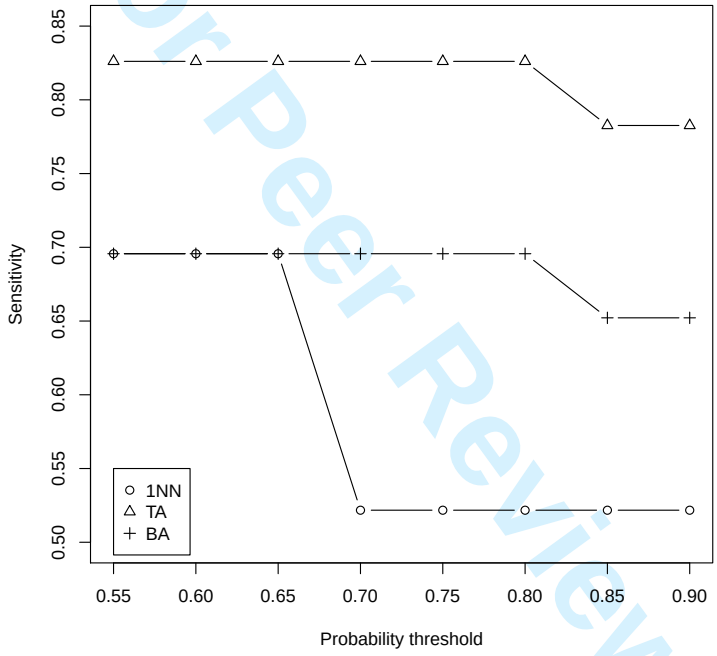


Figure S4. *Sensitivity results for the nuclear Ithomiinae data.* 1NN, TA and BA are the nearest-neighbour classification, TheAssigner and the developed Bayesian assigner.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics, Springer.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004. Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America 101, 14812–14817.

Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18, 337–338.

Huet, S., Bouvier, A., Poursat, M.A., Jolivet, E., 2004. Statistical tools for nonlinear regression : a practical guide with S-PLUS and R examples. Springer Series in Statistics, Springer-Verlag. second edition.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E., Nielsen, R., 2008. Statistical assignment of DNA sequences using bayesian phylogenetics. Systematic Biology 57, 750–757.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. R. Stat. Soc., Ser. B 62, 605–655.

Tavaré, S., Zeitouni, O., 2004. Lectures on probability theory and statistics : Ecole d'été de probabilités de Saint-Flour XXXI - 2001. Lecture notes in mathematics, Springer.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7, 256–276.

Mailing address and contact information of each author

Olivier David
MIA unit
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France
olivier.david@jouy.inra.fr
Phone : (+ 33) (0)1 34 65 28 44
Fax : (+ 33) (0)1 34 65 22 17

Catherine Larédo
MIA unit
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France
catherine.laredo@jouy.inra.fr
Phone : (+ 33) (0)1 34 65 22 26
Fax : (+ 33) (0)1 34 65 22 17

Raphael Leblois
CBGP
Campus International de Baillarguet CS 30016
34988 Montferrier-sur-Lez cedex
France
raphael.leblois@supagro.inra.fr
Phone : (+ 33) (0)4 99 62 33 31
Fax : (+33) (0)4 99 62 33 45

Brigitte Schaeffer
MIA unit
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France
brigitte.schaeffer@jouy.inra.fr
Phone : (+ 33) (0)1 34 65 22 18
Fax : (+ 33) (0)1 34 65 22 17

Nicolas Vergne
Laboratoire de Mathématiques Raphaël Salem
UMR 6085 CNRS-Université de Rouen
Avenue de l'Université, BP.12
Technopôle du Madrillet
76801 Saint-Étienne-du-Rouvray
France
nicolas.vergne@univ-rouen.fr
Phone : (+ 33) (0)2 32 95 52 49
Fax : (+ 33) (0)2 32 95 52 86