



**HAL**  
open science

# Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study

Mohamed Anwar Abouabdallah, Nathalie Peyrard, Alain Franc

## ► To cite this version:

Mohamed Anwar Abouabdallah, Nathalie Peyrard, Alain Franc. Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study. *Molecular Ecology Resources*, 2022, 10.1111/1755-0998.13579 . hal-03546609

**HAL Id: hal-03546609**

**<https://hal.inrae.fr/hal-03546609>**

Submitted on 4 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study

Mohamed Anwar Abouabdallah<sup>1,2</sup>, Nathalie Peyrard<sup>3,\*</sup>, and Alain Franc<sup>1,2</sup>

<sup>1</sup>Université de Bordeaux, INRAE, BIOGECO, 33612 Cestas, France

<sup>2</sup>Pleiade, EPC INRIA-INRAE-CNRS, Université de Bordeaux, 33405 Talence, France

<sup>3</sup>Université de Toulouse, INRAE, UR MIAT, 31320 Castanet-Tolosan, France

\*corresponding author: [nathalie.peyrard@inrae.fr](mailto:nathalie.peyrard@inrae.fr)

<sup>1</sup> **Running title:** Clustering plants' barcodes at order level

## Abstract

Characterising biodiversity is one of the main challenges for the coming decades. Most diversity has not been morphologically described and barcoding is now complementing morphological-based taxonomy to further develop inventories. Both approaches have been cross-validated at the level of species and OTUs. However, many known species are not listed in reference databases. One path to speed up inventories using barcoding is to directly identify individuals at coarser taxonomic levels. We therefore studied in barcoding of plants whether morphological-based and molecular-based approaches are in agreement at genus, family and order levels. We used Agglomerative Hierarchical Clustering (with Ward, Complete and Single Linkage) and Stochastic Block Models (SBM), with two dissimilarity measures (Smith-Waterman scores, kmers). The agreement between morphological-based and molecular-based classifications ranges in most of the cases from good to very good at taxonomic levels above species, even though it decreases when taxonomic levels increase, or when using the tetramer-based distance. Agreement is correlated with the entropy of morphological-based classification and with the ratio of the mean within- and mean between-groups dissimilarities. The Ward method globally leads to the best agreement whereas Single Linkage can show poor behaviours. SBM provides a useful tool to test whether or not the dissimilarities are structured by the botanical groups. These results suggest that automatic clustering and group identification at taxonomic levels above species are possible in barcoding.

**Keywords:** taxonomy; barcoding; clustering; Stochastic Block Model; Ward method; French Guianan Trees

## 1 Introduction

Numerical taxonomy and hierarchical clustering have coevolved since the 1960s' (Cole, 1969; Sneath and Sokal, 1973). Both approaches rely on the assumption that the diversity

28 of life for taxonomy, or patterns in distances between some items in clustering, are organized  
29 as a nested hierarchy, modelled as a tree. This approach has survived the revolution of  
30 molecular-based taxonomy (Hillis et al., 1996) and molecular phylogenies (Felsenstein, 2004;  
31 Yang, 2006), with a current revival due to barcoding (Floyd et al., 2002; Hebert et al., 2003),  
32 and metabarcoding (López-García et al., 2001; Sogin et al., 2006; Hajibabaei et al., 2011;  
33 Taberlet et al., 2012; Kermarrec et al., 2013). As far as morphological-based taxonomy is  
34 concerned, most of the diversity in many clades of organisms is still unknown. Leray and  
35 Knowlton (2015) point out that between 33% and 91% of all marine biodiversity has never  
36 been named. Currently many effort are devoted to speeding up the process of producing  
37 large inventories with metabarcoding by bypassing identified obstacles (Bik et al., 2012).

38 The notion of OTU (Operational Taxonomic Unit) has been coined (Floyd et al., 2002;  
39 Blaxter et al., 2005). Such units are produced by clustering sets of barcodes by aggregation  
40 at a level assumed to be similar to the level of species in morphological-based classifications.  
41 The authors in Blaxter et al. (2005) emphasize that they are "agnostic" as to whether OTU  
42 are species or not. Identifying OTUs in an environmental sample and organising molecular  
43 diversity as the frequency of OTUs make it possible to produce molecular-based inventories  
44 at previously unparalleled speed.

45 A classical approach is therefore to build OTUs and to map them on reference databases  
46 that contain reference barcodes. A standard tool for mapping is BLAST (Altschul et al.,  
47 1990), but other more sophisticated solutions exist (e.g., the use of Bayesian Phylogenetics,  
48 Munch et al., 2008). When taxonomic expertise and references exist at the species level,  
49 the agreement between molecular and morphological-based classification can be excellent  
50 (Ji et al., 2013), even if sometimes like for plants, introgression may blur the distinction  
51 between species (Petit and Excoffier, 2009). It may happen that such a comparison is not  
52 feasible when morphological-based taxonomy is unknown or when only partial references  
53 exist. Leray and Knowlton (2015) report in their study that less than 12% of their OTUs  
54 matched with GenBank or BOLD. The same observation was made in White et al. (2010)

55 regarding intestinal microbial flora. Hence most inventories with supervised learning are  
56 made at a grain often much coarser than the genus/species level.

57 Trying to complete databases at the species level is highly time-consuming. Another  
58 solution is to build groups larger than OTUs, e.g. at the scale of families or orders, by  
59 clustering<sup>1</sup> the barcodes. Then each group could be annotated as a taxon at this higher  
60 taxonomic level by looking for a match for one or several sequences of the group, in the  
61 reference database. This is in line with the conclusion of the study by Meiklejohn et al.  
62 (2019), on the accuracy of BOLD and GenBank: the authors suggest that a solution to  
63 address concerns with incorrect species identifications observed in their experiments would  
64 be to report the taxonomy at a higher level. This raises the question of the agreement  
65 between morphological-based and molecular-based taxonomy when clusters of sequences  
66 are built at a level coarser than species, e.g., class or order. Comparing morphological-  
67 based classifications and OTUs produced by barcode clustering has been thoroughly studied  
68 (see, e.g., White et al., 2010). Several methods have been recently designed and widely  
69 used for delineating species on the basis of barcodes (Pons et al., 2006; Fontaneto et al.,  
70 2008; Puillandre et al., 2012; Talavera et al., 2013; Zhang et al., 2013). However, to our  
71 knowledge, the question has seldom been addressed directly at coarser taxonomic levels  
72 such as orders.

73 Our objective here is to study whether the clustering of barcodes in molecular-based  
74 taxonomy makes it possible to directly recover the taxa present in a sample, for a given  
75 taxonomic level coarser than species, and, if so, with which tool, accuracy and robustness.  
76 More precisely, we consider the clustering of the barcodes in a reduced number of groups  
77 compared to a clustering into species, and we ask the question whether the classification  
78 obtained is similar or not with the botanical classification at genus, family or order levels.  
79 This comparison is performed without annotating the classes: we only aim at comparing

---

<sup>1</sup>In this article, the term *clustering* makes reference to any numerical method for the unsupervised grouping of the individuals, while the term *classification* designates the method's output, i.e. the partition of individuals into classes.

80 the two partitions of the sequences, the botanical one and the molecular-base one.

81 We have selected for this study a dataset of barcodes of trees in "Piste de Sainte Elie"  
82 research station in French Guiana. The corresponding plot has been inventoried botanically  
83 for decades (Madelaine et al., 2007). The data set represents about one third of the diversity  
84 of the French Guianan tree flora (1458 sequences, from 20 orders, 56 families, 182 genera and  
85 428 species) . We selected flowering plants because the botanical classification is well known,  
86 both morphologically (it is organised as a nested system of different taxonomic levels as  
87 a classification system) and molecularly with the Angiosperm Phylogeny Group initiative,  
88 even if it is under continuous revision (The Angiosperm Phylogeny Group et al., 2016).  
89 The dataset itself is composed of some 1,500 trees from French Guiana that have been  
90 botanically identified and sequenced with chloroplastic marker *trnH-psbA* using Sanger  
91 technology which produces high quality sequences (Caron et al., 2019). By selecting a  
92 small data set and a long resolutive sequence (*trnH-psbA* is about 450 bp long, with high  
93 variability), we are not confronted to the computational burden of treatment of massive  
94 data sets as in metabarcoding data, and we can therefore concentrate on the analysis of  
95 agreement. The question of the scaling to metabarcoding with massive data sets of shorter  
96 reads of the clustering methods will be the object of further studies.

97 It can be expected that there is not a clear answer to the degree of agreement between  
98 the two types of classification (morphological-based or molecular-based). There may be  
99 favourable situations where the agreement is strong, and others where the two classifica-  
100 tions are surprisingly quasi-independent of each other. Moreover this can depend on the  
101 taxonomic level. To identify potential factors that may explain variations in agreement in  
102 our study: (i) we varied the taxonomic level at which the clustering is performed (order,  
103 family, genus, species), (ii) we used two definitions of dissimilarity between sequences; and,  
104 finally, (iii) we considered four numerical methods for the clustering of the molecular data.  
105 Altogether, this leads to 32 possible combinations

106 More specifically, we first worked with 30 non random sub-samples of the whole dataset,

107 each sub-sample comprising either all the individuals of an order or of a family. In each case,  
108 we compared the botanical classification of the individuals at the next finer taxonomic level  
109 with the molecular-based classifications. In a second step, we studied whether the mean  
110 behaviour observed from these replicates is recovered when the set of individuals to be  
111 classified is larger and more diverse, by comparing the botanical classification of the whole  
112 dataset into orders with the molecular-based classification for the same number of classes.  
113 We also performed the comparison at the family, gender and species levels.

114 Dissimilarities between sequences have been computed as edit distances (Levenstein,  
115 1966; Gusfield, 1997). The score of local pairwise alignment (Smith and Waterman, 1981)  
116 has been preferred to global pairwise alignment (Needleman and Wunsch, 1970) to avoid the  
117 cost of slight lengths variations due to technological reasons in Sanger sequencing (Gusfield,  
118 1997). Even if this algorithm relies on dynamic programming, thus making it very efficient  
119 (and exact), its complexity is in  $\mathcal{O}(n^2\ell^2)$  if  $n$  is the number of barcodes or reads, and  $\ell$   
120 their length. This becomes prohibitive for large datasets. A classical way to circumvent  
121 this difficulty is to use kmer-based distances (Sun et al., 2009; Mahé et al., 2014), a priori  
122 with a decrease in the quality of the estimation of the dissimilarity, but much faster to  
123 compute. A comparison between Smith-Waterman scores and kmer-based distances can  
124 be found in Sun et al. (2009). The question here is to explore whether the loss in quality  
125 remains acceptable and does not lead to a decrease in agreement between the botanical and  
126 the molecular-based classifications. This is a preliminary step for developing further studies  
127 on metabarcoding which require investment in scaling and accelerating the computation of  
128 distances.

129 If the morphological-based taxonomic classification is a priori unique, this is not true  
130 for a molecular-based classification. A diversity of softwares for implementing hierarchical  
131 clustering has been proposed for more than a decade in metabarcoding with the objective  
132 of efficient scaling with respect to the growing size of environmental datasets. This in-  
133 cludes Uclust (Edgar, 2010), ESPRIT (Sun et al., 2009) and SWARM (Mahé et al., 2014,

134 2015, 2019), which make it possible to cluster millions of barcodes on a laptop. Nearly  
135 all of the hierarchical clustering algorithms mentioned above rely at one step or another  
136 on heuristics (like computing kmer-based distances, considering short distances only i.e.  
137 working with sparse distance matrices) to make computation feasible within a reasonable  
138 time with reasonable memory. SWARM uses kmers only as a first step to filter out pairs  
139 of sequences which are distant and cannot belong to a same compact community. In this  
140 study, we focus on understanding the agreement (or not) between molecular-based clas-  
141 sification from clustering and botanical classification, without computational constraints.  
142 We therefore consider Aggregative Hierarchical Clustering (AHC), whose above-mentioned  
143 algorithms can be seen as heuristic versions for scaling up, with three different aggrega-  
144 tion methods: Single Linkage, Complete Linkage, Ward (Murtagh, 1983; Müllner, 2013).  
145 Statistical models like Bayesian classifiers with mixture models have also been considered  
146 in the literature to cluster sequences (Hao et al., 2011). To extend the scope of statistical  
147 modeling in molecular-based taxonomy, we explore here the potential interest of a model-  
148 based clustering method, the Stochastic Block Model (SBM, Holland et al., 1983; Daudin  
149 et al., 2008; Lee and Wilkinson, 2019) as an alternative to AHC. SBMs are already widely  
150 applied with success in domains like the social sciences (Barbillon et al., 2017), the anal-  
151 ysis of ecological interaction networks (Miele and Matias, 2017) and neurology (Faskowitz  
152 et al., 2018). They rely on a more flexible definition of a cluster than AHC (searching for  
153 general groups and not just communities), and we hypothesised that SBM and AHC could  
154 be complementary in their capacity to distinguish meaningful groups of individuals in an  
155 inventory.

156 In the following section, we provide a brief description of the dataset. We also de-  
157 scribe the method. Results on the quality of the agreement between molecular-based and  
158 morphological-based classifications obtained on replicates are presented in Section 3.2, the  
159 results obtained on the whole dataset are presented in Section 3.3.



## 2 Materials and methods

### 2.1 Dataset and computation of dissimilarities

This study relies on a dataset built from a collection of some 1,500 trees located in the "Piste de Saint-Elie" experimental plot in French Guiana, mainly composed of lowland tropical rainforest (Sabatier et al., 1997). The data used here are part of a dataset gathered for the study published in Caron et al. (2019), which focused on agreement or not between botanical-based and molecular-based classification at the species level over a wide range of diversity along the angiosperms tree. The main result in Caron et al. (2019) is that molecular-based clustering is highly consistent with species delineation in a majority of cases, and that introgression or incomplete lineage sorting are the most likely explanations in the case of non-agreement. We focus here on a similar question but at the level of genera, families and orders. The main elements for the material are recalled here, and the reader can refer to Caron et al. (2019) for details. Among this dataset, 1,458 individual trees were selected for this study. For each tree, we used the botanical name as given by field botanists working with the Cayenne Herbarium of the French National Research Institute for Sustainable Development, and a sequence of chloroplastic marker *trnH-psbA*, which is highly resolute, despite the fact that it is variable in length. This drawback is mitigated because no multiple alignment is done: we work with pairwise distances only, computed either by local alignment or comparison of histogram of tetramer histograms. *trnH-psbA* has been used in several studies or benchmarks in plant metabarcoding (Hollingsworth et al., 2009, 2011; Pang et al., 2012). These trees encompass 20 orders, 56 families, 182 genera and 428 species.

Three  $1458 \times 1458$  matrices of pairwise distances or dissimilarities between sequences were built, a first one using the Smith-Waterman algorithm for local sequence alignment (Smith and Waterman, 1981), and two other ones for the distance between kmers distributions

186 ( $k = 4$  and  $k = 6$ ). The local alignment score is the most precise quantification of genetic  
187 dissimilarities between sequences, but it is time consuming. Several methods for building  
188 OTUs therefore rely on alternatives to local alignment scores. A classical way to circumvent  
189 this computational burden is to build kmer counts for each sequence, and then compute  
190 the distance between the normalised counts. A kmer is a contiguous sub-sequence of length  
191  $k$  in a given sequence. We selected short kmers here with  $k = 4$ : there are  $4^4 = 64$  dif-  
192 ferent tetramers which is a good compromise between longer ones with more resolution,  
193 but too sparse histograms of counts, or smaller ones with coarse resolution and less empty  
194 categories. If  $k = 6$ , there are  $4^6 = 4096$  different hexamers. The length of the sequences  
195 is about 450 bp, which means that at least 9 hexamers out of 10 have 0 count. For  $k = 8$ ,  
196 this increases up to 993 out of 1000. Moreover, for short sequences with bases labelled  
197  $N$ , there may be no hexamer without a  $N$  (met once in the dataset, this sequence has  
198 been eliminated). We designed an efficient algorithm that counts the frequency of each  
199 kmer in each sequence, and a short program that computes a distance between any pair  
200 of frequency distributions as the  $\ell^1$  norm, i.e., the sum of absolute values of difference of  
201 frequencies per kmer. We give here as an illustration the computation times on a standard  
202 laptop. For Smith-Waterman scores, exact local alignment with dynamical programming,  
203 programmed in C language: 5 hours, 39 minutes and 34 seconds. For kmer distances, with  
204  $k = 4$ , programmed in C language as well: 13 minutes and 4 seconds. The time for  $k = 6$   
205 is 32 minutes and 6 seconds.

206

207 The dataset used in the rest of this paper is composed of three files (see Frigerio et al.,  
208 2021):

- 209 • a csv file of botanical names for each sequence for order, family, genus and species;
- 210 • a csv file of pairwise dissimilarities computed with the Smith-Waterman algorithm;
- 211 • a csv file of pairwise distances based on the comparison of tetramer and hexamer

212 histograms (same format as Smith-Waterman dissimilarities)

## 213 **2.2 Visualisation of the whole dataset using MDS and t-SNE**

214 A preliminary step is to propose a global picture of the dataset based on the dissimilarity  
215 matrices, without a classification objective. Multidimensional Scaling (MDS) is a method  
216 that, once a dissimilarity matrix between items is given, builds a point cloud in a Euclidean  
217 space of prescribed dimension where each point corresponds to an item (here a sequence),  
218 and such that the Euclidean distance between any two points is as close as possible to  
219 the dissimilarity given in the matrix (Cox and Cox, 2001; Izenman, 2008). In our case,  
220 we selected the so-called Classical Scaling, as proposed initially in Torgerson (1952). It is  
221 expected that the projection of the point cloud on the first axis encompasses much of the  
222 information about the structure of the point cloud. MDS was run with the dissimilarity  
223 matrices built with the Smith-Waterman algorithm and as distances between tetramer  
224 histograms. We also applied the t-SNE algorithm (van der Maaten and Hinton, 2008) to  
225 obtain a complementary 2D representation of the point cloud. The t-SNE algorithm is  
226 another technique for reduction dimension. It is based on the minimisation of a divergence  
227 between a distribution probability on points' neighbours in the original space and in the  
228 visualisation space. While MDS approximates at best the global structure of the distance  
229 array, t-SNE gives a better summary of local structures (van der Maaten and Hinton, 2008).  
230 MDS and t-SNE have been run on the whole data set (1458 sequences).

## 231 **2.3 General approach**

232 Depending on the specific question addressed, we selected a different sample of the whole  
233 dataset. However, in all cases, the general approach for comparing two classifications was  
234 the same and can be broken down into four steps.

235 In step 1, we selected the sub-sample: either the whole dataset with filters, or only

236 the individuals of a particular order, or of a particular family. We then extracted the  
237 sub-matrix corresponding to the  $n$  individuals in the sample, from the global dissimilarity  
238 matrix based on the Smith-Waterman score. We also extracted the sub-matrix of the global  
239 kmer-based dissimilarity matrix, for  $k = 4$  and  $k = 6$ . The next steps were applied for each  
240 sub-matrix.

241 In step 2, we built the classifications corresponding to AHC with the three aggregation  
242 methods, Ward, Complete Linkage (CL) and Single Linkage (SL), and to SBM (see SI for  
243 a description of these methods). The number of clusters  $K$  was provided by the botanical  
244 classification of the individuals of the sub-sample. For instance if we wanted to study  
245 agreement between the classification into families and the molecular-based ones, we cut the  
246 AHC hierarchy of classifications at  $K$  equal to the number of families in the sample, and  
247 we ran SBM for the same value. At the end of step 2, we had five different classifications  
248 of the individuals in the sub-sample.

249 In step 3, we compared the classifications, two by two, for each possible pair of classifi-  
250 cations (10 pairs in total). We used a visual tool for preliminary analysis of the agreement  
251 between two classifications: Sankey plots. A Sankey plot is a flow chart in which the  
252 width of an arrow is proportional to the flow. For instance, if there are  $n_{kk'}$  sequences that  
253 are in class  $k$  for the botanical classification and  $k'$  for a molecular-based classification,  
254 there is a flow of width proportional to  $n_{kk'}$  between those two clusters. We computed  
255 an index as well, to quantify the agreement. Classification comparison is equivalent to  
256 the comparison of two partitions of the same set, a dynamic research area with several  
257 surveys (Pfitzner et al., 2009). Several indices were proposed and we chose the Normalised  
258 Mutual Information (NMI1 in Pfitzner et al., 2009, see the Supplementary Information  
259 for a formal definition). It is not empirical and has a sound basis in information theory,  
260 as opposed to indices based on counting pairs of elements that may be non-symmetric or  
261 non-bounded or even be dependent on  $K$  or  $n$ , making comparison difficult. The Nor-  
262 malised Mutual Information is normalised and, as such, bounded by 0 and 1, facilitating

263 interpretation and comparison of indices. A Normalised Mutual Information of 0 indicates  
264 independence between the two classifications, while a Normalised Mutual Information of 1  
265 indicates a perfect agreement. For an easier interpretation, we also defined threshold on  
266 the Normalised Mutual Information values, to define domains of very good, good, average,  
267 poor and very poor agreement between two classifications. The method to compute the  
268 thresholds is based on simulated partition. It is presented in the Supplementary Material,  
269 together with the thresholds values (section 4 and Figure 1.

270 Finally, when replicates of the experiment are performed like in Section 2.4, in a fourth  
271 step, we analysed the distributions of the Normalised Mutual Information for a given pair  
272 of classifications in order to study trends in the agreement using histograms and boxplot  
273 representations.

## 274 **2.4 Comparison of botanical and molecular-based classification** 275 **at the family and genus levels, on replicates**

276 In order to have information on the variability of the results, we created 10 sub-samples  
277 of the whole dataset each of them corresponding to the individuals of a particular order,  
278 and 20 sub-samples each of them corresponding to the individuals of a particular family.  
279 We selected only orders (respectively families) composed of at least 15 individuals, and  
280 structured into more than one family (respectively genus). The number of individuals in  
281 the sub-samples at order level varies between 15 and 321. For the sub-samples at family level  
282 it varied between 17 and 127. Then, for the samples at the order level, we performed the  
283 four molecular-based clustering with  $K$  equals to the number of families in that order. For  
284 the samples at the family level, we chose  $K$  equals to the number of genera. The orders are  
285 Malpighiales, Ericales, Sapindales, Laurales, Myrtales, Magnoliales, Gentianales, Rosales,  
286 Oxalidales and Malvales.

287 We structured the empirical analysis of the Normalised Mutual Information obtained

288 (30 × 10 values) into four different analyses addressing the following questions: (i) What is  
289 the level of agreement between the botanical classification and each of the four molecular-  
290 based ones? (ii) Are the classifications provided by the four molecular-based clusterings  
291 similar? (iii) Can we identify elements of the dissimilarity matrix that explain the vari-  
292 ability observed in the answer to question (i) and that would be indicators of the agree-  
293 ment/independence between the botanical classification and the molecular-based ones? (iv)  
294 How does the agreement change between the botanical classification and the molecular-  
295 based ones when substituting kmer-based distances for Smith-Waterman dissimilarities?  
296 In practice, for question (i), we only considered the Normalised Mutual Information in-  
297 volving the botanical classification and any of the four molecular-based ones, whereas for  
298 question (ii), we only considered the Normalised Mutual Information between any pairs of  
299 the molecular-based classifications. For question (iii), we studied three factors: the taxo-  
300 nomic level of the groups, the entropy of the botanical classification (defined as the entropy  
301 of the normalised vector of the groups sizes), and the structure of the dissimilarity matrix.  
302 The latter was measured by three different ratios between the within-group dissimilarities  
303 and the between-group dissimilarities (see Supplementary Information). We only present  
304 here the one for which we observed a relationship with the Normalised Mutual Informa-  
305 tion values on the data:  $r_{mean}$ , defined as the mean of the larger within-class dissimilarity  
306 over the mean of the smaller between-class dissimilarity. Intuitively when the dissimilarity  
307 matrix is well structured into several groups each with a small within-class dissimilarity,  
308 then  $r_{mean}$  will be lower than 1. On contrary, when there are no clearly delimited groups  
309 of similar individuals then  $r_{mean}$  will be larger than 1. This is illustrated of Figure 2 in the  
310 Supplementary Information.

## 311 2.5 Comparison of botanical and molecular-based classification 312 on the whole data set

313 We looked at whether or not clustering on the whole dataset could directly retrieve botanical  
314 classification at levels higher than species (genera, families, orders). In addition, the same  
315 comparison was performed for species as well, as a benchmark. Since several taxa are  
316 singletons, regardless of the level, or have a very small number of sequences (e.g. Apiales  
317 are represented by three sequences only in the whole sample), we built one sub-sample  
318 for each taxonomic level by filtering out taxa with less sequences than a given threshold.  
319 The size of those sub-samples are given in Table 1, with the number of sequences and of  
320 different taxa per level, and the threshold selected for filtering sequences.

321 For a given taxonomic level, we ran SBM and AHC with Ward, CL and SL, on the  
322 sub-matrix of the associated sub-sample and for  $K$  equal to the number of taxa present  
323 in this sub-sample. This was done both on the matrix of dissimilarities between scores  
324 of the Smith-Waterman algorithm and on distances between tetramer frequencies. We  
325 compared each of these four classifications with the botanical one using Normalised Mutual  
326 Information. Note that a good Normalised Mutual Information at a low taxonomic level  
327 does not automatically imply a good Normalised Mutual Information at a higher level. If  
328 there are  $K_s$  species and  $K_g$  genera, the SBM classification into  $K_g$  classes is build without  
329 using the SBM classification into  $K_s$  classes. By construction the AHC classification into  
330  $K_s$  classes is embedded into the one into  $K_g$  but depending on the structure of the dis-  
331 similarity matrix, the successive merges can make the AHC move away from the botanical  
332 classification.

333 As for the study of the replicates, we also computed the entropy and the  $r_{mean}$  ratio of  
334 the botanical orders, families, genera and species classifications. For each of the taxonomic  
335 levels, we produced a visual graphical analysis by generating Sankey plots.

## 336 **3 Results**

### 337 **3.1 Visualisation of the whole dataset using MDS and t-SNE**

338 We represented the shape of the point cloud on the first two axes built with MDS on  
339 the dissimilarity matrix, with points coloured according to the order that they belong to  
340 (see Figure 1, left). For Smith-Waterman-based dissimilarities, axis 1 clearly distinguishes  
341 Ericales (in purple) and Sapindales (dark green), and axis 2, Malpighiales (in light green).  
342 Axis 3 distinguishes Fabales (blue), and the set of Laurales and Magnoliales (red and  
343 orange), which are primitive Eudicots. When using t-SNE (see Figure 1, right), clusters of  
344 sequences appears more clearly, with less overlapping than with MDS. These clusters are in  
345 general composed of sequences of the same order. However an order can be split into several  
346 clusters. This phenomenon is reduced for families (see Figure 3, right, in Supplementary  
347 Information), which indicated a stronger link between dissimilarities and families, than  
348 between dissimilarities and orders.

349 The organisation of the point cloud is different for tetramer-based dissimilarities (see  
350 Figure 4 in Supplementary Information). For MDS, the point cloud is more compact. Axis  
351 1 distinguishes the same set of Laurales and Magnoliales, and axis 2 distinguishes Fabales.  
352 With t-SNE also, the separation between groups is less obvious when using tetramer-based  
353 dissimilarities. Clearly, the shape of the point cloud based on Smith-Waterman distances is  
354 more closely related to the organisation of specimens in botanical orders. Such a connection  
355 is blurred for tetramer-based distances. This allowed us to predict that agreement between  
356 the botanical classification and the molecular-based ones will be lower when using tetramer-  
357 based distances.



## 3.2 Comparison of botanical and molecular-based classification at the family and genus levels on replicates

We present first the results obtained with Smith-Waterman scores for points (i) to (iii) raised in Section 2.4. We then show how results change when working with kmer-based distances (point (iv)).

**(i) Level of agreement between the botanical classification and the molecular-based ones.** For SBM, Ward, and CL, the shape of the histogram of the 30 Normalised Mutual Information is the same (see Figure 5 of Supplementary Material). The mode is observed at large values and 50 % of the values correspond to good to very good agreement, according to our definition of Normalised Mutual Information categories (see Figure 2). Then, intermediate values of the Normalised Mutual Information (corresponding to an average agreement according to our thresholds) are not often observed. In the case of the Normalised Mutual Information between SL and the botanical classification, the mode is also observed at values corresponding to very good agreement, however the second mode is for values of very poor agreement. So globally we observe a range of values that correspond to good to very good agreement between the botanical and the molecular-based classification, with better performance for the Ward method.

**(ii) Mutual agreement of the responses of the four molecular-based clustering methods.** There is a good agreement between the three AHC methods (see Figure 3). We observed larger Normalised Mutual Information between Ward and CL than between Ward and SL or CL and SL, but the median values are all in the categories good or very good. The SBM classification is globally in good agreement with Ward, in average agreement with CL and in poor agreement with SL, if we consider the median value of the Normalised Mutual Information.

**(iii) Factors explaining variability in the results.** We observed no clear difference in the distribution of the Normalised Mutual Information (between the botanical classifi-

384 cation and the molecular-based ones) when computed on replicates whose groups are at  
385 the family level or those at the genus level or when pooling the replicates (see Table 2).

386 We observed a trend towards an increase in agreement between botanical and molecular-  
387 based classifications when the entropy of the sub-sample increases (Figure 4 left). We also  
388 observed a clear decrease of the agreement when the ratio  $r_{mean}$  increases (see Figure 4  
389 right). When a dissimilarity matrix is associated with a ratio larger than 1, it can be the  
390 case that several sequences exist that are closer to sequences belonging to a different genus  
391 or family than to sequences in their own genus or family. This can lead to low Normalised  
392 Mutual Information.

393 *(iv) Influence of the choice of dissimilarity.* We observed a decrease of the Nor-  
394 malised Mutual Information when substituting the Smith-Waterman dissimilarity with the  
395 tetramer-based or 6mer-based distances (Table 2). For  $k = 4$ , this decrease ranged between  
396 6 % to 39 % depending on the taxonomic level of the groups and the molecular-based clus-  
397 tering method. For  $k = 6$  it is lower and ranged between 0 % and 28 %. As with the  
398 Smith-Waterman dissimilarity, the agreement with the botanical classification is the high-  
399 est for the Ward-based classification, and we still observed the influence of the entropy of  
400 the botanical classification and of  $r_{mean}$  on the agreement (Figures 6 and 7 in Supplemen-  
401 tary Material). From now on, we present only results for the Smith-Waterman dissimilarity  
402 and for tetramer-based distances, to illustrate the best and the worst case.

403

404 In conclusion, agreement between the botanical classification and molecular-based ones  
405 can be good to very good. However, there are also situations where the agreement is low.  
406 We have identified several factors that can influence the level of agreement: the choice  
407 of the clustering method, with Ward leading to the greatest agreement; the choice of the  
408 dissimilarity, with a greater agreement for Smith-Waterman dissimilarities than for kmer-  
409 based distances; the entropy of the botanical classification, with greater agreement for  
410 larger entropies;  $r_{mean}$ , with greater agreement for lower ratios.

### 3.3 Comparison of botanical and molecular-based classification on the whole data set

The results presented here extend the results on the replicates with four new experiments: we compared, on the one hand, the botanical classifications of the whole dataset partitioned into 11 orders, 20 families and 36 genera, as well as 55 species as a benchmark, (see Table 1) and, on the other hand, the molecular-based classifications obtained for the same number of classes.

(i) **Level of agreement between the botanical classification and the molecular-based ones.** On Figure 5 one curve is associated to one numerical method and gives the value of the Normalised Mutual Information for the taxonomic levels ordered from the finer to the coarser: species, genera, families and orders. All curves, regardless of the molecular-based clustering method and the dissimilarity, display a decrease from species to orders. All of the methods are excellent for identifying species (Normalised Mutual Information are in categories good or very good), and decreases depend on the method: a slight decrease for the Ward method, a sharp decrease for the SL method, and an intermediate decrease for CL or SBM. When groups are at orders or even families levels, SL seems to lead to the lower indices, regardless of the dissimilarity used. This result illustrates that it is not granted that the aggregation from fine to coarse level follows the same path in botanics and in the dendrogram of the AHC. The cut of the dendrogram at  $K_s$  groups,  $K_s$  being the number of species, can be in good agreement with the botanical classification into species, but the next merging steps of AHC may not be consistent with families and orders.

The correspondence between botanical, Ward and SBM classifications obtained with Smith-Waterman dissimilarities are graphically visualised in Figure 6 for orders and Figure 7 for families, with Sankey plots. We can note two types of behaviour: a botanical group is split into several groups in Ward or SBM classifications or, on the contrary a Ward or SBM group is composed of individuals from several botanical groups. The latter is more

437 problematic when interpreting molecular-based classifications. On Figure 8, we can observe  
438 that the low Normalised Mutual Information for SL at the order level is due to the creation  
439 of a giant cluster formed by almost all of the orders present in the dataset.

440 (iii) **Factors explaining variability in the results.** The fact that agreement be-  
441 tween the molecular-based and the botanical-based classifications decreases when the tax-  
442 onomic level of the groups searched increases is in agreement with the influences of the  
443 entropy and of the  $r_{mean}$  observed on the replicates. Indeed entropy here decreases and  
444  $r_{mean}$  increases when moving from the classification into species and genera towards families  
445 and orders (see Table 3).

446 (iv) **Influence of the choice of dissimilarity.** Regardless of the clustering method,  
447 when groups are species or genera, the Normalised Mutual Information is equivalent for  
448 Smith-Waterman-based dissimilarities and for kmer-based distances (the variation is at  
449 most of 6%). When the groups are families or orders there is a decrease in the Normalised  
450 Mutual Information when performing HAC with tetramer-based distances : Normalised  
451 Mutual Information varied between 2% and 60% with the larger decrease observed for  
452 SL. On contrary, for SBM, we observe a larger Normalised Mutual Information with the  
453 tetramer-based distance, when groups are families or orders.

454 Note that for AHC, the running times varied between 1 and 3 seconds, whatever the  
455 subset of sequences considered and the level of the groups searched. For clustering with  
456 SBM on tetramer distances, we used the Gaussian distribution and the running time was  
457 about 5 minutes for clustering the whole data set into orders and about 1 hour for clustering  
458 the whole data set into families. Running time was multiplied by two when using SBM on  
459 the Smith-Waterman dissimilarity with the Poisson distribution.

## 460 **4 Discussion**

461 In this study, several numerical methods were compared on a dataset of approximately 1,500  
462 specimens of trees in a French Guianese forest for the purpose of quantifying the agreement  
463 between, on the one hand, botanical classification and, on the other hand, molecular-based  
464 classification on an array of genetic distances, on deep taxonomic levels of the classification.  
465 We discuss here the results obtained.

### 466 **4.1 Agreement between botanical and molecular-based classifi-** 467 **cations**

468 There is one pattern common to the study based on the clusterings of the 30 replicates and  
469 the clusterings performed on the whole dataset: regardless of the combination between tax-  
470 onomic level and dissimilarity, AHC with the Ward aggregation criterion provides the best  
471 agreement. Other methods rank differently depending on these combinations. Agreement  
472 can be high (good or very good values of Normalised Mutual Information), in particular  
473 when the molecular-based clustering is based on the Smith-Waterman dissimilarity. How-  
474 ever, we also occasionally observed low agreement, and we will discuss the reasons for this.  
475 When interpreting Normalised Mutual Information values, it is important to have in mind  
476 that Normalised Mutual Information is conservative in the sense that a strong agreement  
477 is required to obtain a large Normalised Mutual Information value. The strength of the  
478 agreement could be higher with another choice of index, but we selected Normalised Mutual  
479 Information partly for this conservative behaviour.

480 A strong assumption in our study is that the number of groups  $K$  in the botanical  
481 classification is known when performing the molecular-based clustering. This is obviously  
482 not the case in real situations, like in metabarcoding of environmental samples. When  $K$   
483 is not available, the Integrated Classification Likelihood criterion (Biernacki et al., 2000)  
484 for model selection can be used to estimate a number of groups that lead to a trade-off

485 between a good explanation of the dissimilarity matrix and parsimony. This criterion  
486 has the advantage to take into account the objective of clustering when comparing two  
487 models (i.e. two values for  $K$ ). A version for selecting  $K$  in a SBM has been proposed  
488 in Daudin et al. (2008). For AHC, choosing  $K$  amounts to choosing where to cut the  
489 dendrogram, and heuristics have been proposed (Husson et al., 2010; Zumel and Mount,  
490 2014) However, these approaches do not include a goal of agreement with the botanical  
491 classification. In White et al. (2010), to compare molecular-based clustering at the OTU  
492 level and the taxonomic classification, the authors used partial assignment of the sequences  
493 and the VI-cut algorithm (Navlakha et al., 2010) to automatically determine the number  
494 of OTUs that optimally matches this partial knowledge. The method relies on the Value of  
495 Information to compare two classifications, which we did not select for our study because it  
496 is not normalised. However, the VI-cut method could easily be extended to the Normalised  
497 Mutual Information and therefore provide a way to estimate the number of groups, driven  
498 by the partial taxonomic knowledge that is available on some sequences of the inventory.

499 Although neighbor-joining (Saitou and Nei, 1987) is one of the reference methods in  
500 phylogenies, and based on distances, we have not retained it in our study for two rea-  
501 sons. First, the agreement between orders and clades<sup>2</sup> (monophyly of orders) in the tree  
502 is not excellent (see section 5 and Figure 8 in Supplementary Information), and second,  
503 neighbor-joining is not a clustering method (Limpiti et al., 2014): the outcome cannot be  
504 automatically organized as a partition into clusters.

## 505 **4.2 Agreement of botanical classification and the AHC classifi-** 506 **cations**

507 In our result, a variability of agreement is observed according to the linkage method. If  
508 the dataset is organised as a set of isolated clusters, all linkage methods will find them

---

<sup>2</sup>A clade here is an internal node with its descent.

509 and provide the same classification. If not, different linkage methods will yield differ-  
510 ent classifications. Not surprisingly, we recover these behaviours in our experiments on  
511 molecular-based clustering of the tree specimens.

512 **Ward method:** The Ward method nearly always led to the best agreement with botan-  
513 ical classification, regardless of the measurement of distance (Smith-Waterman or kmers)  
514 and the taxonomic level of the groups (Sections 3.2 and 3.3).

515 **Complete linkage Method** The CL method generally led to the second-best agreement  
516 with the botanical classification. It provided classifications very similar to those obtained  
517 with the Ward method (see Table 3).

518 **Single linkage method:** In contrast, agreement between the classification provided by  
519 the SL method and the botanical classification was highly variable and could be either very  
520 good or very poor. The agreement was very poor with the classification into orders of the  
521 whole dataset (the Normalised Mutual Information is equal to 0.06 for Smith-Waterman dis-  
522 similarities and to 0.02 for tetramer-based distances, which is very close to independence),  
523 better but still low for the classification into families (the Normalised Mutual Information  
524 is equal to 0.44 for Smith-Waterman dissimilarities, and to 0.34 for tetramer-based dis-  
525 tances). As we explained, reason for that can be seen on Figure 8: the SL classification is  
526 composed of a huge cluster, containing sequences from all orders, and a set of much smaller  
527 clusters, each containing one, seldom two, orders. The creation of the huge cluster may  
528 be due to low dissimilarities existing between the orders. By nature, the SL criterion will  
529 link these orders by the well known "chaining effect" which produces long and thin clusters  
530 which are not compact (Ros and Guillaume, 2019).

### 531 4.3 Interest of SBM models for molecular-based classification

532 Even if the SBM clustering and the botanical one are in very good agreement in some of  
533 the experiments, globally, the Normalised Mutual Information values for SBM are lower  
534 than the Normalised Mutual Information for the best AHC method (see Table 2 and Figure  
535 5). When agreement with the botanical classification is good, then the SBM classification  
536 resembles the one obtained with the Ward method. This is the case when the dissimilarity  
537 matrix is well structured into communities, and all clustering methods will perform well.  
538 When agreement is low, our interpretation is the following. The main difference between  
539 AHC and SBM clustering is that AHC looks for groups with small within-group dissimi-  
540 larity (communities), while SBM does not impose such a constraint on the groups. It seeks  
541 for groups such that (i) all individual in group  $k$  share the same pattern of connections  
542 with the other groups, and (ii) members of group  $k$  are almost at the same distance to  
543 each others. However, this distance is not necessarily small, meaning that SBM groups  
544 should not be systematically interpreted as communities. When the matrix of the pairwise  
545 dissimilarities is not clearly structured according to the botanical groups, SBM clustering  
546 can create groups with individuals that are far from each other. This is what we observed  
547 on the SBM classification of the whole dataset into orders (both for the Smith-Waterman  
548 and the kmer-based clustering). For several SBM groups, the estimated parameter char-  
549 acterising the mean within group distance was larger than the lower mean distance with  
550 the other groups. In these situations, the Normalised Mutual Information between the  
551 botanical and the SBM classification is obviously low, and the ratio  $r_{mean}$  is large. A SBM  
552 classification with groups of large within-group mean distances should be a warning that  
553 the matrix of dissimilarities is not entirely structured according to the botanical classes.  
554 For similar reasons, SBM is also able to detect outlier individuals by gathering them into a  
555 group, while methods looking for communities will force them to enter a community. For  
556 these reasons, we think SBM should be considered as a valuable tool for (meta)barcoding.



#### 557 **4.4 Factors explaining the variations in the agreement.**

558 One of the two main factors influencing the quality of agreement between the botanical  
559 and the molecular-based classifications is the relative difference between the dissimilarities  
560 within and between groups in botanical classification. This notion was well captured by  
561 the  $r_{mean}$  ratio and, we obtained a clear tendency for Normalised Mutual Information  
562 to decrease when the ratio increases on the 30 replicates (Figure 4). When considering  
563 the clustering of the whole dataset, the same tendency was observed. The other factor  
564 influencing the quality of agreement is the value of the entropy of the distribution of  
565 the group sizes in the botanical classification. We observed a tendency for an increase in  
566 agreement when entropy increases, both on the 30 replicates and when clustering the whole  
567 dataset at different taxonomic levels.

568 In the latter experiments we obtained a clear decrease of agreement for high taxonomic  
569 levels, whereas in the experiments on the 30 replicates, agreement was better at the family  
570 level than at the genus level. These apparent contradictory results are actually explained by  
571 the fact that they correspond to two different protocols. On the 30 replicates the targeted  
572 set of sequences to cluster is different for each replicate: we did not search for families and  
573 genera among the same set of individuals. We instead searched for families (respectively  
574 genera) of sequences of the same order (respectively family).

575 The negative influence of the  $r_{mean}$  ratio and the positive influence of the entropy are  
576 global tendencies. We also observed variations around these main tendencies, which means  
577 that they are probably not the only factors explaining the Normalised Mutual Information  
578 values. Still, they are strong signals.

#### 579 **4.5 Biological interpretation**

580 There may be several approaches to analyse the reasons for agreement/disagreement be-  
581 tween botanical and molecular-based numerical classification. We first examine possible

582 reasons arising from the structure of the molecular data, and second we propose some in-  
583 terpretations arising from the literature in plant molecular phylogenies.

584

585 In our study on the whole dataset, the agreement between the molecular-based and the  
586 botanical-based classification is better when groups are at a low taxonomic level, hence  
587 more numerous, regardless of the method and the distance (see Figure 5). As discussed  
588 in Section 4.4, the  $r_{mean}$  ratio, involving distances within a group over distances between  
589 groups, is smaller at the family level than at the order level. This suggests that families are  
590 better delineated than orders by pairwise distances. The results shown in Figure 5 extend  
591 this observation to species over genera, and show that molecular-based delineation of taxa  
592 is more accurate at fine taxonomic levels than at coarse ones.

593 This is consistent with the evolution of plant classification system, where confidence  
594 about delineation of higher groups, like orders, is lower than for lower groups, like gen-  
595 era. APG (Angiosperm Phylogeny Group) regularly updates phylogenetic classification of  
596 plants, focusing on families and orders. Initial goal in APG has been to classify families  
597 in orders, and later revisions have focused on delineations. In the first proposal, in 1998,  
598 there were 40 orders and 462 families. In the fourth one, called APG IV (The Angiosperm  
599 Phylogeny Group et al., 2016), there were respectively 64 orders and 416 families. This  
600 is consistent with a stabilisation of family delineations, while there is still ongoing work  
601 for stabilising orders. This might be an explanation for the drop in agreement for orders,  
602 whereas the quality of agreement for families is similar to the one for genera and species  
603 for some methods (see figure 5).

604 The commonly accepted notion for molecular-based classification is monophyly in molec-  
605 ular phylogenies (Hillis et al., 1996). The evolutionary distance between two species is the  
606 age of their Most Recent Common Ancestor. It is related to genetic distance as measured  
607 here by Smith-Waterman score, provided that the molecular clock hypothesis is accepted  
608 (see Yang (2006), chapter 7, for an overview). Even if the marker selected here is neutral

609 (intergenic spacer), it is highly likely that evolution rates over tens of millions of years across  
610 all lineages have not remained constant. Main clades of angiosperms have radiated quickly  
611 in Late Cretaceous (this is Darwin's "abominable mystery", see Friedman (2009) for a his-  
612 torical perspective), whereas they have diverged earlier in late Jurassic / Early Cretaceous.  
613 Diversification occurred with heterogeneities in space and time (Ramirez-Barahona et al.,  
614 2020). It is highly likely that such heterogeneities have been reflected even partially in  
615 evolution rates of markers, which may in turn lead to heterogeneities in agreement between  
616 molecular based and botanical classifications at the level of orders. As a consequence, as-  
617 suming that botanical classification reflects monophyletic clades can lead to a decrease of  
618 agreement between botanical and molecular- based classification for higher taxa, especially  
619 at the order level.

620 Our interpretation is that uncertainties on classification of plants (e.g. APG system)  
621 are currently higher at high levels of taxonomy (orders and higher), and that this is shared  
622 by clustering of barcodes (our numerical result).

## 623 **4.6 Comparison between Smith-Waterman and kmer-based dis-** 624 **similarities**

625 Computing Smith-Waterman dissimilarity between two sequences is the most precise way  
626 to compare them. However, it is time-consuming. Computing kmer-based distances is  
627 much quicker, but at the cost of approximations. The histograms of Smith-Waterman dis-  
628 similarities and kmer-based ( $k = 4$  and  $6$ ) distances are provided in Figures 9 and 10 of  
629 Supplementary Information. A coarse correlation can be observed between both quantifi-  
630 cations of dissimilarities (see Figure 11 in the Supplementary Information), stronger for  
631 small dissimilarities. However, a significant number of pairs of sequences exists with a  
632 very low Smith-Waterman dissimilarity and a significant tetramer-based distance. This  
633 is due to the high variability in length of the *trnH-psbA* marker. For instance, a small

634 Smith-Waterman dissimilarity means that the smallest sequence is nearly identical to a  
635 contiguous sub-sequence of the larger one. However, due to the dissimilarity in length of  
636 the two sequences, the kmer histograms cannot be similar, and the kmer-based distance is  
637 large. Therefore, some small values of the Smith-Waterman dissimilarity can be associated  
638 with median values of the kmer-based distance. Since the AHC classification (regardless  
639 of the linkage) builds groups of individuals with small within group distances, it can be  
640 expected that the Smith-Waterman-based and the kmer-based classifications will be dif-  
641 ferent. In practice, as expected, we observed that agreement decreases when substituting  
642 kmers for Smith-Waterman regardless of the combination between the taxonomic level and  
643 the clustering method (but for SBM sometimes). However, substituting kmer-based dis-  
644 tances for Smith-Waterman dissimilarities did not strongly modify the agreement between  
645 the molecular-based classifications and the botanical one.

## 646 4.7 Perspectives for metabarcoding

647 The dataset is sufficiently small for all calculations to be run on a laptop in a reasonable  
648 time, making it possible to focus on the comparison of the methods. Some methods are  
649 clearly more accurate than others to retrieve orders or families in our dataset. The expec-  
650 tation is that those methods are those that will permit inventories or clustering at higher  
651 taxonomic levels such as families, orders or phyla in metabarcoding studies. However, we  
652 underline two issues.

653 We have worked with *trnH-psbA* which is highly resolute but longer than markers  
654 currently used in metabarcoding of environmental samples or with degraded DNA. It is  
655 an issue to study whether the results found here can be extended either to other groups  
656 than plants in barcoding or with shorter markers for metabarcoding. A second issue is the  
657 scaling of the clustering methods used in the study, to data sets with hundreds of thousands  
658 of reads.

659 We recommend first using AHC with the Ward method for clustering regardless of the

660 taxonomic level, and not using AHC with Single Linkage which may produce poor results,  
661 despite the observation that current softwares scaling up with NGS massive data sets make  
662 it possible to use it (like MOTHUR) or yield results very close to it (like SWARM). It  
663 can be observed that SWARM has a step for preventing the formation of giant clusters  
664 by irrelevant aggregation between two clusters from different seeds, see Mahé et al. (2014,  
665 2015).

666 Second we recommend using SBM classification to detect, via the estimated distribution  
667 of within cluster distances, situations where the molecular-based classifications may be  
668 poorly related to the morphological-based one (because the dissimilarity matrix is not  
669 clearly structured into communities).

670 These results and observations lead us to recommend the pursuit of methodological  
671 efforts to analyse metabarcoding data for building inventories at the coarse level (i.e.,  
672 between phyla and orders). Inventories at the coarse level are a first step towards the  
673 global exploration of diversity of unknown groups. This can be done in two ways. First,  
674 nearly all surveys about clustering emphasize that there is no method that is perfect and  
675 better than some others for all evaluation criteria (see, e.g., Fahad et al., 2014). Therefore,  
676 it may be useful to produce classifications by several numerical methods and to extract  
677 the shared elements. These are the ones in which the user can be more confident that  
678 they actually reflect an actual structure in the data. Second, scaling-up methods that  
679 have proven themselves to properly perform on well-known datasets, like AHC with Ward  
680 linkage or SBM-based clustering, is a key issue. A very efficient method for clustering may  
681 be to "divide and conquer": first, dividing the problem by building connected components  
682 in a graph built from pairwise distances and, second, conquering by implementing AHC  
683 Ward or SBM in each connected component. More globally, connecting these efforts with  
684 studies on wider classes of methods under development for clustering for big-data (Fahad  
685 et al., 2014) is a challenging issue for metabarcoding.

## 686 **5 Acknowledgements**

687 The authors acknowledge support of an “Investissement d’Avenir” grant managed by the  
688 Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01). M.A. Abouabdallah  
689 PhD is funded by INRAE and by INRIA. The authors acknowledge the work done by H.  
690 Caron (sequencing), D. Sabatier and J.F. Molino (botanical assignation of trees) for a  
691 previous work that we have used here. We thank Rémy Petit for key discussions during  
692 the preparation of this manuscript, Jean-Marc Frigerio for many useful discussions for  
693 data management and analysis, as well as Philippe Chaumeil for his contribution to the  
694 programs to compute dissimilarities. We thank as well three anonymous reviewers for their  
695 comments which led to significant improvements in the content of the manuscript.

## 696 **6 Author contributions**

697 A.F. and N.P. designed the study. M.A.A., A.F. and N.P. performed the research and  
698 analyzed the data. The paper was drafted by M.A.A., A.F. and N.P. and written by A.F.  
699 and N.P. All authors commented on and approved the final manuscript.

## 700 **7 Data accessibility**

701 Sequences used to compute distances are in NCBI under accession number KX247940–KX249593.  
702 A fasta file with these sequences, a file with taxonomical assignation for each tree, as well  
703 as pairwise Smith-Waterman and kmer distances are publicly available at  
704 <https://doi.org/10.15454/XSJ079> in Inrae Data Portal.

## 705 **References**

706 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local  
707 alignment search tool. *Journal of Molecular Biology*, 215:403–410.

- 708 Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017). Stochastic block-models for  
709 multiplex networks : An application to a multilevel network of researchers. *Journal of*  
710 *the Royal Statistical Society: Series A Statistics in Society*, 180(1):295–314.
- 711 Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering  
712 with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and*  
713 *Machine Intelligence*, 22(7):719–725.
- 714 Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K.  
715 (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends*  
716 *in Ecology and Evolution*, 27:233–243.
- 717 Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E.  
718 (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical*  
719 *Transactions of the Royal Society B*, 360:1935–1943.
- 720 Caron, H., Molino, J.-F., Sabatier, D., Léger, P., Chaumeil, P., Scotti-Saintagne, C.,  
721 Frigério, J.-M., Scotti, I., Franc, A., and Petit, R. J. (2019). Chloroplast DNA vari-  
722 ation in a hyperdiverse tropical tree community. *Ecology and Evolution*, 9(8):4897–4905.
- 723 Cole, A. J., editor (1969). *Numerical Taxonomy*. Academic Press, London & New York.
- 724 Cox, T. and Cox, M. A. A. (2001). *Multidimensional Scaling - Second edition*, volume 88  
725 of *Monographs on Statistics and Applied Probability*. Chapman & al.
- 726 Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs.  
727 *Statistics and Computing*, 18:173–183.
- 728 Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioin-*  
729 *formatics*, 26:2460–2461.
- 730 Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Fofou, S., and  
731 Bouras, B. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and  
732 Empirical Analysis. *IEEE Transactions on Emerging Topics on Computing*, 2(4):267–  
733 279.
- 734 Faskowitz, J., Yan, X., Zuo, X.-N., and Sporns, O. (2018). Weighted stochastic block  
735 models of the human connectome across the life span. *Scientific Reports*, 8.
- 736 Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer.
- 737 Floyd, R., Abebe, E., Papert, A., and Blaxter, M. (2002). Molecular barcodes for soil  
738 nematode identification. *Molecular Ecology*, 11:839–850.
- 739 Fontaneto, D., Boschetti, C., and Ricci, C. (2008). Cryptic diversification in ancient asex-  
740 uals: evidence from the bdelloid rotifer *Philodina flaviceps*. *Journal of Evolutionary*  
741 *Biology*, 21:580–587.

- 742 Friedman, W. E. (2009). The meaning of Darwin's "abominable mystery". *American*  
743 *journal of Botany*, 96(1):5–21.
- 744 Frigerio, J.-M., Caron, H., Sabatier, D., Molino, J.-F., and Franc, A. (2021). Guiana Trees.  
745 Portail Data INRAE, V1, DOI: 10.15454/XSJ079.
- 746 Gusfield, D. (1997). *Algorithms on strings, trees, and sequences*. Cambridge University  
747 Press, Cambridge, UK.
- 748 Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environ-  
749 mental barcoding: a next generation sequencing approach for biomonitoring applications  
750 using river benthos. *PLoS One*, 6(4):e17497.
- 751 Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a  
752 method of unsupervised Bayesian clustering. *Bioinformatics*, 27(5):611–618.
- 753 Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifica-  
754 tions through DNA barcodes. *Proceedings of the Royal Society of London B*, 270:313–321.
- 755 Hillis, D. M., Moritz, C., and Mable, B. (1996). *Molecular Systematics*. Sinauer, Sunder-  
756 land, Mass.
- 757 Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps.  
758 *Social Networks*, 5(2):109–137.
- 759 Hollingsworth, P. M., Forrest, L., Spouge, J. L., and al., H. . (2009). A dna barcode for  
760 land plants. *PNAS*, 106:12794–12797.
- 761 Hollingsworth, P. M., Graham, S. W., and Little, D. P. (2011). Choosing and Using a Plant  
762 DNA Barcode. *PLoS ONE*, 6(5):e19254.
- 763 Husson, F., Pagès, J., and Lê, S. (2010). *Exploratory Multivariate Analysis by Example*  
764 *Using R*. Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press Taylor  
765 & Francis.
- 766 Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer, NY.
- 767 Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R.,  
768 Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. X., Benedick, S.,  
769 Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C.,  
770 and Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via  
771 metabarcoding. *Ecology Letters*, 16(10):1245–1257.
- 772 Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.-F., and Bouchez, A. (2013).  
773 Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities:  
774 a test for freshwater diatoms. *Molecular Ecology Resources*, 13:607–619.
- 775 Lee, C. and Wilkinson, D. (2019). A review of stochastic block models and extensions for  
776 graph clustering. *Applied Network Science*, 4(122).



- 777 Leray, M. and Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized  
778 samples reveal patterns of marine benthic diversity. *PNAS*, 112(7):2076–2081.
- 779 Levenstein, V. I. (1966). Binary codes capable of correcting insertions and reversals. *Sov.*  
780 *Phys. Dokl.*, 10:707–710.
- 781 Limpiti, T., Amornbunchornvej, C., Intarapanich, A., Assawamakin, A., and Tongsimma, S.  
782 (2014). injclust: Iterative neighbor-joining tree clustering framework for inferring popu-  
783 lation structure. *Computational Biology and Bioinformatics, IEEE/ACM Transactions*  
784 *on*, 11:903–914.
- 785 López-García, P., Rodríguez-Valera, F., Pedros-Alio, C., and Moreira, D. (2001). Unex-  
786 pected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, 409:603–607.
- 787 Madelaine, C., Péliissier, R., Vincent, G., Molino, J. F., Sabatier, D., Prévost, M. F., and  
788 de Namur, C. (2007). Mortality and recruitment in a lowland tropical rain forest of french  
789 guiana: Effects of soil type and species guild. *Journal of Tropical Ecology*, 23:277–287.
- 790 Mahé, F., Czech, L., Stamatakis, A., Quince, C., de Vargas, C., Dunthorn, M., and Rognes,  
791 T. (2019). Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics*.
- 792 Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust  
793 and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593.
- 794 Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2:  
795 highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420.
- 796 Meiklejohn, K. A., Damaso, N., and Robertson, J. (2019). Assessment of BOLD and  
797 Genbank – Their accuracy and reliability for the identification of biological materials.  
798 *PLoS ONE*, 14(6).
- 799 Miele, V. and Matias, C. (2017). Revealing the hidden structure of dynamic ecological  
800 networks. *Royal Society Open Science*, 4(6).
- 801 Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., and Nielsen, R. (2008).  
802 Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics. *Systematic*  
803 *Biology*, 57(5):750–757.
- 804 Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The*  
805 *Computer Journal*, 26(4):354–359.
- 806 Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for  
807 R and Python. *Journal of Statistical Software*, 53(9):1–18.
- 808 Navlakha, S., White, J., Nagarajan, N., Pop, M., and Kingsford, C. (2010). Finding  
809 biologically accurate clusterings in hierarchical tree decompositions using the variation  
810 of information. *Journal of Computational Biology*, 17(3):503–516.

- 811 Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to search for  
812 similarities in the amino-acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.
- 813 Pang, X., Liu, C., Shi, L., Liu, R., Liang, D., Li, H., Cherny, S. S., and Chen, S. (2012).  
814 Utility of the *trnH-psbA* Intergenic Spacer Region and its combinations as Plant DNA  
815 Barcodes: a Meta-Analysis. *PLoS ONE*, 7(11):e48833.
- 816 Petit, R. J. and Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology  
817 and Evolution*, 24(7):386–393.
- 818 Pfitzner, D., Leibbrandt, R., and Powers, D. (2009). Characterization and evaluation of  
819 similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(361).
- 820 Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S.,  
821 Kamoun, S., Sumlin, W. D., and Vogler, A. (2006). Sequence-Based Species Delimitation  
822 for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*, 55(4):595–609.
- 823 Puillandre, N., Modica, M. V., Zhang, Y., Sirovich, L., Boisselier, M.-C., Cruaud, C., Hol-  
824 ford, M., and Samadi, S. (2012). Large-scale species delimitation method for hyperdiverse  
825 groups. *Molecular Ecology*, 21:2671–2691.
- 826 Ramirez-Barahona, S., Sauquet, H., and Magallon, S. (2020). The delayed and geographi-  
827 cally heterogeneous diversification of flowering plant families. *Nature Ecology & Evolu-  
828 tion*, 4:1232–1238.
- 829 Ros, F. and Guillaume, S. (2019). A hierarchical clustering algorithm and an improvement  
830 of the single linkage criterion to deal with noise. *Expert Systems with Applications*,  
831 128:96–108.
- 832 Sabatier, D., Grimaldi, M., Prévost, M., Guillaume, J., Godron, M., Dosso, M., and Curmi,  
833 P. (1997). The influence of soil cover organization on the floristic and structural hetero-  
834 geneity of a guianan rain forest. *Plant Ecology*, 131:81–108.
- 835 Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for recon-  
836 structing phylogenetic trees. *Molecular Biology and Evolution*, (4):406–425.
- 837 Smith, P. D. and Waterman, M. S. (1981). Identification of common molecular subse-  
838 quences. *Journal of Molecular Biology*, 147:195–197.
- 839 Sneath, R. H. A. and Sokal, R. R. (1973). *Numerical taxonomy*. Freeman, San Francisco.
- 840 Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R.,  
841 Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the  
842 underexplored “rare biosphere”. *PNAS*, 103(32):12115–12120.
- 843 Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., and Farmerie, W. (2009).  
844 ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences.  
845 *Nucleic Acids Research*, 37(10):e76.

- 846 Taberlet, P., Coissac, E., Pompanon, F., Brochman, C., and Willerslev, E. (2012). Towards  
847 next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*,  
848 21:2045–2050.
- 849 Talavera, G., Dinca, V., and Vila, R. (2013). Factors affecting species delimitations with  
850 the GMYC model: insights from a butterfly survey. *Methods in Ecology and Evolution*,  
851 4:1101–1110.
- 852 The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F.,  
853 Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S.,  
854 and Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification  
855 for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean*  
856 *Society*, 181(1):1–20.
- 857 Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika*,  
858 17(4):401–419.
- 859 van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne.  
860 *Journal of Machine Learning Research*, 9:2579–2605.
- 861 White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M.  
862 (2010). Alignment and clustering of phylogenetic markers - implications for microbial  
863 diversity studies. *BMC Bioinformatics*, 11(152):1471–2105.
- 864 Yang, Z. (2006). *Computational Molecular Evolution*. Oxford Series in Ecology and Evo-  
865 lution. Oxford University Press.
- 866 Zhang, J., Kapli, P., Pavlidis, P., and Satamakis, A. (2013). A general species delimitation  
867 method with applications to phylogenetic placements. *Bioinformatics*, 29(22):2869–2876.
- 868 Zumel, N. and Mount, J. (2014). *Practical data science with R*. Manning.

Taxonomic level	Number of sequences	Number of taxa	Minimal size of a taxon
Species	313	55	5
Genera	845	36	10
Families	1349	30	10
Orders	1357	11	15

Table 1: Characteristics of the four subsamples of sequences, one per taxonomic level. The number of sequences in the sample is lower for low taxonomic levels because we selected only taxa composed of a minimal number of sequences, and there are less sequences of a given species than of a given genera, etc. Each subsample is used for a comparison between the molecular-based clustering methods and the botanical classification.

Method		Families			Genera			Pooled		
		SW	4mer	6mer	SW	4mer	6mer	SW	4mer	6mer
AHC	Ward	1	0.61	0.72	0.83	0.73	0.74	0.87	0.71	0.74
	SL	0.51	0.48	0.65	0.75	0.59	0.72	0.70	0.58	0.68
	CL	0.85	0.63	0.66	0.75	0.71	0.75	0.75	0.67	0.68
SBM		0.67	0.52	0.51	0.82	0.62	0.71	0.73	0.61	0.68

Table 2: Normalised Mutual Information between the botanical classification (into families or into genera) and the four molecular-based classifications (row) for two dissimilarities (column). SW stands for Smith-Waterman, 4mer for kmer-based distances computed with kmers of length  $k = 4$  and 6mer for kmer-based distances computed with kmers of length  $k = 6$

). Results for families are median values over 10 samples and results for genera are median values over 20 samples. A sample is the set of sequences of an order (10 of them) or a family (20 of them).

	Orders	Families	Genus	Species
Entropy	2.15	3.01	3.39	3.98
$r_{mean}$ with SW	2.22	1.3	0.60	0.30
$r_{mean}$ with kmer	1.89	1.29	0.77	0.14

Table 3: Entropy and  $r_{mean}$  ratio (describing the ratio between mean larger within-group and mean smaller between-group dissimilarities) for the botanical classifications of the dataset into orders, families, genera and species. SW stands for Smith-Waterman and kmer for kmer-based distances computed with kmers of length  $k = 4$ . Samples (one per taxonomic level) are those which have been built with the filters presented in Table 1.

870 **9** Figures

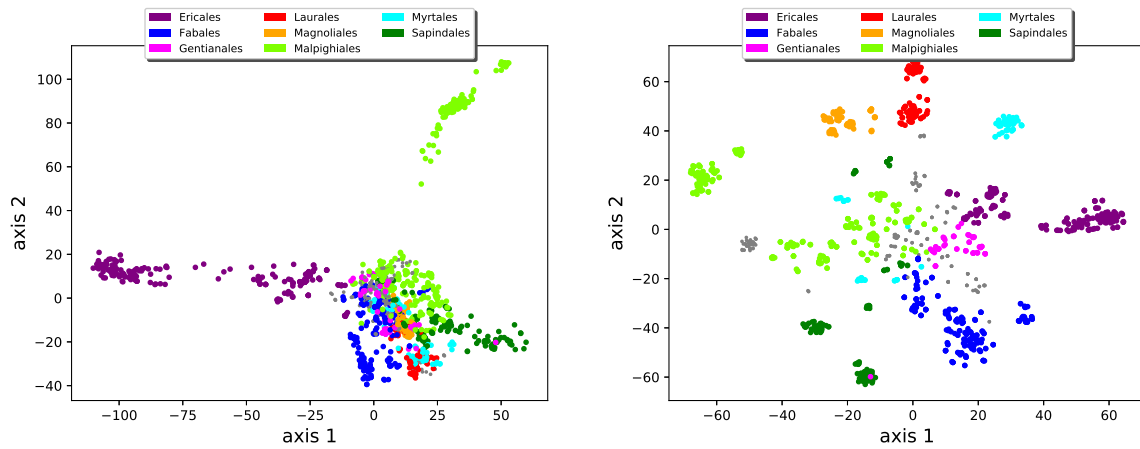


Figure 1: Visualisation of the sequences of the whole data set, as a point cloud. One dot is one sequence. The points of the eight more numerous orders are coloured, while the others are in grey. Dissimilarities are computed with the Smith-Waterman algorithm. Left: MDS, projected on axis 1 and 2. Right, t-SNE.

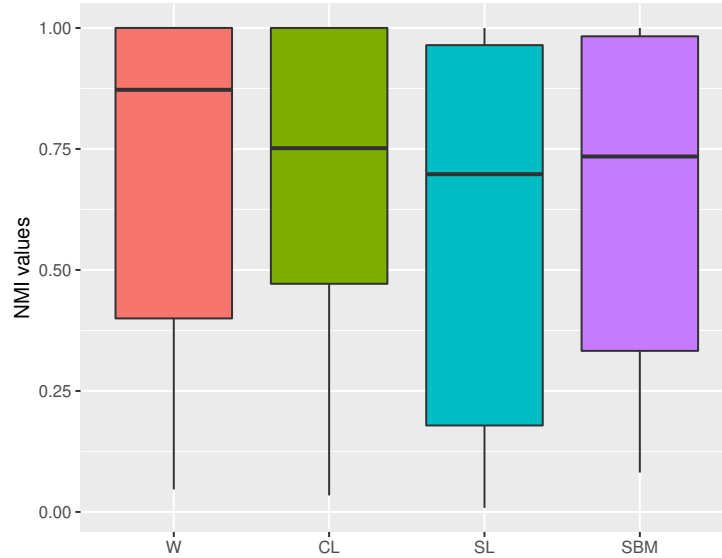


Figure 2: Boxplots on the distribution of the Normalised Mutual Information computed between each molecular-based classification and the botanical one. The data are the Normalised Mutual Information obtained on 30 replicates (10 classifications into families and 20 into genera). A replicate is the set of sequences of an order (10 of them) or a family (20 of them). Results obtained using the Smith-Waterman dissimilarity.

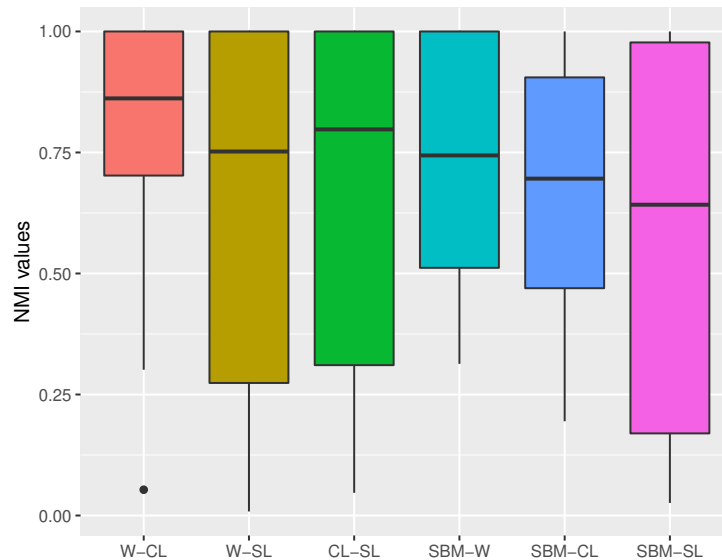


Figure 3: Boxplots on the distribution of the Normalised Mutual Information computed between each pair of molecular-based classifications. The data are the Normalised Mutual Information obtained on 30 replicates (10 classifications into families and 20 into genera). A replicate is the set of sequences of an order (10 of them) or a family (20 of them). Results obtained using the Smith-Waterman dissimilarity.

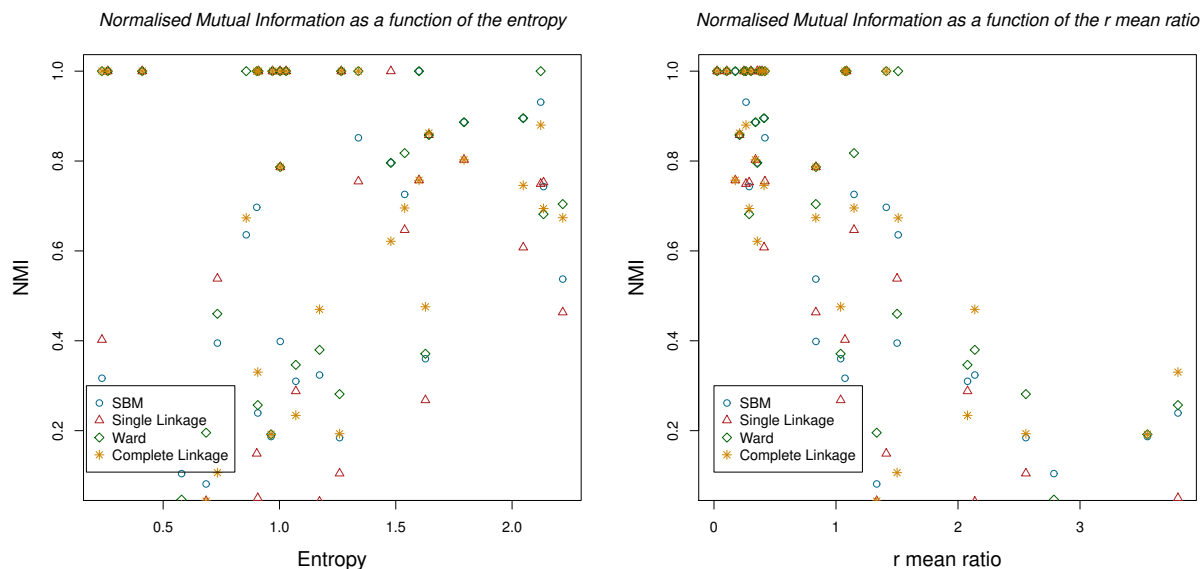


Figure 4: Normalised Mutual Information as a function of the entropy (left) and the ratio  $r_{mean}$  (right) computed on the botanical classification. Each point corresponds to one of the four molecular-based clustering method applied to one of the 30 replicates. The  $x$ -axis is the value of the entropy or ratio  $r_{mean}$  computed on the botanical classification, the  $y$ -axis is the Normalised Mutual Information between the botanical classification and the molecular-based one. The Clustering is made using the Smith-Waterman dissimilarity.

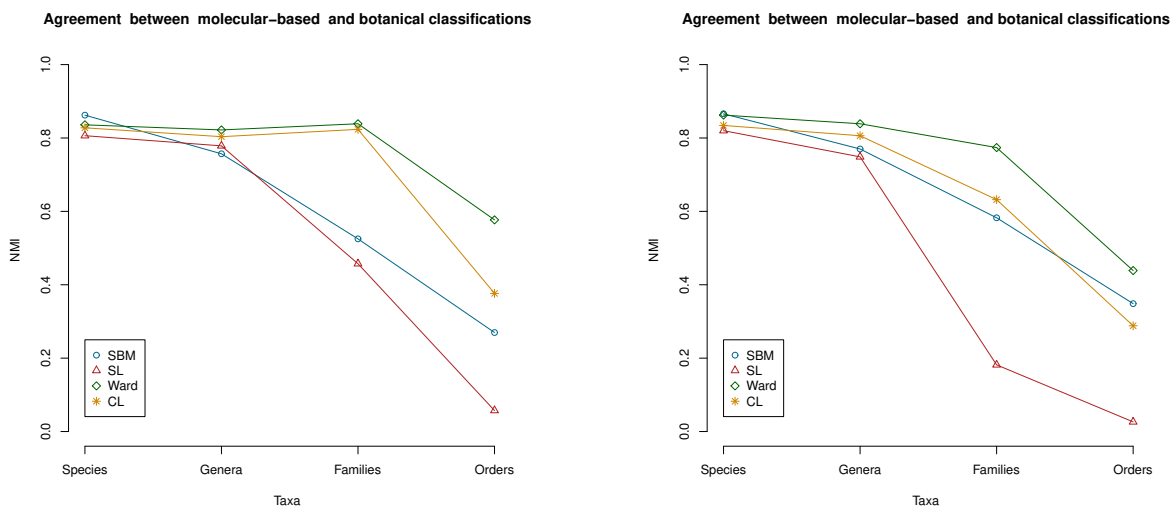


Figure 5: Agreement between molecular-based classifications and botanical classification from low to higher taxonomic levels.  $x$  axis: taxonomic levels,  $y$  axis: Normalised Mutual Information between molecular-based classification and botanical classification. One curve corresponds to one molecular-based classification. The Normalised Mutual Information are computed for classifications obtained on the samples (one per taxonomic level) presented in Table 1. Left: Smith-Waterman dissimilarities. Right: tetramer-based distances.



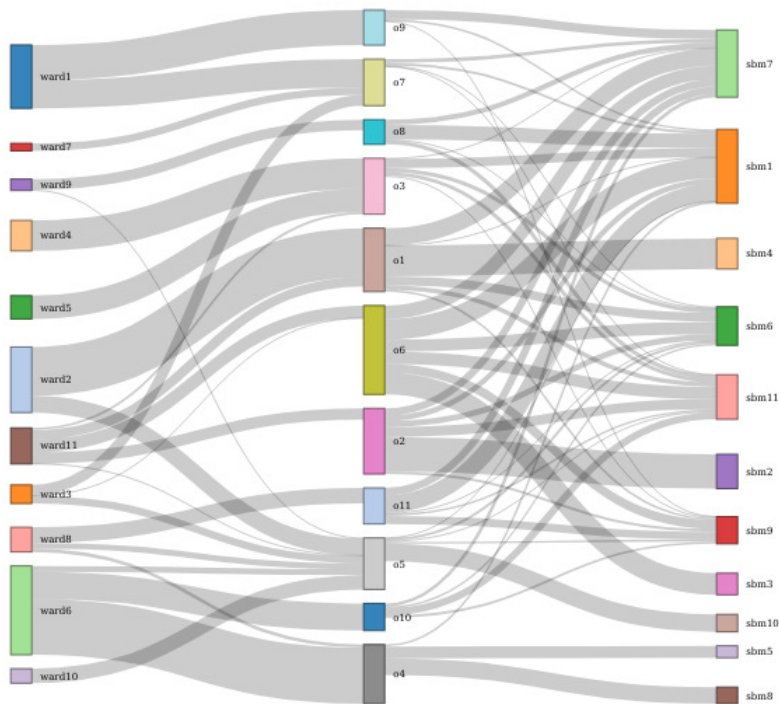


Figure 6: Sankey plot of correspondences between AHC with Ward (left column), botanical (central column) and SBM classification (right column) at the order level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.

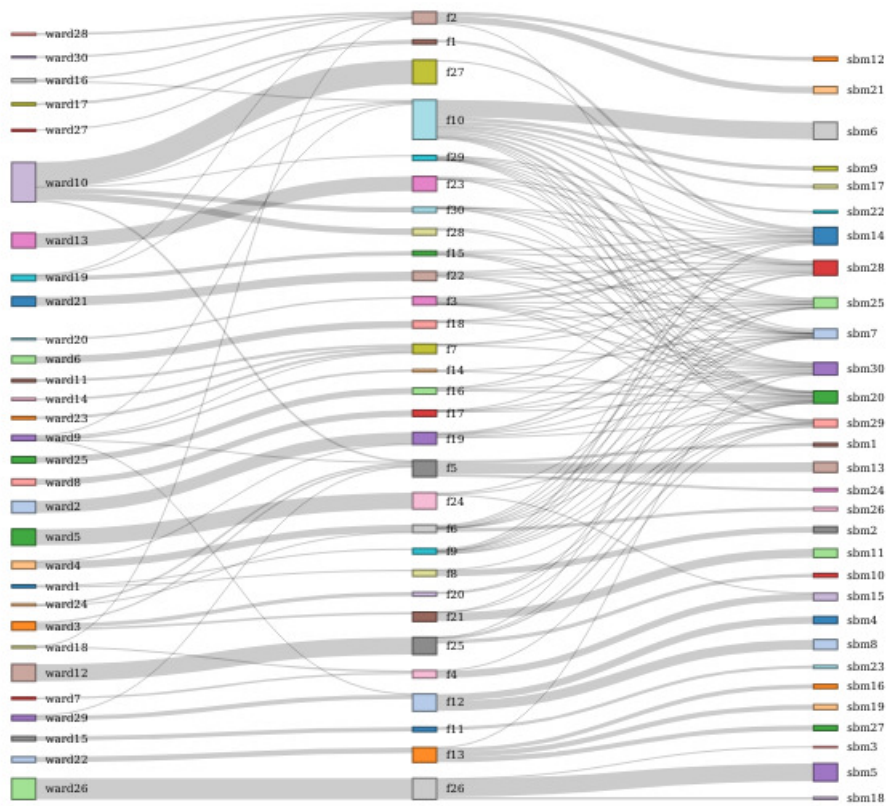


Figure 7: Sankey plot of correspondences between AHC with Ward (left column), botanical (central column) and SBM classification (right column) at the family level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.

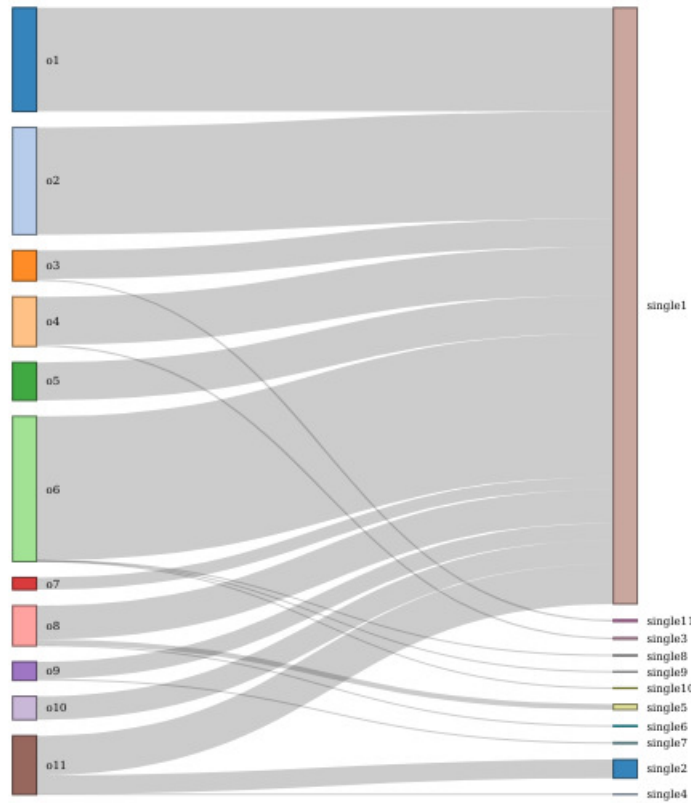


Figure 8: Sankey plot of correspondences between botanical classification (left column) and AHC with Single Linkage (right column), at the order level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.