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

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[1,2], Nathalie Peyrard[3,*], and Alain Franc [1,2]

[1]Université de Bordeaux, INRAE, BIOGECO, 33612 Cestas, France

[2]Pleiade, EPC INRIA-INRAE-CNRS, Université de Bordeaux, 33405 Talence, France

[3]Université de Toulouse, INRAE, UR MIAT, 31320 Castanet-Tolosan, France

[*]corresponding author: nathalie.peyrard@inrae.fr

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

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

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

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

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

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

[1]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.

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

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

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

More specifically, we first worked with 30 non random sub-samples of the whole dataset, each sub-sample comprising either all the individuals of an order or of a family. In each case, we compared the botanical classification of the individuals at the next finer taxonomic level with the molecular-based classifications. In a second step, we studied whether the mean behaviour observed from these replicates is recovered when the set of individuals to be classified is larger and more diverse, by comparing the botanical classification of the whole dataset into orders with the molecular-based classification for the same number of classes. We also performed the comparison at the family, gender and species levels.

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

If the morphological-based taxonomic classification is a priori unique, this is not true for a molecular-based classification. A diversity of softwares for implementing hierarchical clustering has been proposed for more than a decade in metabarcoding with the objective of efficient scaling with respect to the growing size of environmental datasets. This includes Uclust (Edgar, 2010), ESPRIT (Sun et al., 2009) and SWARM (Mahé et al., 2014, 2015, 2019), which make it possible to cluster millions of barcodes on a laptop. Nearly all of the hierarchical clustering algorithms mentioned above rely at one step or another on heuristics (like computing kmer-based distances, considering short distances only i.e. working with sparse distance matrices) to make computation feasible within a reasonable time with reasonable memory. SWARM uses kmers only as a first step to filter out pairs of sequences which are distant and cannot belong to a same compact community. In this study, we focus on understanding the agreement (or not) between molecular-based classification from clustering and botanical classification, without computational constraints. We therefore consider Aggregative Hierarchical Clustering (AHC), whose above-mentioned algorithms can be seen as heuristic versions for scaling up, with three different aggregation methods: Single Linkage, Complete Linkage, Ward (Murtagh, 1983; Müllner, 2013). Statistical models like Bayesian classifiers with mixture models have also been considered in the literature to cluster sequences (Hao et al., 2011). To extend the scope of statistical modeling in molecular-based taxonomy, we explore here the potential interest of a model-based clustering method, the Stochastic Block Model (SBM, Holland et al., 1983; Daudin et al., 2008; Lee and Wilkinson, 2019) as an alternative to AHC. SBMs are already widely applied with success in domains like the social sciences (Barbillon et al., 2017), the analysis of ecological interaction networks (Miele and Matias, 2017) and neurology (Faskowitz et al., 2018). They rely on a more flexible definition of a cluster than AHC (searching for general groups and not just communities), and we hypothesised that SBM and AHC could be complementary in their capacity to distinguish meaningful groups of individuals in an inventory.

In the following section, we provide a brief description of the dataset. We also describe the method. Results on the quality of the agreement between molecular-based and morphological-based classifications obtained on replicates are presented in Section 3.2, the 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 *trn*H-*psb*A, which is highly resolutive, 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. *trn*H-*psb*A has been used is 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 1458×1458 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

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

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

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

histograms (same format as Smith-Waterman dissimilarities)

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

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

2.3 General approach

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

In step 1, we selected the sub-sample: either the whole dataset with filters, or only the individuals of a particular order, or of a particular family. We then extracted the sub-matrix corresponding to the n individuals in the sample, from the global dissimilarity matrix based on the Smith-Waterman score. We also extracted the sub-matrix of the global kmer-based dissimilarity matrix, for $k = 4$ and $k = 6$. The next steps were applied for each sub-matrix.

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

In step 3, we compared the classifications, two by two, for each possible pair of classifications (10 pairs in total). We used a visual tool for preliminary analysis of the agreement between two classifications: Sankey plots. A Sankey plot is a flow chart in which the width of an arrow is proportional to the flow. For instance, if there are $n_{kk'}$ sequences that are in class k for the botanical classification and k' for a molecular-based classification, there is a flow of width proportional to $n_{kk'}$ between those two clusters. We computed an index as well, to quantify the agreement. Classification comparison is equivalent to the comparison of two partitions of the same set, a dynamic research area with several surveys (Pfitzner et al., 2009). Several indices were proposed and we chose the Normalised Mutual Information (NMI1 in Pfitzner et al., 2009, see the Supplementary Information for a formal definition). It is not empirical and has a sound basis in information theory, as opposed to indices based on counting pairs of elements that may be non-symmetric or non-bounded or even be dependent on K or n, making comparison difficult. The Normalised Mutual Information is normalised and, as such, bounded by 0 and 1, facilitating interpretation and comparison of indices. A Normalised Mutual Information of 0 indicates independence between the two classifications, while a Normalised Mutual Information of 1 indicates a perfect agreement. For an easier interpretation, we also defined threshold on the Normalised Mutual Information values, to define domains of very good, good, average, poor and very poor agreement between two classifications. The method to compute the thresholds is based on simulated partition. It is presented in the Supplementary Material, together with the thresholds values (section 4 and Figure 1.

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

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

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

We structured the empirical analysis of the Normalised Mutual Information obtained

(30×10 values) into four different analyses addressing the following questions: (i) What is the level of agreement between the botanical classification and each of the four molecular-based ones? (ii) Are the classifications provided by the four molecular-based clusterings similar? (iii) Can we identify elements of the dissimilarity matrix that explain the variability observed in the answer to question (i) and that would be indicators of the agreement/independence between the botanical classification and the molecular-based ones? (iv) How does the agreement change between the botanical classification and the molecular-based ones when substituting kmer-based distances for Smith-Waterman dissimilarities? In practice, for question (i), we only considered the Normalised Mutual Information involving the botanical classification and any of the four molecular-based ones, whereas for question (ii), we only considered the Normalised Mutual Information between any pairs of the molecular-based classifications. For question (iii), we studied three factors: the taxonomic level of the groups, the entropy of the botanical classification (defined as the entropy of the normalised vector of the groups sizes), and the structure of the dissimilarity matrix. The latter was measured by three different ratios between the within-group dissimilarities and the between-group dissimilarities (see Supplementary Information).We only present here the one for which we observed a relationship with the Normalised Mutual Information values on the data: r_{mean}, defined as the mean of the larger within-class dissimilarity over the mean of the smaller between-class dissimilarity. Intuitively when the dissimilarity matrix is well structured into several groups each with a small within-class dissimilarity, then r_{mean} will be lower than 1. On contrary, when there are no clearly delimited groups of similar individuals then r_{mean} will be larger than 1. This is illustrated of Figure 2 in the Supplementary Information.

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

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

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

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

3 Results

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

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

The organisation of the point cloud is different for tetramer-based dissimilarities (see Figure 4 in Supplementary Information). For MDS, the point cloud is more compact. Axis 1 distinguishes the same set of Laurales and Magnoliales, and axis 2 distinguishes Fabales. With t-SNE also, the separation between groups is less obvious when using tetramer-based dissimilarities. Clearly, the shape of the point cloud based on Smith-Waterman distances is more closely related to the organisation of specimens in botanical orders. Such a connection is blurred for tetramer-based distances. This allowed us to predict that agreement between the botanical classification and the molecular-based ones will be lower when using tetramer-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 classification and the molecular-based ones) when computed on replicates whose groups are at the family level or those at the genus level or when pooling the replicates (see Table 2).

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

(iv) **Influence of the choice of dissimilarity.** We observed a decrease of the Normalised Mutual Information when substituting the Smith-Waterman dissimilarity with the tetramer-based or 6mer-based distances (Table 2). For $k = 4$, this decrease ranged between 6 % to 39 % depending on the taxonomic level of the groups and the molecular-based clustering method. For $k = 6$ it is lower and ranged between 0 % and 28 %. As with the Smith-Waterman dissimilarity, the agreement with the botanical classification is the highest for the Ward-based classification, and we still observed the influence of the entropy of the botanical classification and of r_{mean} on the agreement (Figures 6 and 7 in Supplementary Material). From now on, we present only results for the Smith-Waterman dissimilarity and for tetramer-based distances, to illustrate the best and the worst case.

In conclusion, agreement between the botanical classification and molecular-based ones can be good to very good. However, there are also situations were the agreement is low. We have identified several factors that can influence the level of agreement: the choice of the clustering method, with Ward leading to the greatest agreement; the choice of the dissimilarity, with a greater agreement for Smith-Waterman dissimilarities than for kmer-based distances; the entropy of the botanical classification, with greater agreement for 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 dendogram 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 problematic when interpreting molecular-based classifications. On Figure 8, we can observe that the low Normalised Mutual Information for SL at the order level is due to the creation of a giant cluster formed by almost all of the orders present in the dataset.

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

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

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

4 Discussion

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

4.1 Agreement between botanical and molecular-based classifications

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

A strong assumption in our study is that the number of groups K in the botanical classification is known when performing the molecular-based clustering. This is obviously not the case in real situations, like in metabarcoding of environmental samples. When K is not available, the Integrated Classification Likelihood criterion (Biernacki et al., 2000) for model selection can be used to estimate a number of groups that lead to a trade-off between a good explanation of the dissimilarity matrix and parsimony. This criterion has the advantage to take into account the objective of clustering when comparing two models (i.e. two values for K). A version for selecting K in a SBM has been proposed in Daudin et al. (2008). For AHC, choosing K amounts to choosing where to cut the dendogram, and heuristics have been proposed (Husson et al., 2010; Zumel and Mount, 2014) However, these approaches do not include a goal of agreement with the botanical classification. In White et al. (2010), to compare molecular-based clustering at the OTU level and the taxonomic classification, the authors used partial assignment of the sequences and the VI-cut algorithm (Navlakha et al., 2010) to automatically determine the number of OTUs that optimally matches this partial knowledge. The method relies on the Value of Information to compare two classifications, which we did not select for our study because it is not normalised. However, the VI-cut method could easily be extended to the Normalised Mutual Information and therefore provide a way to estimate the number of groups, driven by the partial taxonomic knowledge that is available on some sequences of the inventory.

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

4.2 Agreement of botanical classification and the AHC classifications

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

[2]A clade here is an internal node with its descent.

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

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

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

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

4.3 Interest of SBM models for molecular-based classification

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

4.4 Factors explaining the variations in the agreement.

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

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

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

4.5 Biological interpretation

There may be several approaches to analyse the reasons for agreement/disagreement between botanical and molecular-based numerical classification. We first examine possible reasons arising from the structure of the molecular data, and second we propose some interpretations arising from the literature in plant molecular phylogenies.

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

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

The commonly accepted notion for molecular-based classification is monophyly in molecular phylogenies (Hillis et al., 1996). The evolutionary distance between two species is the age of their Most Recent Common Ancestor. It is related to genetic distance as measured here by Smith-Waterman score, provided that the molecular clock hypothesis is accepted (see Yang (2006), chapter 7, for an overview). Even if the marker selected here is neutral (intergenic spacer), it is highly likely that evolution rates over tens of millions of years across all lineages have not remained constant. Main clades of angiosperms have radiated quickly in Late Cretaceous (this is Darwin's "abominable mystery", see Friedman (2009) for a historical perspective), whereas they have diverged earlier in late Jurassic / Early Cretaceous. Diversification occurred with heterogeneities in space and time (Ramirez-Barahona et al., 2020). It is highly likely that such heterogeneities have been reflected even partially in evolution rates of markers, which may in turn lead to heterogeneities in agreement between molecular based and botanical classifications at the level of orders. As a consequence, assuming that botanical classification reflects monophyletic clades can lead to a decrease of agreement between botanical and molecular- based classification for higher taxa, especially at the order level.

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

4.6 Comparison between Smith-Waterman and kmer-based dissimilarities

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

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

4.7 Perspectives for metabarcoding

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

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

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

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

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

5 Acknowledgements

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

6 Author contributions

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

7 Data accessibility

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

References

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

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017). Stochastic block-models for multiplex networks : An application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A Statistics in Society*, 180(1):295–314.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution,*, 27:233–243.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B*, 360:1935–1943.

Caron, H., Molino, J.-F., Sabatier, D., Léger, P., Chaumeil, P., Scotti-Saintagne, C., Frigério, J.-M., Scotti, I., Franc, A., and Petit, R. J. (2019). Chloroplast DNA variation in a hyperdiverse tropical tree community. *Ecology and Evolution*, 9(8):4897–4905.

Cole, A. J., editor (1969). *Numerical Taxonomy.* Academic Press, London & New York.

Cox, T. and Cox, M. A. A. (2001). *Multidimensional Scaling - Second edition*, volume 88 of *Monographs on Statistics and Applied Probability.* Chapman & al.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18:173–183.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26:2460–2461.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., and Bouras, B. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics on Computing*, 2(4):267–279.

Faskowitz, J., Yan, X., Zuo, X.-N., and Sporns, O. (2018). Weighted stochastic block models of the human connectome across the life span. *Scientific Reports*, 8.

Felsenstein, J. (2004). *Inferring phylogenies.* Sinauer.

Floyd, R., Abebe, E., Papert, A., and Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology*, 11:839–850.

Fontaneto, D., Boschetti, C., and Ricci, C. (2008). Cryptic diversification in ancient asexuals: evidence from the bdelloid rotifer *Philodina flaviceps*. *Journal of Evolutionary Biology*, 21:580–587.

Friedman, W. E. (2009). The meaning of Darwin's "abominable mystery". *American journal of Botany*, 96(1):5–21.

Frigerio, J.-M., Caron, H., Sabatier, D., Molino, J.-F., and Franc, A. (2021). Guiana Trees. Portail Data INRAE, V1, DOI: 10.15454/XSJ079.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences.* Cambridge University Press, Cambridge, UK.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environmental barcoding: a next generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, 6(4):e17497.

Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, 27(5):611–618.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, 270:313–321.

Hillis, D. M., Moritz, C., and Mable, B. (1996). *Molecular Systematics.* Sinauer, Sunderland, Mass.

Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Hollingsworth, P. M., Forrest, L., Spouge, J. L., and al., H. . (2009). A dna barcode for land plants. *PNAS*, 106:12794–12797.

Hollingsworth, P. M., Graham, S. W., and Little, D. P. (2011). Choosing and Using a Plant DNA Barcode. *PLoS ONE*, 6(5):e19254.

Husson, F., Pagès, J., and Lê, S. (2010). *Exploratory Multivariate Analysis by Example Using R.* Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press Taylor & Francis.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques.* Springer, NY.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. X., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C., and Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10):1245–1257.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.-F., and Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, 13:607–619.

Lee, C. and Wilkinson, D. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(122).

Leray, M. and Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *PNAS*, 112(7):2076–2081.

Levenstein, V. I. (1966). Binary codes capable of correcting insertions and reversals. *Sov. Phys. Dokl.*, 10:707–710.

Limpiti, T., Amornbunchornvej, C., Intarapanich, A., Assawamakin, A., and Tongsima, S. (2014). injclust: Iterative neighbor-joining tree clustering framework for inferring population structure. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11:903–914.

López-García, P., Rodriguez-Valera, F., Pedros-Alio, C., and Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, 409:603–607.

Madelaine, C., Pélissier, R., Vincent, G., Molino, J. F., Sabatier, D., Prévost, M. F., and de Namur, C. (2007). Mortality and recruitment in a lowland tropical rain forest of french guiana: Effects of soil type and species guild. *Journal of Tropical Ecology*, 23:277–287.

Mahé, F., Czech, L., Stamatakis, A., Quince, C., de Vargas, C., Dunthorn, M., and Rognes, T. (2019). Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics*.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420.

Meiklejohn, K. A., Damaso, N., and Robertson, J. (2019). Assessment of BOLD and Genbank – Their accuracy and reliability for the identification of biological materials. *PLoS ONE*, 14(6).

Miele, V. and Matias, C. (2017). Revealing the hidden structure of dynamic ecological networks. *Royal Society Open Science*, 4(6).

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., and Nielsen, R. (2008). Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics. *Systematic Biology*, 57(5):750–757.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.

Navlakha, S., White, J., Nagarajan, N., Pop, M., and Kingsford, C. (2010). Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. *Journal of Computational Biology*, 17(3):503–516.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to search for similarities in the amino-acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Pang, X., Liu, C., Shi, L., Liu, R., Liang, D., Li, H., Cherny, S. S., and Chen, S. (2012). Utility of the *trn*H-*psb*A Intergenic Spacer Region and its combinations as Plant DNA Barcodes: a Meta-Analysis. *PLoS ONE*, 7(11):e48833.

Petit, R. J. and Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology and Evolution*, 24(7):386–393.

Pfitzner, D., Leibbrandt, R., and Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(361).

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., and Vogler, A. (2006). Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*, 55(4):595–609.

Puillandre, N., Modica, M. V., Zhang, Y., Sirovich, L., Boisselier, M.-C., Cruaud, C., Holford, M., and Samadi, S. (2012). Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology*, 21:2671–2691.

Ramirez-Barahona, S., Sauquet, H., and Magallon, S. (2020). The delayed and geographically heterogeneous diversification of flowering plant families. *Nature Ecology & Evolution*, 4:1232–1238.

Ros, F. and Guillaume, S. (2019). A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications*, 128:96–108.

Sabatier, D., Grimaldi, M., Prévost, M., Guillaume, J., Godron, M., Dosso, M., and Curmi, P. (1997). The influence of soil cover organization on the floristic and structural heterogeneity of a guianan rain forest. *Plant Ecology*, 131:81–108.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, (4):406–425.

Smith, P. D. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197.

Sneath, R. H. A. and Sokal, R. R. (1973). *Numerical taxonomy*. Freeman, San Francisco.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *PNAS*, 103(32):12115–12120.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., and Farmerie, W. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, 37(10):e76.

Taberlet, P., Coissac, E., Pompanon, F., Brochman, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21:2045–2050.

Talavera, G., Dinca, V., and Vila, R. (2013). Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. *Methods in Ecology and Evolution*, 4:1101–1110.

The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S., and Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20.

Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17(4):401–419.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M. (2010). Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics*, 11(152):1471–2105.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press.

Zhang, J., Kapli, P., Pavlidis, P., and Satamakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22):2869–2876.

Zumel, N. and Mount, J. (2014). *Pratical data science with R*. Manning.

8 Tables

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.

		Families			Genera			Pooled		
Method		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.

9 Figures

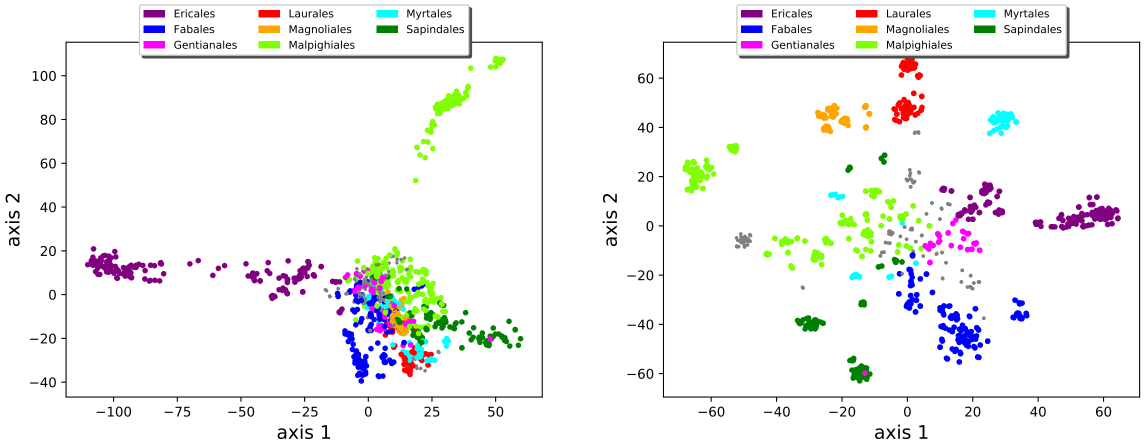

Figure 1: Visualisation of the sequences of the whole data set, as a point cloud. One dot is one sequence. The points of the eight more numerous orders are coloured, while the others are in grey. Dissimilarities are computed with the Smith-Waterman algorithm. Left: MDS, projected on axis 1 and 2. Right,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

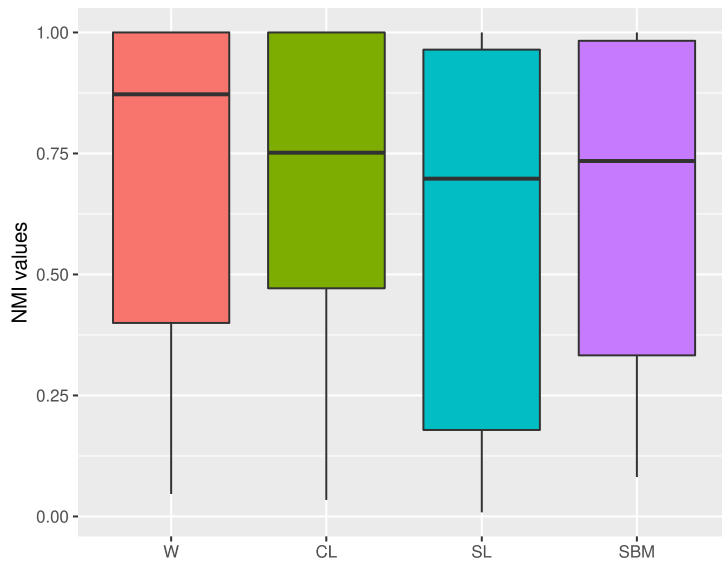

Figure 2: Boxplots on the distribution of the Normalised Mutual Information computed between each molecular-based classification and the botanical one. The data are the Normalised Mutual Information obtained on 30 replicates (10 classifications into families and 20 into genera). A replicate is the set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

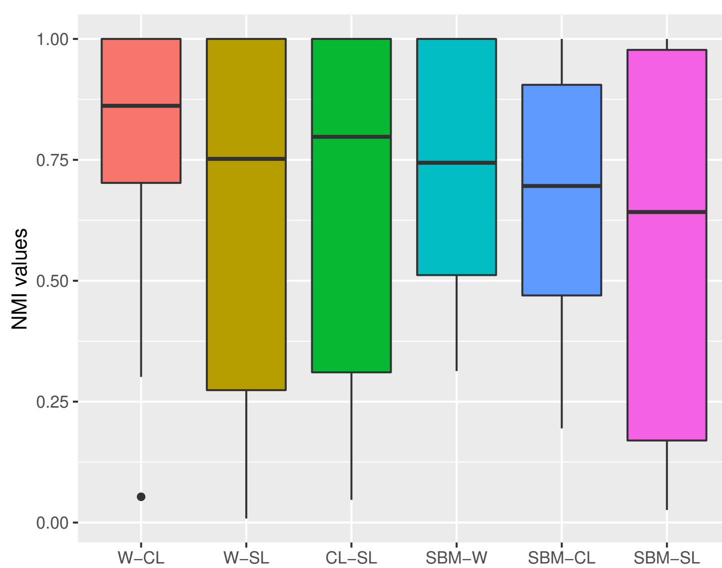

Figure 3: Boxplots on the distribution of the Normalised Mutual Information computed between each pair of molecular-based classifications. The data are the Normalised Mutual Information obtained on 30 replicates (10 classifications into families and 20 into genera). A replicate is the set of sequences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

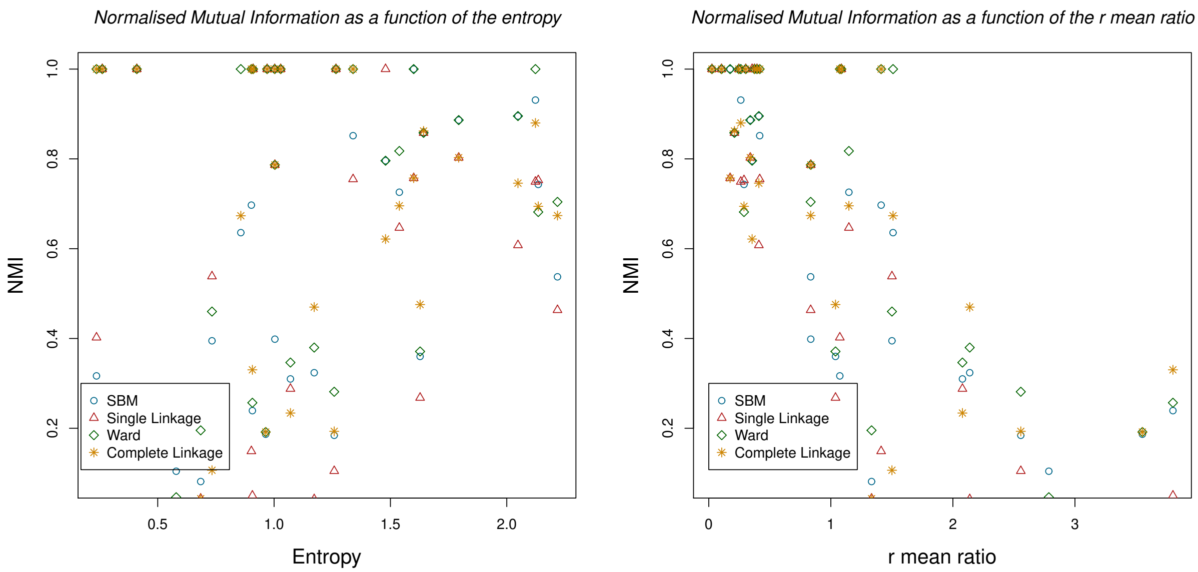

Figure 4: Normalised Mutual Information as a function of the entropy (left) and the ratio r_{mean} (right) computed on the botanical classification. Each point corresponds to one of the four molecular-based clustering method applied to one of the 30 replicates. The x-axis is the value of the entropy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

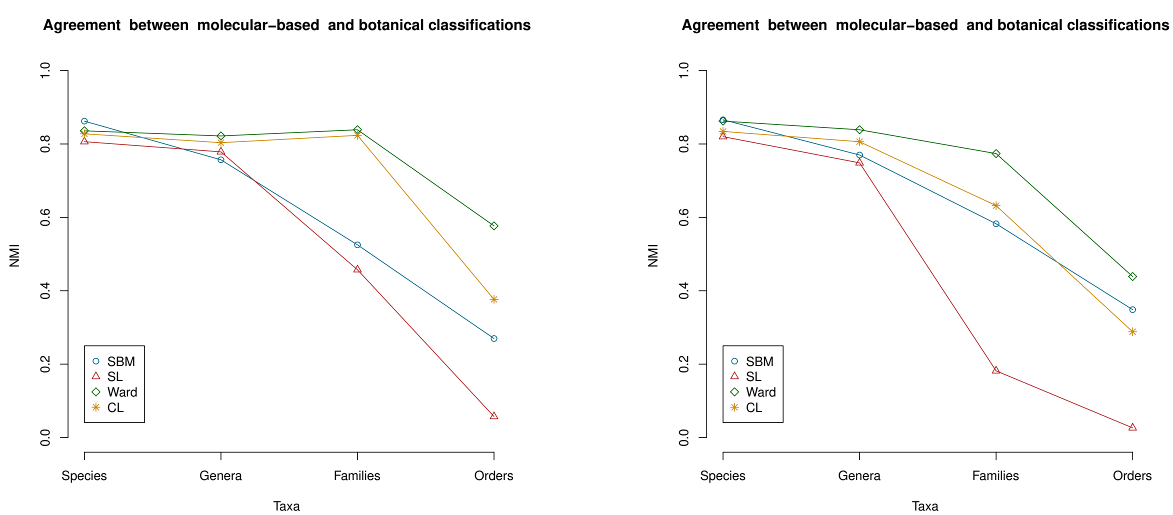

Figure 5: Agreement between molecular-based classifications and botanical classification from low to higher taxonomic levels. x axis: taxonomic levels, y axis: Normalised Mutual Information between molecular-based classification and botanical classification. One curve corresponds to one molecular-based 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.

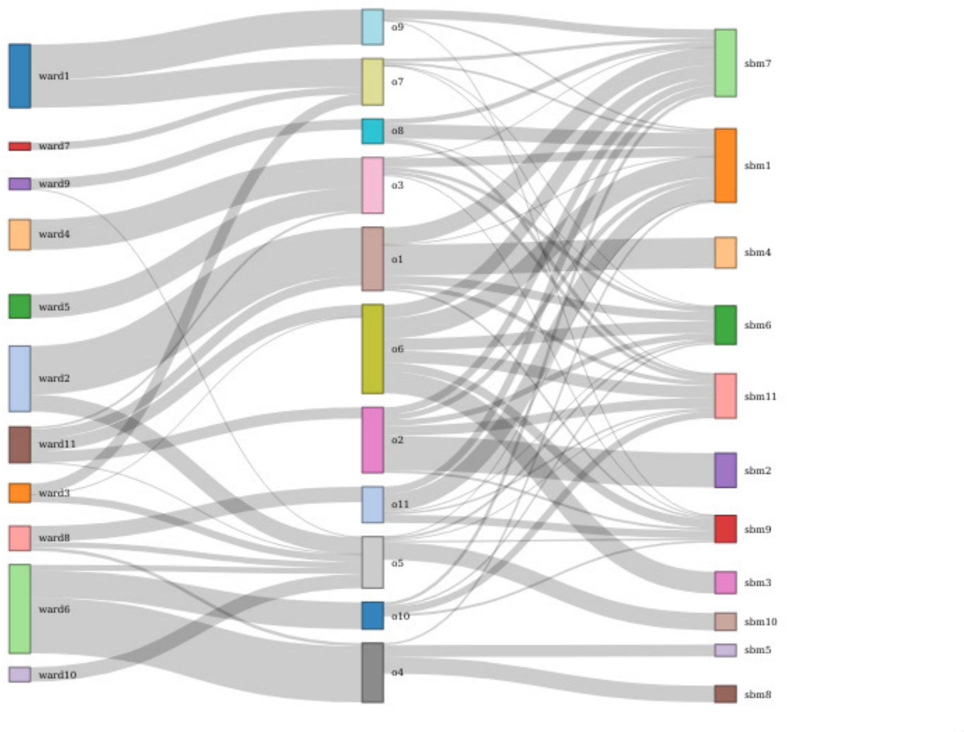

Figure 6: Sankey plot of correspondences between AHC with Ward (left column), botanical (central column) and SBM classification (right column) at the order level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.

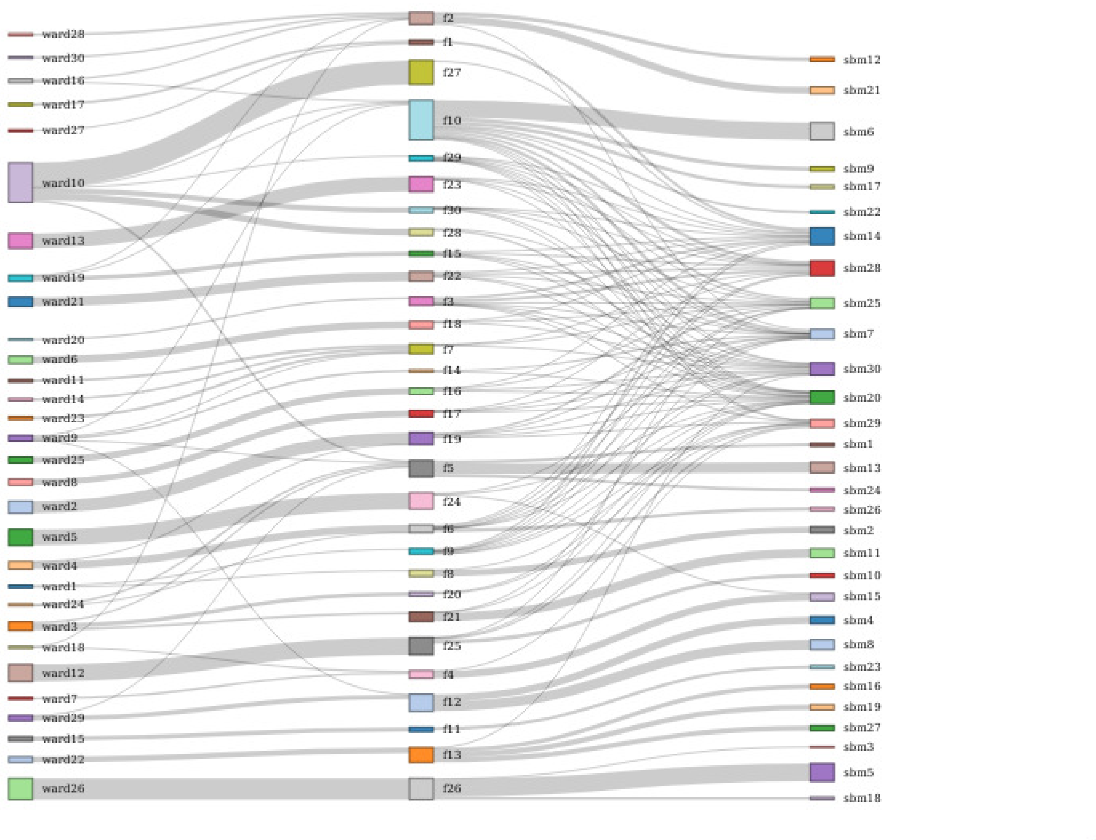

Figure 7: Sankey plot of correspondences between AHC with Ward (left column), botanical (central column) and SBM classification (right column) at the family level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.

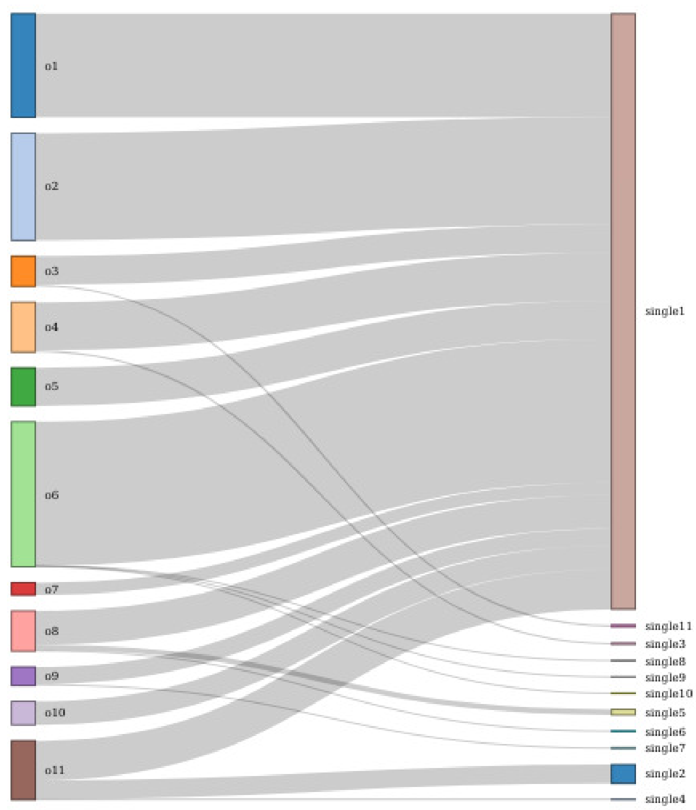

Figure 8: Sankey plot of correspondences between botanical classification (left column) and AHC with Single Linkage (right column), at the order level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes.