Integrating *TARA* Oceans datasets using unsupervised multiple kernel learning

Jérôme Mariette[1,*], Nathalie Villa-Vialaneix[1,**]

1 MIAT, Université de Toulouse, INRA, 31326 Castanet-Tolosan, France

*** jerome.mariette@inra.fr**
**** nathalie.villa-vialaneix@inra.fr**

Abstract

In metagenomic analysis, the integration of various sources of information is a difficult task since produced datasets are often of heterogeneous types. These datasets can be composed of species counts, which need to be analysed with distances, but also species abundances, interaction networks or phylogenetic information which have been shown relevant to provide a better comparison between communities. Standard integration methods can take advantage of external information but do not allow to analyse heterogenous multi-omics datasets in a generic way.

We propose a multiple kernel framework that allows to integrate multiple datasets of various types into a single exploratory analysis. Several solutions are provided to learn either a consensus meta-kernel or a meta-kernel that preserves the original topology of the datasets. This kernel is subsequently used in kernel PCA to provide a fast and accurate visualisation of similarities between samples, in a non linear space and from the multiple source point of view. A generic procedure is also proposed to improve the interpretability of the kernel PCA in regards with the original data. We applied our framework to the multiple metagenomic datasets collected during the *TARA* Oceans expedition. We demonstrate that our method is able to retrieve previous findings in a single analysis as well as to provide a new image of the sample structures when a larger number of datasets are included in the analysis.

Proposed methods are available in the R package **mixKernel**, released on CRAN. It is fully compatible with the **mixOmics** package and a tutorial describing the approach can be found on **mixOmics** web site `http://mixomics.org/mixkernel/`.

1 Introduction

The development of high-throughput sequencing technologies has substantially improved our ability to estimate complex microbial communities composition, even for organisms that cannot be cultured. The sequence reads, produced by amplicon sequencing such as 16S rRNA sequencing, can be taxonomically classified into taxa or clustered into operational taxonomic units (OTUs). Important insights have been gained from the analysis of such data by profiling microbial communities and differences between communities in a wide range of applications from the human enterotypes [3] to the plankton [5]. In microbiome studies, differences among various samples are often extracted to understand associations between organisms and external factors [10, 44]), or to characterize microbial diversity patterns [14, 17].

However, the analysis of metagenomic datasets is complex due to their sparse and compositional structure: OTU counts are often converted to relative, rather than absolute, abundances because the sequencing depth strongly varies between samples. The resulting measures are constrained to a simplex space and the standard Euclidean distance is thus irrelevant to compare samples [1]. As a consequence, directly using standard statistical methods on these data may lead to spurious results [28]. The most widely used approaches to address this issue include transforming the compositional datasets using log-ratio in order to release the simplex constrain [21] or using β-diversity measures to assess the dissimilarity between communities. These dissimilarity measures compute absolute [18] or relative [7] overlaps between two communities. In microbiome studies, they are often used as inputs for an ordination analysis, such as the Principal Coordinates Analysis (PCoA, or Multidimensional Scaling), to identify features that explain differences between studied communities.

However, [30] shows that integrating information about differences among species in the analysis (*i.e.*, by means of phylogenetic dissimilarity) is relevant to reveal phylogenetic patterns in comparing communities. Integrating the philogenetic information is usually performed by using specific dissimilarities, such as the Unifrac and weighted Unifrac measures [26, 27] in ordination methods. Alternatively, [31] propose the DPCoA to analyze the relations between the abundance data and external information corresponding to differences among species (phylogenetic, morphological, biological...). [13] extend this approach to also integrate external variables measured on communities, using a prior clustering of the communities based on these variables. [34] and [9] show that these methods can be generalized by using a kernel framework and extend them to incorporate context-dependent non-Euclidean structures with abundance data into a regression framework.

In the present work, we use a similar kernel framework to propose a generic approach that can incorporate various types of external information to metagenomic data or that can integrate multiple metagenomic datasets. More precisely, β-diversity measures or phylogenetic-based dissimilarities or any other dissimilarity measuring a specific kind or dissemblance between two samples are viewed as kernels and integrated using an unsupervised multiple kernel approach. Such a kernel can be subsequently used in combination with KPCA [39] for exploratory analysis. To improve the interpretability of our approach, indexes of the importance of the various features of the samples are proposed. The method is evaluated on the *TARA* Oceans expedition datasets [5, 19]. Results show that not only our approach allows to retrieve the main conclusions stated in the different *TARA* Oceans papers in a single and fast analysis, but that, integrating a larger number of information, it can also provide a different overview of the datasets.

2 Methods

2.1 Unsupervised multiple kernel learning

2.1.1 Kernels and notations

For a given set of observations $(x_i)_{i=1,\ldots,N}$, taking values in an arbitrary space $\mathcal{X}$, we call “kernel” a function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that provides pairwise similarities between the observations: $K_{ij} := K(x_i, x_j)$. Moreover, this function is assumed to be symmetric ($K_{ij} = K_{ji}$) and positive ($\forall\, n \in \mathbb{N},\ \forall\, (\alpha_i)_{i=1,\ldots,n} \subset \mathbb{R},\ \forall\, (x_i)_{i=1,\ldots,n} \subset \mathcal{X},\ \sum_{i,i'=1}^{n} \alpha_i \alpha_{i'} K_{ii'} \geq 0$). According to [2], this ensures that K is the dot product in a uniquely defined Hilbert space $(\mathcal{H}, \langle ., . \rangle)$ of the images of $(x_i)_i$ by a uniquely defined feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$: $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$. In the sequel, the notation K will be used to denote either the kernel itself or the evaluation matrix $(K_{ij})_{i,j=1,\ldots,N}$ depending on the context.

This setting allows us to deal with multiple source datasets in a uniform way, provided that a relevant kernel can be calculated from each dataset (examples are given in Section 3.2 for standard numeric datasets, philogenetic tree, ...). Suppose now that M datasets $(x_i^m)_{i=1,\ldots,N}$ (for $m = 1, \ldots, M$) are given instead of just one, all obtained on the same samples $i = 1, \ldots, N$. M different kernels $(K^m)_{m=1,\ldots,M}$ provide different views of the datasets, each related to a specific aspect.

Multiple kernel learning (MKL) refers to the process of linearly combining the M given kernels into a single kernel K^*:

$$K^* = \sum_{m=1}^{M} \beta_m K^m \quad \text{subject to} \quad \left\{ \begin{array}{l} \beta_m \geq 0, \ \forall\, m = 1, \ldots, M \\ \sum_{m=1}^{M} \beta_m = 1 \end{array} \right. . \tag{1}$$

By definition, the kernel K^* is also symmetric and positive and thus induces a feature space and a feature map (denoted by ϕ^* in the sequel). This kernel can thus be used in subsequent analyses (SVM, KPCA, ...) as it is supposed to provide an integrated summary of the samples.

A simple choice for the coefficients β_m is to set them all equal to $1/M$. However, this choice treats all the kernels similarly and does not take into account the fact that some of the kernels can be redundant or, on the contrary, atypical. Sounder choices aim at solving an optimization problem so as to better integrate all informations. In a supervised framework, this mainly consists in choosing weights that minimize the prediction error [15]. For clustering, a similar strategy is used in [45], optimizing the margin between the different clusters. However, for other unsupervised analyses (such as exploratory analysis, KPCA for instance), such criteria do not exist and other strategies have to be used to choose relevant weights.

As explained in [46], propositions for unsupervised multiple kernel learning (UMKL) are less numerous than the ones available for the supervised framework. Most solutions (see, *e.g.*, [25, 46]) seek at providing a kernel that minimizes the distortion between all training data and/or that minimizes the approximation of the original data in the kernel embedding. However, this requires that the datasets $(x_i^m)_i$ $(m = 1, \ldots, M)$ are standard numerical datasets: the distortion between data and the approximation of the original data are then directly computed in the input space (which is $\mathbb{R}^d$) using the standard Euclidean distance as a reference. Such a method is not applicable when the input dataset is not numerical (*i.e.*, is a phylogenetic tree for instance) or when the different datasets $(x_i^m)_i$ $(m = 1, \ldots, M)$ do not take value in a common space.

In the sequel, we propose two solutions that overcome this problem: the first one seeks at proposing a consensual kernel, which is the best consensus of all kernels. The second one uses a different point of view and, similarly to what is suggested in [46], computes a kernel that minimizes the distortion between all training data. However, this distortion is obtained directly from the M kernels, and not from an Euclidean input space. Moreover, it is used to provide a kernel representation that preserve the original data topology. Two variants are described: a sparse variant, which also selects the most relevant kernels, and a non sparse variant, when the user does not want to make a selection among the M kernels.

2.1.2 A consensus multiple kernel

Our first proposal, denoted by STATIS-UMKL, relies on ideas similar to STATIS [20, 23]. STATIS is an exploratory method designed to integrate multi-block datasets when the blocks are measured on the same samples. STATIS finds a consensus matrix, which is obtained as the matrix that has the highest average similarity with the relative positions of the observations as provided by the different blocks. We propose to use a similar idea to learn a consensus kernel.

More precisely, a measure of similarity between kernels can be obtained by computing their cosines[1] according to the Frobenius dot product: $\forall\, m,\, m' = 1,\, \ldots,\, M$,

$$C_{mm'} = \frac{\langle K^m, K^{m'} \rangle_F}{\|K^m\|_F \|K^{m'}\|_F} = \frac{\text{Trace}(K^m K^{m'})}{\sqrt{\text{Trace}((K^m)^2)\text{Trace}((K^{m'})^2)}}. \tag{2}$$

$C_{mm'}$ can be viewed as an extension of the RV-coefficient [37] to the kernel framework, where the RV-coefficient is computed between $(\phi^m(x_i^m))_i$ and $(\phi^{m'}(x_i^{m'}))_i$ (where ϕ^m is the feature map associated to K^m).

The similarity matrix $\mathbf{C} = (C_{mm'})_{m,m'=1,\ldots,M}$ provides information about the resemblance between the different kernels and can be used as such to understand how they complement each other or if some of them provide an atypical information. It also gives a way to obtain a summary of the different kernels by choosing a kernel K^* which maximizes the average similarity with all the other kernels:

$$\begin{aligned} \text{maximize}_\beta \qquad & \sum_{m=1}^{M} \left\langle K^*_{\mathbf{v}}, \frac{K^m}{\|K^m\|_F} \right\rangle_F = \mathbf{v}^\top \mathbf{C} \mathbf{v} \\ & \text{for } K^*_{\mathbf{v}} = \sum_{m=1}^{M} v_m K^m \\ & \text{and } \mathbf{v} \in \mathbb{R}^M \text{ such that } \|\mathbf{v}\|_2 = 1. \end{aligned} \tag{3}$$

The solution of the optimization problem of Equation (3) is given by the eigen-decomposition of $\mathbf{C}$. More precisely, if $\mathbf{v} = (v_m)_{m=1,\ldots,M}$ is the first eigenvector (with norm 1) of this decomposition, then its entries are all positive (because the matrices K^m are positive) and are the solution of the maximization of $\mathbf{v}^\top \mathbf{C} \mathbf{v}$. Setting $\beta = \frac{\mathbf{v}}{\sum_{m=1}^M v_m}$ thus provides a solution satisfying the constrains of Equation (1) and corresponding to a consensual summary of the M kernels.

Note that this method is equivalent to performing multiple CCA between the multiple feature spaces, as suggested in [43] in a supervised framework, or in [35] for multiple kernel PCA. However, only the first axis of the CCA is kept and a L^2-norm constrain is used to allow the solution to be obtained by a simple eigen-decomposition. This solution is better adapted to the case where the number of kernels is small.

2.1.3 A sparse kernel preserving the original topology of the data

Because it focuses on consensual information, the previous proposal tends to give more weights to kernels that are redundant in the ensemble of kernels and to discard the information given by kernels that provide complementary informations. However, it can also be desirable to obtain a solution which weights the different images of the dataset provided by the different kernels more evenly. A second solution is thus proposed, which seeks at preserving the original topology of the data. This method is denoted by sparse-UMKL in the sequel.

More precisely, weights are optimized such that the local geometry of the data in the feature space is the most similar to that of the original data. Since the input datasets are not Euclidean and do not take values in a common input space, the local geometry of the original data cannot be measured directly as in [46]. It is thus approximated using only the information given by the M kernels. To do so, a graph, the k-nearest neighbor graph (for a given $k \in \mathbb{N}^*$), $\mathcal{G}^m$, associated with each kernel K^m is built. Then,

[1] Cosines are usually preferred over the Frobenius dot product itself because they allow to re-scale the different matrices at a comparable scale. It is equivalent to using the kernel $\widetilde{K}^m = \frac{K^m}{\|K^m\|_F}$ instead of K^m.

a $(N \times N)$-matrix $\mathbf{W}$, representing the original topology of the dataset is defined such that W_{ij} is the number of times the pair (i, j) is in the edge list of $\mathcal{G}^m$ over $m = 1, \ldots, m$ (*i.e.*, the number of times, over $m = 1, \ldots, M$, that x_i^m is one of the k nearest neighbors of x_j^m or x_j^m is one of the k nearest neighbors of x_i^m).

The solution is thus obtained for weights that ensure that $\phi^*(x_i)$ and $\phi^*(x_j)$ are "similar" (in the feature space) when W_{ij} is large. To do so, similarly as [25], we propose to focus on some particular features of $\phi^*(x_i)$ which are relevant to our problem and correspond to their similarity (in the feature space) with all the other $\phi^*(x_j)$. More precisely for a given $\beta \in \mathbb{R}^M$, we introduce the N-dimensional vector $\Delta_i(\beta) = \left\langle \phi^*_\beta(x_i), \begin{pmatrix} \phi^*_\beta(x_1) \\ \vdots \\ \phi^*_\beta(x_N) \end{pmatrix} \right\rangle = \begin{pmatrix} K^*_\beta(x_i, x_1) \\ \vdots \\ K^*_\beta(x_i, x_N) \end{pmatrix}$. But, contrary to [25], we do not rely on a distance in the original space to measure topology preservation but we directly use the information provided by the different kernels through $\mathbf{W}$. The following optimization problem is thus solved:

$$\begin{aligned} \text{minimize}_\beta \quad & \sum_{i,j=1}^N W_{ij} \left\| \Delta_i(\beta) - \Delta_j(\beta) \right\|^2 \\ & \text{for } K^*_\beta = \sum_{m=1}^M \beta_m K^m \\ & \text{and } \beta \in \mathbb{R}^M \text{ such that } \beta_m \geq 0 \text{ and } \sum_{m=1}^M \beta_m = 1. \end{aligned} \tag{4}$$

The optimization problem of Equation (4) expands as

$$\begin{aligned} \text{minimize}_\beta \quad & \sum_{m,m'=1}^M \beta_m \beta_{m'} S_{mm'} \\ & \text{for } \beta \in \mathbb{R}^M \text{ such that } \beta_m \geq 0 \text{ and } \sum_{m=1}^M \beta_m = 1, \end{aligned} \tag{5}$$

for $S_{mm'} = \sum_{i,j=1}^N W_{ij} \langle \Delta_i^m - \Delta_j^m, \Delta_i^{m'} - \Delta_j^{m'} \rangle$ and $\Delta_i^m = \begin{pmatrix} K^m(x_i, x_1) \\ \vdots \\ K^m(x_i, x_N) \end{pmatrix}$. The matrix $\mathbf{S} = (S_{mm'})_{m,m'=1,\ldots,M}$ is positive and the problem is thus a standard Quadratic Programming (QP) problem with linear constrains, which can be solved by using the R package **quadprog**. Since the constrain $\sum_{m=1}^M \beta_m = 1$ is an L^1 constrain in a QP problem, the produced solution will be sparse: a kernel selection is performed because only some of the obtained $(\beta_m)_m$ are non zero. While desirable when the number of kernels is large, this property can be a drawback when the number of kernels is small and that using all kernels in the integrated exploratory analysis is expected. To address this issue, a modification of Equation (5) is proposed in the next section.

2.1.4 A full kernel preserving the original topology of the data

To get rid of the sparse property of the solution of Equation (5), an L^2 constrain can be used to replace the L^1 constrain, similarly to Equation (3):

$$\begin{aligned} \text{minimize}_{\mathbf{v}} \qquad & \sum_{m,m'=1}^{M} v_m v_{m'} S_{mm'} \\ & \mathbf{v} \in \mathbb{R}^M \text{ such that } v_m \geq 0 \text{ and } \|\mathbf{v}\|_2 = 1, \end{aligned} \tag{6}$$

and to finally set $\beta = \frac{\mathbf{v}}{\sum_m v_m}$. This problem is a Quadratically Constrained Quadratic Program (QCQP), which is known to be hard to solve. For a similar problem, [25] propose to relax the problem into a semidefinite programming optimization problem. However, a simpler solution is provided by using ADMM (Alterning Direction Method of Multipliers; [6]). More precisely, the optimization problem of Equation (6) is re-written as

$$\begin{aligned} \text{minimize}_{\mathbf{x} \text{ and } \mathbf{z}} \qquad & \mathbf{x}^T \mathbf{S} \mathbf{x} + \mathbb{I}_{\{\mathbf{x} \geq 0\}}(\mathbf{x}) + \mathbb{I}_{\{\mathbf{z} \geq 1\}} \\ & \text{such that } \mathbf{x} - \mathbf{z} = 0 \end{aligned}$$

and is solved with the method of multipliers. Final weights are then obtained by re-scaling the solution $\beta := \frac{\mathbf{z}}{\sum_m z_m}$. The method is denoted by full-UMKL in the sequel.

2.2 Kernel PCA (KPCA) and enhanced interpretability

2.2.1 Short description of KPCA

KPCA, introduced in [39], is a PCA analysis performed in the feature space induced by the kernel K^*. It is equivalent to standard MDS (*i.e.*, metric MDS or PCoA; [41]) for Euclidean dissimilarities. Without loss of generality, the kernel K^* is supposed centered[2]. KPCA simply consists in an eigen-decomposition of K^*: if $(\boldsymbol{\alpha}_k)_{k=1,\ldots,N} \in \mathbb{R}^N$ and $(\lambda_k)_{k=1,\ldots,N}$ respectively denote the eigenvectors and corresponding eigenvalues (ranked in decreasing order) then the PC axes are, for $k = 1, \ldots, N$, $a_k = \sum_{i=1}^N \alpha_{ki} \phi^*(x_i)$, where $\boldsymbol{\alpha}_k = (\alpha_{ki})_{i=1,\ldots,N}$. $\boldsymbol{a}_k = (a_{ki})_{i=1,\ldots,N}$ are orthonormal in the feature space induced by the kernel: $\forall\, k, k'$, $\langle a_k, a_{k'} \rangle = \boldsymbol{\alpha}_k^\top K^* \boldsymbol{\alpha}_{k'} = \delta_{kk'}$ with $\delta_{kk'} = \begin{cases} 0 & \text{if } k \neq k' \\ 1 & \text{otherwise} \end{cases}$. Finally, the coordinates of the projections of the images of the original data, $(\phi^*(x_i))_i$, onto the PC axes are given by: $\langle a_k, \phi^*(x_i) \rangle = \sum_{j=1}^N \alpha_{kj} K^*_{ji} = K^*_{i.} \boldsymbol{\alpha}_k = \lambda_k \alpha_{ki}$, where $K^*_{i.}$ is the i-th row of the kernel K^*.

These coordinates are useful to represent the samples in a small dimensional space and to better understand their relations. However, contrary to standard PCA, KPCA does not come with a variable representation, since the samples are described by their relations (via the kernel) and not by standard numeric descriptors. PC axes are defined by their similarity to all samples and are thus hard to interpret.

2.2.2 Interpretation

There is few attempts, in the literature, to help understand the relations of KPCA with the original measures. When the input datasets take values in $\mathbb{R}^d$, [36] propose to add a representation of the variables to the plot, visualizing their influence over the results

[2] if K^* is not centered, it can be made so by computing $K^* - \frac{1}{N} K^* \mathbf{I}_N + \frac{1}{N^2} \mathbf{I}_N^\top K^* \mathbf{I}_N$, with $\mathbf{I}_N$ a vector with N entries equal to 1.

from derivative computations. However, this approach would make little sense for datasets like ours, *i.e.*, described by discrete counts.

We propose a generic approach that assesses the influence of variables and is based on random permutations. More precisely, for a given measure j, that is used to compute the kernel K^m, the values observed on this measure are randomly permuted between all samples and the kernel is re-computed: $\widetilde{K}^{m,j}$. For species abundance datasets, the permutation can be performed at different phylogeny levels, depending on the user interest. Then, using the weights found with the original (non permuted) kernels, a new meta-kernel is obtained $\widetilde{K}^* = \sum_{l \neq m} \beta_l K^l + \beta_m \widetilde{K}^{m,j}$. The influence of the measure j on a given PC subspace is then assessed by computing the Crone-Crosby distance [11] at the axis level: $\forall\, k = 1, \ldots, N$, $D_{cc}(\alpha_k, \tilde{\alpha}_k) = \frac{1}{\sqrt{2}} \|\alpha_k - \tilde{\alpha}_k\|$, where α_k and $\tilde{\alpha}_k$ respectively denote the eigenvectors of the eigen-decomposition of K^* and $\tilde{K}^*$.[3]

Finally, the KPCA interpretation is done similarly as for a standard PCA: the interpretation of the axes $(a_k)_{k=1,\ldots,N}$ is done with respect to the observations $(x_i)_{i=1,\ldots,N}$ which contribute the most to their definition, when important variables are the ones leading to the largest Crone-Crosby distances.

Methods presented in the paper are available in the R package **mixKernel**, released on CRAN. Further details about implemented functions are provided in Supplementary Section S1.

3 Implementation on *TARA* Oceans datasets

3.1 Overview on *TARA* Oceans

The *TARA* Oceans expedition [5, 19] facilitated the study of plankton communities by providing oceans metagenomic data combined with environmental measures to the scientific community. During the expedition, 579 samples were collected for morphological, genetic and environmental analyses, from 75 stations in epipelagic and mesopelagic waters across eight oceanic provinces. The *TARA* Oceans consortium partners analyzed prokaryotic [40], viral [8] and eukaryotic-enriched [12] size fractions and provided an open access to the raw datasets and processed materials.

Some integrated analyses have already been performed with these datasets: by integrating prokaryotic, eukaryotic and viral datasets, [24] created the global plankton interactome, *i.e.*, a taxon-taxon co-occurrence network. This integrated network, associated to a sparse partial least square analysis, allowed [16] to detect associations between genomic datasets and carbon export. A similar co-occurrence strategy is used in [42] to perform an integrated analysis across domains of life to study the environmental characteristics of the Agulhas rings.

So far, all articles related to *TARA* Oceans that aim at integrating prokaryotic, eukaryotic and viral communities, took advantage of the datasets only by using co-occurrence associations. The integration analysis of the whole material aims at providing a more complete overview of the relations between all collected informations.

3.2 Dissimilarities and kernels for *TARA* Oceans datasets

Selected samples are precisely described in Supplementary Section S2. Using these samples, 8 (dis)similarities were computed:

- The **phychem** kernel is a similarity measure obtained from environmental variables. To compute this kernel, 22 numerical features were used, including, *e.g.*,

[3] Note that a similar distance can be computed at the entire projection space level but, since axes are naturally ordered in PCA, we chose to restrict to axis-specific importance measures.

temperature, salinity, ... This dataset was extracted from Table W8, available on the companion website of [40][4]. Missing values were previously imputed using a k-nearest neighbor approach, as implemented in the R package **DMwR** (for $k = 5$). Finally, the linear kernel, $K(x_i, x_j) = x_i^T x_j$, was computed between pairs of ocean samples from this dataset;

- The **pro.phylo** dissimilarity describes the phylogenetic dissimilarities between ocean samples. The companion website of [40][2] gives access to the abundance table of 35,650 OTUs summarized at different taxonomic levels as well as to the OTUs of 16S ribosomal RNA gene sequences. A phylogenomic tree was built from these data using fasttree [33]. The weighted Unifrac distance was then computed using the R package **phyloseq** [29]: $d_{wUF}(x_i, x_j) = \frac{\sum_e l_e |p_e - q_e|}{\sum_e (p_e + q_e)}$, in which, for each branch e, l_e is the branch length and p_e (respectively q_e) is the fraction of the community of ocean sample x_j (respectively of ocean sample x_j) below branch e;

- The **pro.NOGs** dissimilarity provides a measure of prokaryotic functional processes dissimilarities between ocean samples. It was obtained using the Bray-Curtis dissimilarity

$$d_{BC}(x_i, x_j) = \frac{\sum_s |n_{is} - n_{js}|}{\sum_s (n_{is} + n_{js})}, \tag{7}$$

computed on the gene abundances of 39,246 bacterial genes. In Equation (7), n_{is} is the number of counts of bacterial gene number s in ocean sample x_i. Genes were annotated using the ocean microbial reference gene catalog[2] and summarized at eggNOG gene families (genes annotated by eggNOG version 3 database: [32]). The gene abundance table is freely available from the companion website of [40][2];

- The ocean eukaryotic aspect is assessed by four dissimilarities, one for each eukaryotic organism size collected: **euk.pina** for piconanoplankton, **euk.nano** for nanoplankton, **euk.micro** for microplankton and **euk.meso** mesoplankton. The Bray-Curtis dissimilarity, defined in Equation (7), is computed on the abundance table of $\sim 150,000$ eukaryotic plankton OTUs. The dataset can be downloaded from the companion website of [12] [5];

- The **vir.VCs** dissimilarity measures ocean viral communities and was computed using the Bray-Curtis dissimilarity, defined in Equation (7), on the abundance table of 867 Viral Clusters (VCs) available from the supplementary materials of [38].

All dissimilarities, d, described above (**pro.phylo**, **pro.NOGs**, **euk.pina**, **euk.nano**, **euk.micro**, **euk.meso** and **vir.VCs**) were transformed into similarities as suggested in [22]: $K_{ij} = -\frac{1}{2}\left(d(x_i, x_j) - \frac{1}{N}\sum_{k=1}^{N}\left(d(x_i, x_k) + d(x_k, x_j)\right) + \frac{1}{N^2}\sum_{k,\,k'=1}^{N} d(x_k, x_{k'})\right)$, where d is the weighted Unifrac distance or the Bray-Curtis dissimilarity. The eight similarities obtained are all positive and are thereby kernels, which are all centred by definition. To avoid scaling effects in kernel integration, all kernels were scaled using the standard cosine transformation [4]: $\tilde{K}_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$.

4 Results and discussion

This section is divided into two parts: Section 4.1 performs the exploratory analysis only with the datasets studied in [40]. The results described in this paper are used as a

[4] http://ocean-microbiome.embl.de/companion.html
[5] http://taraoceans.sb-roscoff.fr/EukDiv/

ground truth to validate the relevance of our strategy. A further step is taken in Section 4.2 in which a larger set of datasets are analyzed to illustrate the use of the method and its efficiency to perform an integrated exploratory analysis. In both sections, analyses are performed with the full-UMKL approach presented in Section 2.1. An analysis of the correlation between kernels is provided in Supplementary Section S3 and a comparison with the other multiple kernel strategies that explains the choice of full-UMKL is discussed in Supplementary Section S4.

4.1 Proof of concept with a restricted number of datasets

In the present section, only the datasets analyzed in [40] are analyzed. These kernels are the environmental kernel, **phychem**, and the two prokaryotic kernels, **pro.phylo** and **pro.NOGs**, all computed on the 139 prokaryotic samples described in Section 3. Figure 1 (left) provides the sample projection of the first two axes of the KPCA (full-UMKL kernel). The 10 most important variables for each dataset are displayed in Figure 1 (first axis) and in Supplementary Figure S4 (second axis). Both figures were obtained by randomly permuting the 22 environmental variables, the eggNOG gene families at 23 functional levels of the gene ontology and the *proteobacteria* abundances at 102 order levels. Additionally, the explained variance supported by the first 15 axes is provided in Supplementary Figure S5.

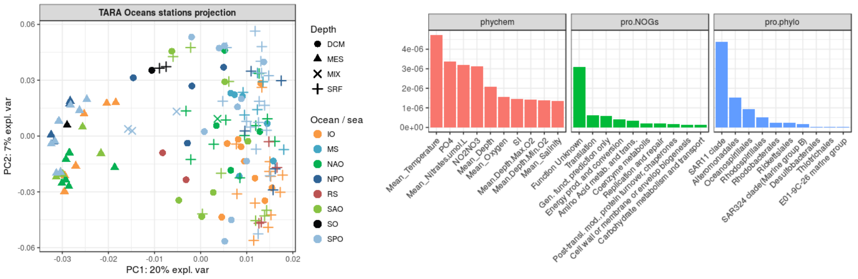

Figure 1. Only datasets of [40]. Left: Projection of the observations on the first two KPCA axes. Colors represent the oceanic regions and shapes the depth layers. Right: The 10 most important variables for the first KPCA axis, ranked by decreasing Crone-Crosby distance.

First, note that Figure 1 shows very similar results to the ones returned by the PCA performed on community composition dissimilarities (Bray-Curtis) presented in [40]: samples are separated by their depth layer of origin, *i.e.*, SRF, DCM or MES, with stronger differences for MES samples.

Figure 1 exhibits that both the abundance of *clade SAR11* and the temperature lead to the largest Crone-Crosby distances, meaning that they contribute the most to the first KPCA axis definition. This result is validated by displaying the values of this variable on the KPCA projection (see Supplementary Figures S6 and S7). On both figures, a gradient can be observed on the first KPCA axis between the left (lowest abundances of *clade SAR11* and lowest temperatures), and the right (highest values of these variables). Those results are similar to the ones presented in [40]: the vertical stratification of prokaryotic communities is mostly driven by temperature and *proteobacteria* (more specifically *clade SAR11* and *clade SAR86*) dominate the sampled areas.

Similarly, Supplementary Figure S4 shows that *cyanobacteria* abundance and the nitracline mean depth (*i.e.* water layer in which the nitrate concentration changes rapidly with depth) contribute the most to the second KPCA axis definition. The display of the nitracline mean depth on KPCA projection (Supplementary Figure S8) shows a gradient on the second KPCA axis. Supplementary Figure S9, displaying *cyanobacteria* abundance, shows a gradient between the top-left and the bottom-right of the KPCA projection, because *cyanobacteria* abundance also ranks as the third important variable on the first axis (see Supplementary Figure S10). Those results are consistent with findings of [40]: *cyanobacteria* were found abundant and the nitracline strongly correlated to the taxonomic composition (p-value ¡ 0.001). On both first two axes of the KPCA, unknown functions lead to the largest Crone-Crosby distances between variables used to compute the **pro.NOGs** kernel. Again, this result is in agreement with a conclusion made in [40]: a large fraction of the ocean gene families encode for unknown functions.

These results demonstrate that the proposed method gives a fast and accurate insight to the main variability explaining the differences between the different samples, viewed through different omics datasets. In particular, for both **pro.phylo** and **phychem** kernels, the most important variables are those used in [40] to state the main conclusions.

4.2 Integrating environmental, prokaryotic, eukaryotic and viral datasets

In this section, environmental, prokaryotic, eukaryotic and viral datasets are integrated together into a meta-kernel obtained using the full-UMKL method. Figure 2 (left) displays the projection of the samples on the first two axes of the KPCA. Figure 2 (right) and Supplementary Figure S11 provide the 5 most important variables for each datasets, respectively for the first and the second axes of the KPCA. To obtain these figures, abundance values were permuted at 56 prokaryotic phylum levels for the **pro.phylo** kernel, at 13 eukaryotic phylum levels for **euk.pina**, **euk.nano**, **euk.micro** and **euk.meso** and at 36 virus family levels for the **vir.VCs** kernel. Variables used for **phychem** and **pro.NOGs** were the same than in Section 4.1. Additionally, the explained variance supported by the first 15 axes is provided in Supplementary Figure S12.

First, note that Figure 2 does not highlight anymore any particular pattern in terms of depth layers but it does in terms of geography. SO samples are gathered in the bottom-center of the KPCA projection and SPO samples are gathered on the top-left side. Second, Figure 2 shows that the most important variables come from the **phychem** kernel (especially the longitude) and from kernels representing the eukaryotic plankton. More specifically, large size organisms are the most important: *rhizaria* phylum for **euk.meso** and *alveolata* phylum for **euk.nano**. The abundance of *rhizaria* organisms also ranks first between important variables of the second KPCA axis, followed by the *opisthokonta* phylum for **euk.nano**. The display of these variables on the KPCA projection reveals a gradient on the first axis for both the *alveolata* phylum abundance (Supplementary Figure S13) and the longitude (Supplementary Figure S14) and on the second axis for *rhizaria* (Supplementary Figure S15) and *opisthokonta* (Supplementary Figure S16) abundances. This indicates that SO and SPO epipelagic waters mainly differ in terms of *Rhizarians* abundances and both of them differ from the other studied waters in terms of *alveolata* abundances.

The integration of *TARA* Oceans datasets shows that the variability between epipelagic samples is mostly driven by geography rather than environmental factors and that this result is mainly explained by the strong geographical structure of large eukaryotic communities. Studied samples were all collected from epipelagic layers, where water temperature does not vary much, which explains the poor influence of the prokaryotic dataset in this analysis.

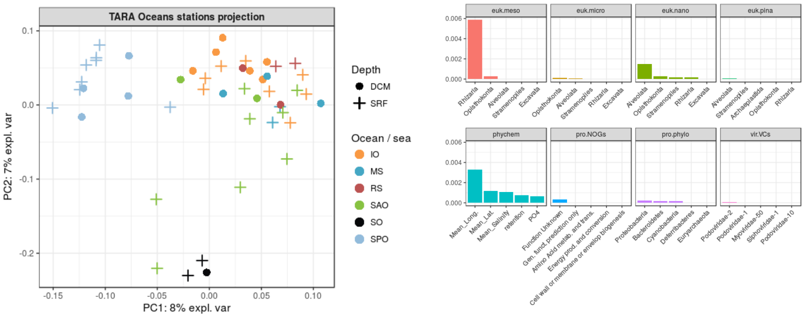

Figure 2. Left: Projection of the observations on the first two KPCA axes. Colors represent the oceanic regions and shapes the depth layers. Right: The 5 most important variables for each of the eight datasets, ranked by decreasing Crone-Crosby distance.

5 Conclusion

The contributions of the present manuscript to the analysis of multi-omics datasets are twofolds: firstly, we have proposed three unsupervised kernel learning approaches to integrate multiple datasets from different types, which either allow to learn a consensus meta-kernel or a meta-kernel preserving the original topology of the data. Secondly, we have improved the interpretability of the KPCA by assessing the influence of input variables in a generic way.

The experiments performed on *TARA* Oceans datasets showed that presented methods allow to give a fast and accurate insight over the different datasets within a single analysis. However, the approach is not restricted to KPCA analyses: the meta-kernel presented in this article could have been used in combination with kernel clustering methods or with kernel supervised models, to integrate multi-omics datasets.

6 Acknowledgments

The authors are grateful to Rémi Flamary (University of Nice, France) for helpful suggestions and comments on the optimization and to the **mixOmics** team, especially to Kim-Anh Le Cao and Florian Rohart, for helping to integrate our method in the **mixOmics** package.

References

1. J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–177, 1982.

2. N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

3. M. Arumugam, , J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473:174–180, 2011.

4. A. Ben-Hur and J. Weston. *Data Mining Techniques for the Life Sciences*, volume 609 of *Methods in Molecular Biology*. Springer-Verlag, 2010.

5. P. Bork, C. Bowler, C. de Vargas, G. Gorsky, E. Karsenti, and P. Wincker. Tara oceans studies plankton at planetary scale. *Science*, 348(6237):873–873, 2015.

6. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alterning direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

7. R. Bray and J. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

8. J. Brum, J. Ignacio-Espinoza, S. Roux, G. Doulcier, S. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. de Vargas, J. Gasol, G. Gorsky, A. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. Poulos, S. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, *Tara* Oceans coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, and M. Sullivan. Patterns and ecological drivers of ocean viral communities. *Science*, 348(6237), 2015.

9. J. Chen and H. Li. Kernel methods for regression analysis of microbiome compositional data. In L. Hu, Y. Liu, and J. Lin, editors, *Topics in Applied Statistics*, volume 55 of *Springer Proceedings in Mathematics & Statistics (PROMS)*, pages 191–201, New York, NY, USA, 2013. Springer.

10. D. L. Cox-Foster, S. Conlan, E. C. Holmes, G. Palacios, J. D. Evans, N. A. Moran, P.-L. Quan, T. Briese, M. Hornig, D. M. Geiser, V. Martinson, D. vanEngelsdorp, A. L. Kalkstein, A. Drysdale, J. Hui, J. Zhai, L. Cui, S. K. Hutchison, J. F. Simons, M. Egholm, J. S. Pettis, and W. I. Lipkin. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, 318(5848):283–287, 2007.

11. L. Crone and D. Crosby. Statistical applications of a metric on subspaces to satellite meteorology. *Technometrics*, 37(3):324–328, 1995.

12. C. de Vargas, S. Audic, N. Henry, J. Decelle, P. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, *Tara* Oceans coordinators, S. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 2015.

13. S. Dray, S. Pavoine, and D. Aguirre de Cárcer. Considering external information to improve the phylogenetic comparison of microbial communities: a new approach based on constrained double principal coordinates analysis (cDPCoA). *Molecular Ecology Resources*, 15(2):242–249, 2014.

14. N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, 109(52):21390–21395, 2012.

15. M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, July 2011.

16. L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, A. S., L. Berline, J. Brum, L. Coelho, J. Cesar, I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, F. Kandels-Lewis, J. Picheral, M. Poulain, S. Searson, T. O. C. Coordinators, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, E. Pesant, J. Weissenbach, P. Wincker, S. Acinas, P. Bork, C. de Vargas, D. Iudicone, M. Sullivan, J. Raes, R. Karsenti, and G. Bowler, C.and Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532:465–470, May 2016.

17. C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, M. G. Giglio, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. Bonazzi, J. Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. Chain, I. M. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. M. Dunne, A. Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. J. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, S. K. Haake, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, N. B. King, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C. C. Lo, C. A. Lozupone, R. Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavromatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. Pop, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y. H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren,

R. A. Gibbs, S. K. Highlander, B. A. Methe, K. E. Nelson, J. F. Petrosino, G. M. Weinstock, R. K. Wilson, and O. White. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

18. P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

19. E. Karsenti, S. Acinas, P. Bork, C. Bowler, C. de Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. Reynaud, C. Sardet, M. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, P. Wincker, and Tara Oceans Consortium. A holistic approach to marine eco-systems biology. *PLoS Biology*, 9(10):e1001177, 2011.

20. C. Lavit, Y. Escoufier, R. Sabatier, and P. Traissac. The act (statis method). *Computational Statistics & Data Analysis*, 18(1):97 – 119, 1994.

21. K. Lê Cao, M. Costello, V. Lakis, F. Bartolo, X. Chua, R. Brazeilles, and P. Rondeau. mixMC: a multivariate statistical framework to gain insight into microbial communities. *PloS One*, 11(8):e0160169, 2016.

22. J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York; London, 2007.

23. H. L'Hermier des Plantes. *Structuration des tableaux à trois indices de la statistique*. PhD thesis, Université de Montpellier, 1976. Thèse de troisième cycle.

24. G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. De Meester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, and J. Raes. Determinants of community structure in the global plankton interactome. *Science*, 348(6237), 2015.

25. Y. Lin, T. Liu, and C. Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1147–1160, 2010.

26. C. Lozupone, M. Hamady, S. Kelley, and R. Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

27. C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

28. S. Mandal, W. van Treuren, R. White, M. Eggesbø, R. Knight, and S. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:27663, 2015.

29. P. McMurdie and S. Holmes. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, 8(4):e61217, 2013.

30. S. Pavoine. A guide through a family of phylogenetic dissimilarity measures among sites. *Oikos*, 125:1719–1732, 2016.

31. S. Pavoine, A. Dufour, and D. Chessel. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, 228(4):523–537, 2004.

32. S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, 40(D1):D284, 2012.

33. M. Price, P. Dehal, and A. Arkin. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3):e9490, 2010.

34. T. Randolph, S. Zhao, W. Copeland, M. Hullar, and A. Shojaie. Kernel-penalized regression for analysis of microbiome data. Preprint arXiv:1511.00297, 2017.

35. S. Ren, P. Ling, M. Yang, Y. Ni, and Z. Zong. Multi-kernel PCA with discriminant manifold for hoist monitoring. *Journal of Applied Sciences*, 13(20):4195–4200, 2013.

36. F. Reverter, E. Vegas, and J. Oller. Kernel-PCA data integration with enhanced interpretability. *BMC Systems Biology*, 8, 2014.

37. P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied Statistics*, 25(3):257–265, 1976.

38. S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, S. Pesant, S. Kandels-Lewis, C. Dimier, M. Picheral, S. Searson, C. Cruaud, A. Alberti, C. M. M. Duarte, J. M. M. Gasol, D. Vaqué, P. Bork, S. G. Acinas, P. Wincker, and M. B. Sullivan. Ecogenomics and biogeochemical impacts of uncultivated globally abundant ocean viruses. *Nature*, 537:689–693, 2016.

39. B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

40. S. Sunagawa, L. Coelho, S. Chaffron, J. Kultima, K. Labadie, F. Salazar, B. Djahanschiri, G. Zeller, D. Mende, A. Alberti, F. Cornejo-Castillo, P. Costea, C. Cruaud, F. d'Oviedo, S. Engelen, I. Ferrera, J. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, *Tara* Oceans coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. Acinas, and P. Bork. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.

41. W. Togerson. *Theory & Methods of Scaling*. Wiley, New York, NY, USA, 1958.

42. E. Villar, G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B. Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, F. d'Ortenzio, J.-P. Gattuso, N. Grima, L. Guidi, C. N. Hill, O. Jahn, J.-L. Jamet, H. Le Goff, C. Lepoivre, S. Malviya, E. Pelletier, J.-B. Romagnan, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, T. Vannier, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis,

S. G. Acinas, P. Bork, E. Boss, C. de Vargas, G. Gorsky, H. Ogata, S. Pesant, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, and D. Iudicone. Environmental characteristics of agulhas rings affect interocean plankton transport. *Science*, 348(6237), 2015.

43. Z. Wang, S. Chen, and T. Sun. MultiK-MHKS: a novel multiple kernel learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):348–353, 2008.

44. G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

45. B. Zhao, J. Kwok, and C. Zhang. Multiple kernel clustering. In C. Apte, H. Park, K. Wang, and M. Zaki, editors, *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*, pages 638–649, Philadelphia, PA, 2009. SIAM.

46. J. Zhuang, J. Wang, S. Hoi, and X. Lan. Unsupervised multiple kernel clustering. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 20:129–144, 2011.