Integrating *TARA* Oceans datasets using unsupervised multiple kernel learning: supplementary material

Jérôme Mariette and Nathalie Villa-Vialaneix

1 Unsupervised multiple kernel and KPCA in mixOmics

Methods presented in the paper are available on CRAN in the R package **mixKernel** and a full tutorial on the **mixOmics** R package WEB site at http://mixomics.org/mixkernel/. Kernels can be computed using the function **compute.kernel** that allows to choose between linear, phylogenic and abundance kernels. Unifrac and weighted Unifrac distances are processed using functions taken from the **phyloseq** package [McMurdie and Holmes, 2013]. Bray-Curtis dissimilarities are computed with the **vegan** package. The function **combine.kernels** implements methods described in Section 2.1 and returns a meta-kernel which can be used as an input for the function **kernel.pca**. The KPCA result can then be displayed using the **mixOmics** plot function **plotInd**.

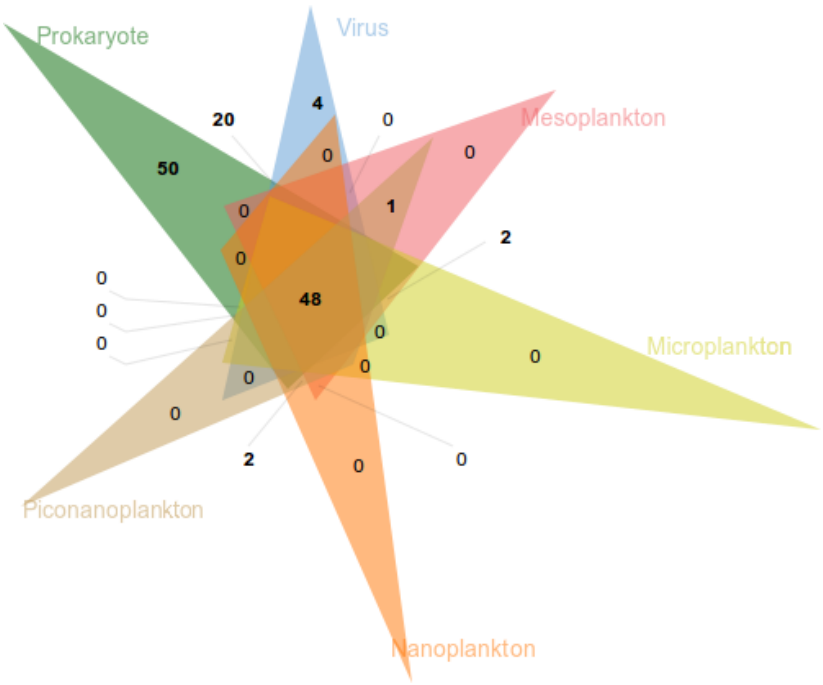
To assess variable influence in the different datasets, the function **kernel.pca.permute** computes Crone-Crosby distances resulting from permutations. In this function, the user can specify the level at which the permutations must be performed. The most important variables can then be plotted using the **plotVar mixOmics** function. A subset of *TARA* Oceans datasets and a tutorial are also provided in the package to help users processing their own data. In addition, the tutorial is also available on the **mixOmics** web site `http://mixomics.org/mixkernel/` and the method is scheduled to be part of the next version of **mixOmics**.

2 Selected samples

Ocean samples used in [Sunagawa et al., 2015, de Vargas et al., 2015, Brum et al., 2015, Roux et al., 2016] were collected at various locations, representing all main oceanic regions at different depth layers. Collected samples were located in height different oceans or seas: indian ocean (IO), mediterranean sea (MS), north atlantic ocean (NAO), north pacific ocean (NPO), red sea (RS), south atlantic ocean (SAO), south pacific ocean (SPO) and south ocean (SO).

[Sunagawa et al., 2015] focused on 139 prokaryotic-enriched samples collected from 68 stations and spread across three depth layers: the surface (SRF), the deep chlorophyll maximum (DCM) layer and the mesopelagic (MES) zones. In [de Vargas et al., 2015], 334 size-fractionated samples were analyzed from 47 stations at two water-column depths of the photic-zone: SRF and DCM. The different size-fractions filters used during the sampling allowed to split samples into four major eukaryotic organism sizes: piconanoplankton, nanoplankton, microplankton and mesoplankton. Finally, [Brum et al., 2015] and [Roux et al., 2016] analyzed respectively 43 and 89 viral-fractioned samples, collected from 45 stations from the SRF, the DCM and the MES layers.

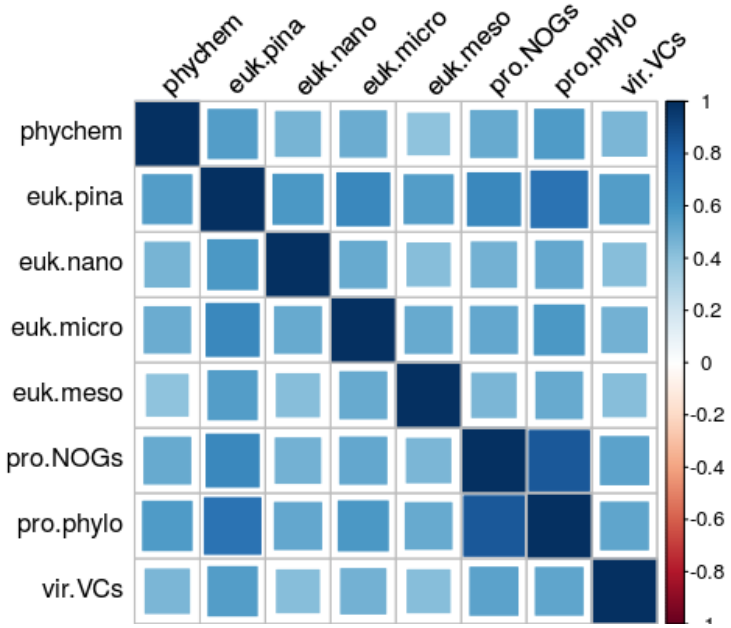
To evaluate the performances of the proposed methods from different points of view, two analyses were performed. First, the 139 prokaryotic samples were used as inputs of the proof-of-concept analysis presented in Section 4.1. Then, a more complete analysis is presented in Section 4.2. This analysis was performed on the whole available material but only samples for which all the prokaryotic, eukaryotic and viral information was available. As shown in Supplementary Figure S1, this resulted in 48 common sampling locations which included two depth layers (SRF and MES) and 31 stations.



Supplementary Figure S1: Common sampling locations among prokaryotic, eukaryotic and viral samples. Figure was obtained using jvenn [Bardou et al., 2014].

3 Similarities between kernels

To have a general overview on the 8 datasets to integrate, the similarity measure between kernels defined in Equation (2) is computed. The pairwise values are displayed in Supplementary Figure S2.



Supplementary Figure S2: Similarities between kernels computed using the STATIS-UMKL approach.

The figure shows that **pro.phylo** and **pro.NOGs** are the most correlated pair of kernels. This result is expected as both kernels provide a summary of prokaryotic communities. Second, the kernel that is the less correlated (in average) with the other ones is **euk.meso** and the kernel that is the most correlated (in average) with the other ones is **euk.pina**. These facts are supported by the

conclusions stated in [de Vargas et al., 2015]: mesoplanktonic communities are strongly geographically structured, according to their basin of origin, whereas piconanoplankton communities are more homogeneous across the world oceans.

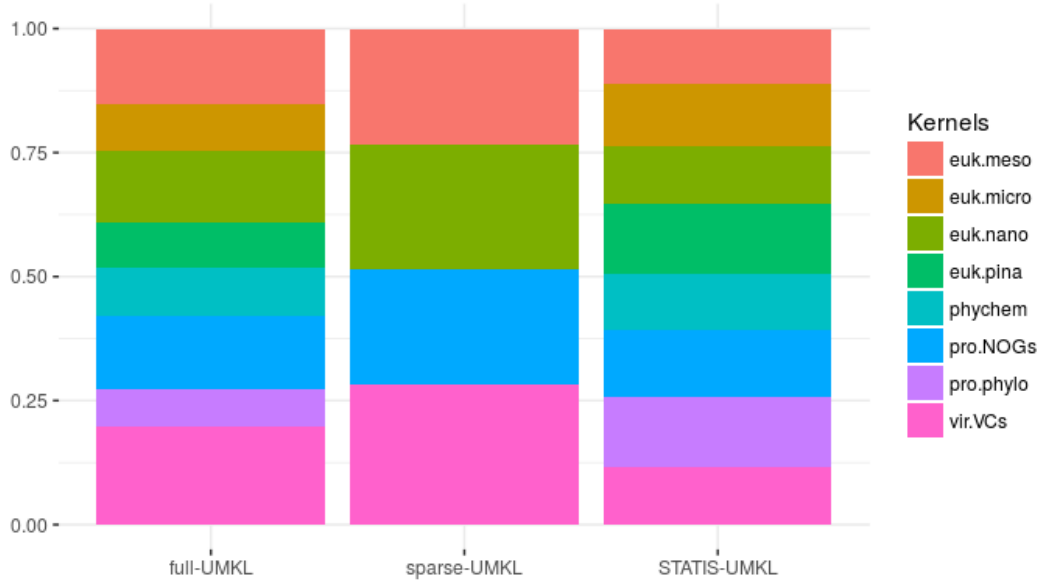
When focusing on similarities to environmental and physical variables, as measured by **phychem**, the figure shows that the kernels that are the most correlated to this kernel are **pro.phylo** and **euk.pina** kernels and that, again, **euk.meso** provides a different image of the oceans. These results are supported by a conclusion made in [Sunagawa et al., 2015] and [de Vargas et al., 2015]: the vertical stratification of the ocean microbiome is mainly driven by temperature rather than geography, but geography plays a strong role to structure communities with respect to the large organism size fractions.

Finally, **vir.VCs** is also more similar to small size organisms kernels than kernels representing larger ones. This is explained by the fact that the biographical structure of viruses is due to host community structure and to a passive transport by oceanic currents [Brum et al., 2015].

These results confirm the discussion reported in Supplementary Section S4: STATIS-UMKL allows to have an overview on the different datasets and should be used when the integrated analysis focuses on correlated informations.

4 Comparison of the different integration options

In the following section, the different methods proposed and especially the relevance of using a specific approach to perform the integration is evaluated. To perform this analysis, environmental, prokaryotic, eukaryotic and viral datasets are integrated together using the three proposed approaches: full-UMKL, sparse-UMKL and STATIS-UMKL. The weights obtained for each methods are presented in Supplementary Figure S3.

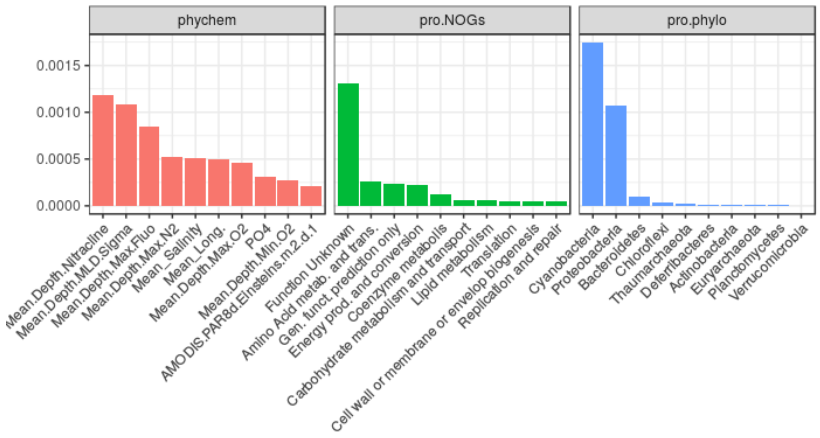


Supplementary Figure S3: Kernels weights obtained for the three proposed approaches: full-UMKL, sparse-UMKL and STATIS-UMKL. Colors represent the different kernels.

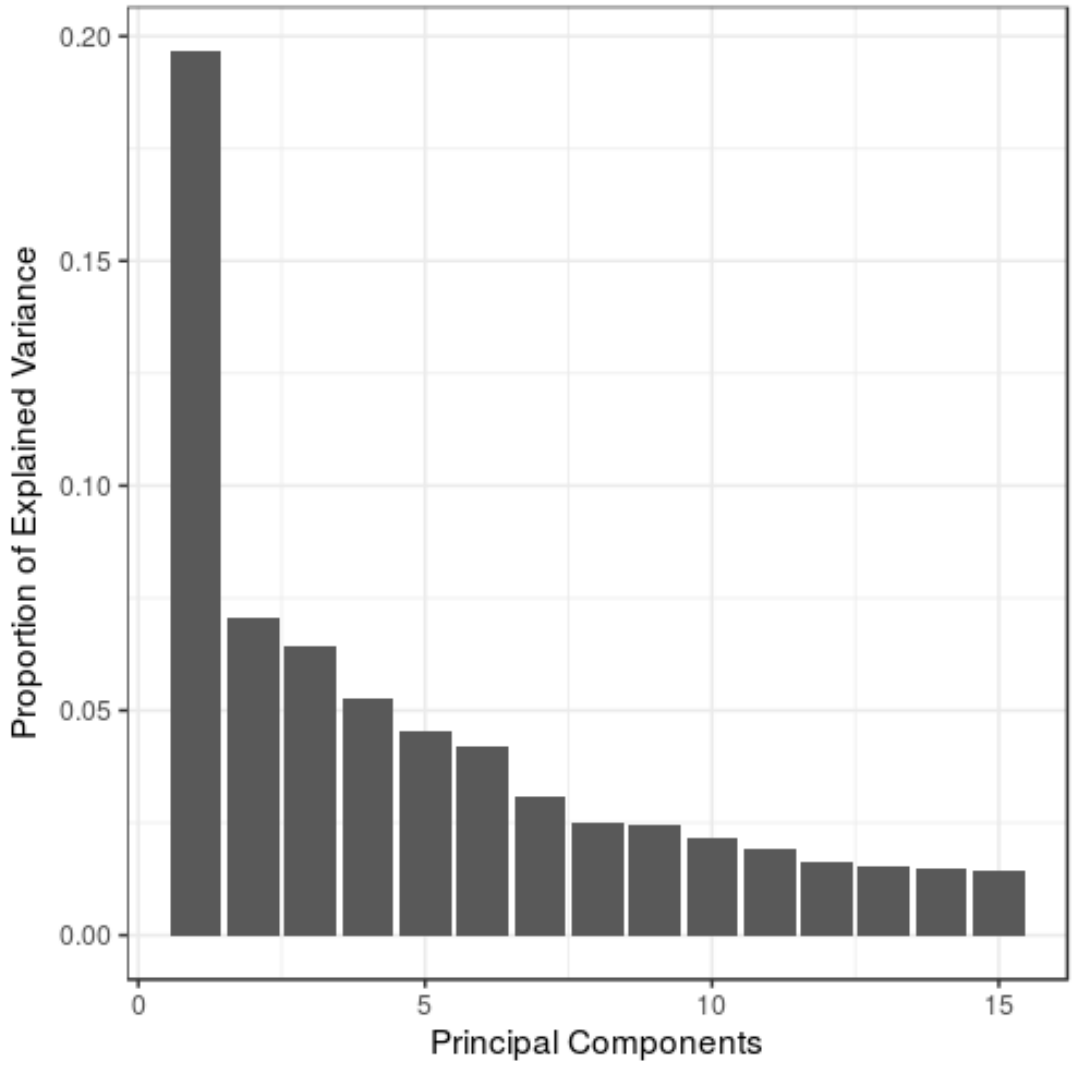
First, note that, Supplementary Figure S3 shows that STATIS-UMKL gives more weights to **euk.micro**, **euk.pina**, **pro.NOGs** and **pro.phylo**, meaning that these kernels are strongly correlated. In the contrary, full-UMKL gives more importance to atypical kernels, *i.e.*, **euk.meso**, **euk.micro**, **pro.NOGs** and **vir.VCs**, which are the only kernels selected by the sparse-UMKL approach, the other ones being discarded from the final meta-kernel.

Results show that the three proposed methods are complementary and can be used depending on the research question and the analysis step. The STATIS-UMKL approach allows to have an overview on the correlation between the different datasets to analyze and to integrate them in a consensual way. sparse-UMKL can be used to focus on a more even contribution of the various images provided by the different kernels and to remove redundant informations. Finally, a similar goal is achieved with the full-UMKL method, that should be preferred when the analysis requires to be performed on the whole material.

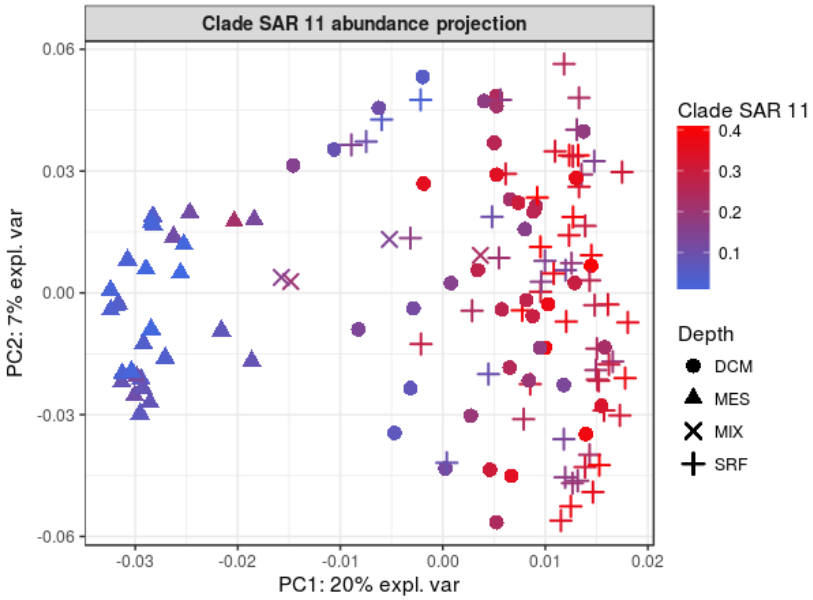
5 Supplementary figures



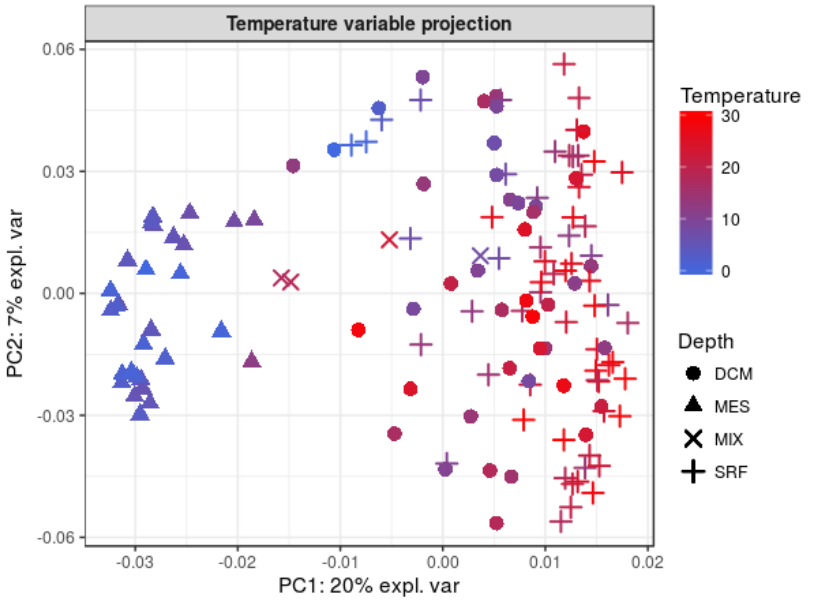
Supplementary Figure S4: **Only datasets of [Sunagawa et al., 2015]**. The 10 most important variables for the second KPCA axis, ranked by decreasing Crone-Crosby distance. Variables of the **pro.phylo** kernel were permuted at the phylum level.



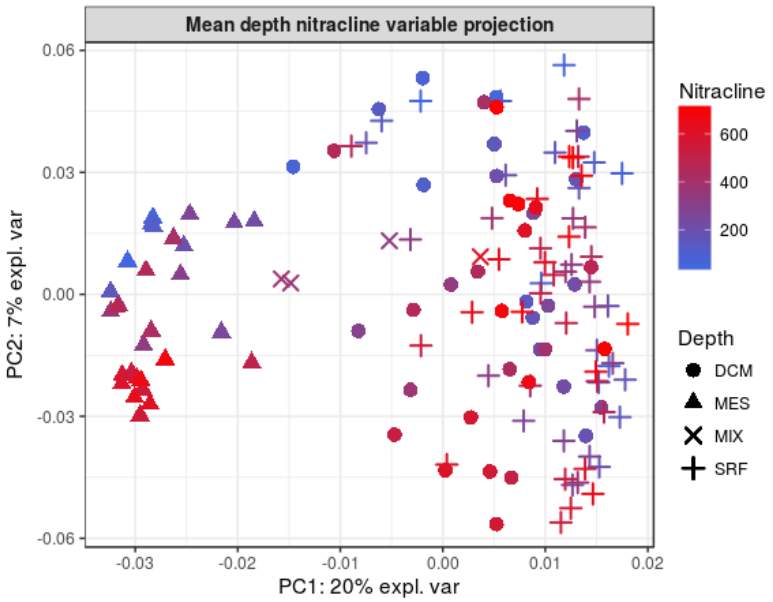
Supplementary Figure S5: **Only datasets of [Sunagawa et al., 2015]**. Entropy preserved by the 15 first axes of the KPCA performed on the meta-kernel obtained using the full-UMKL approach.



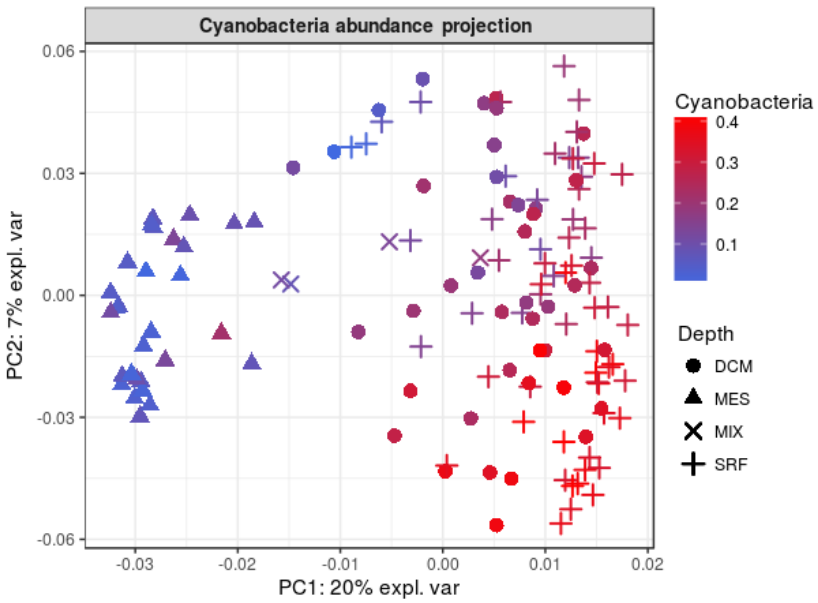
Supplementary Figure S6: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colors represent the relative abundance of *clade SAR11*: blue for low values and red for high values.



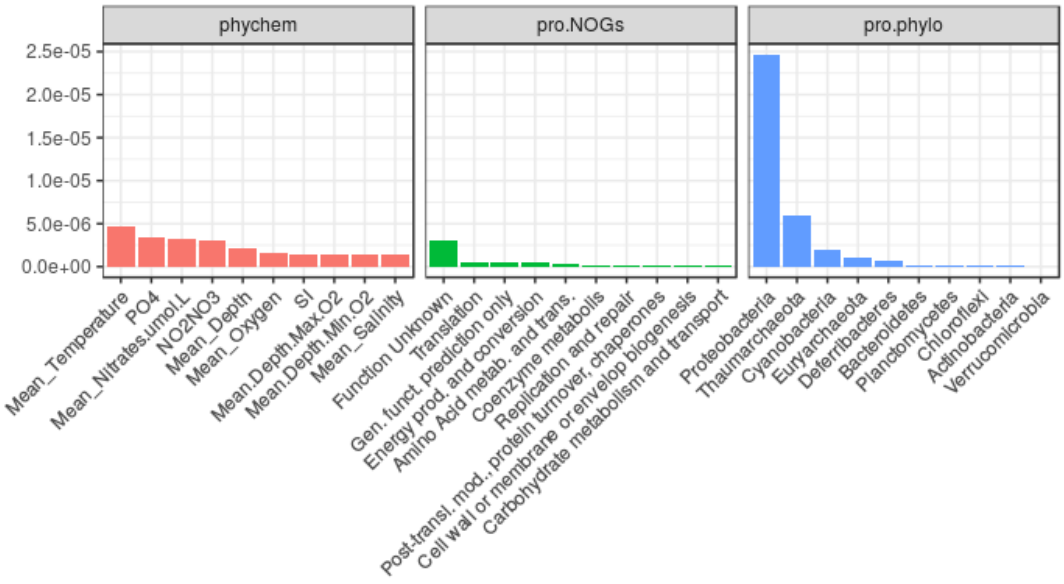
Supplementary Figure S7: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colors represent the temperature: blue for cold waters and red for warm waters.



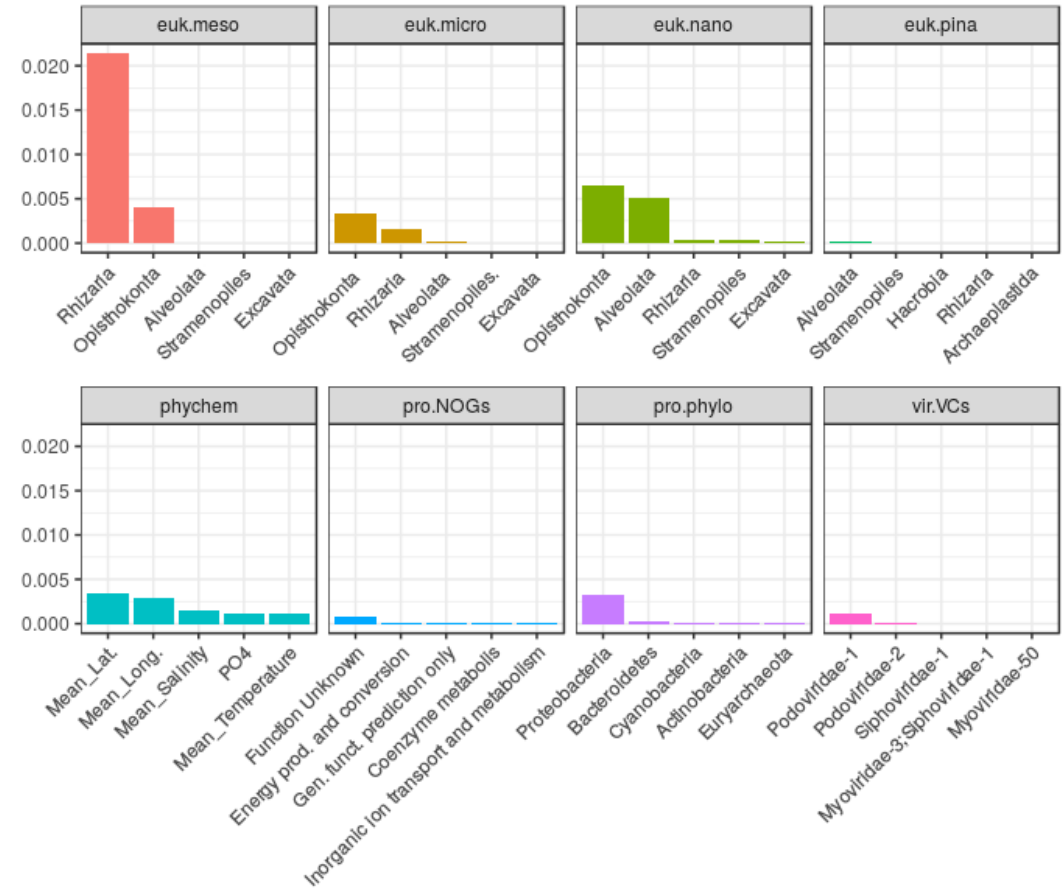
Supplementary Figure S8: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colors represent the nitracline mean depth: blue for low values and red for high values.



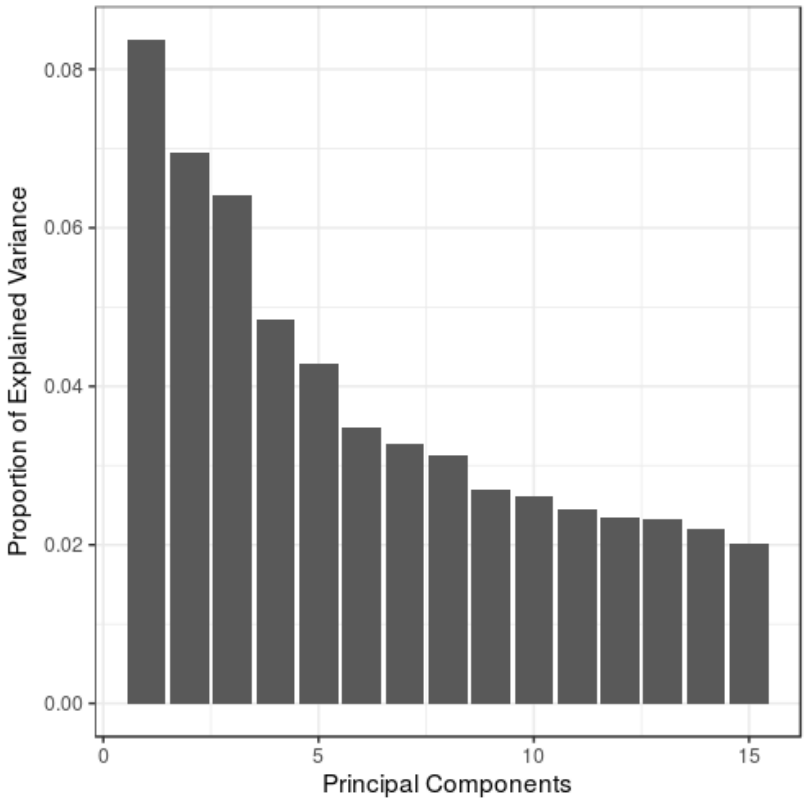
Supplementary Figure S9: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colors represent the relative abundance of *cyanobacteria*: blue for low values and red for high values.



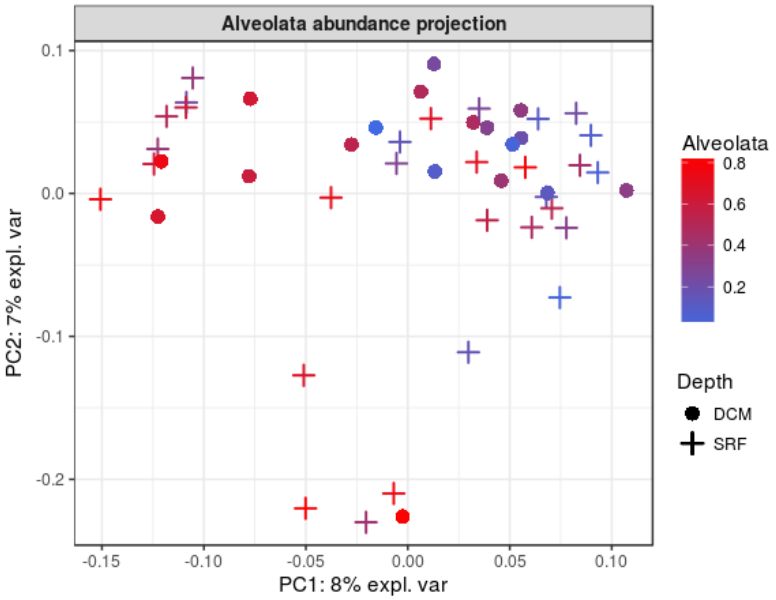
Supplementary Figure S10: **Only datasets of [Sunagawa et al., 2015]**. The 10 most important variables for the second axis of KPCA, ranked by decreasing Crone-Crosby distance. Variables of the **pro.phylo** kernel were permuted at the phylum level.



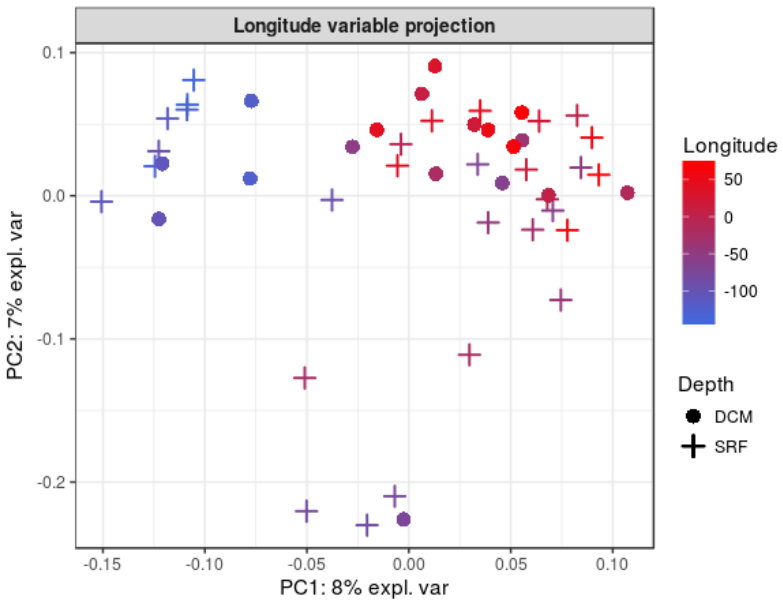
Supplementary Figure S11: The 5 most important variables for the second axis of the KPCA and for each of the 8 datasets, ranked by decreasing Crone-Crosby distance.



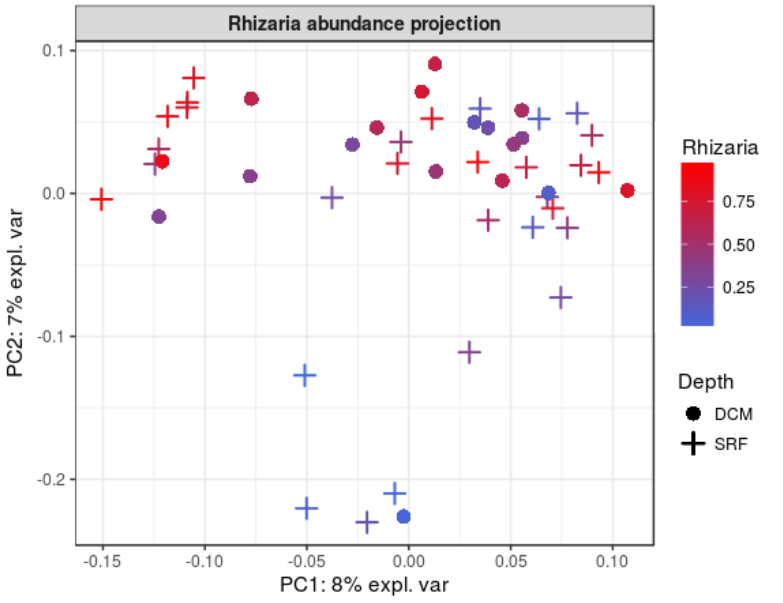
Supplementary Figure S12: Entropy preserved by the 15 first axes of the KPCA performed on the meta-kernel obtained using the full-UMKL approach and environmental, prokaryotic, eukaryotic and viral datasets.



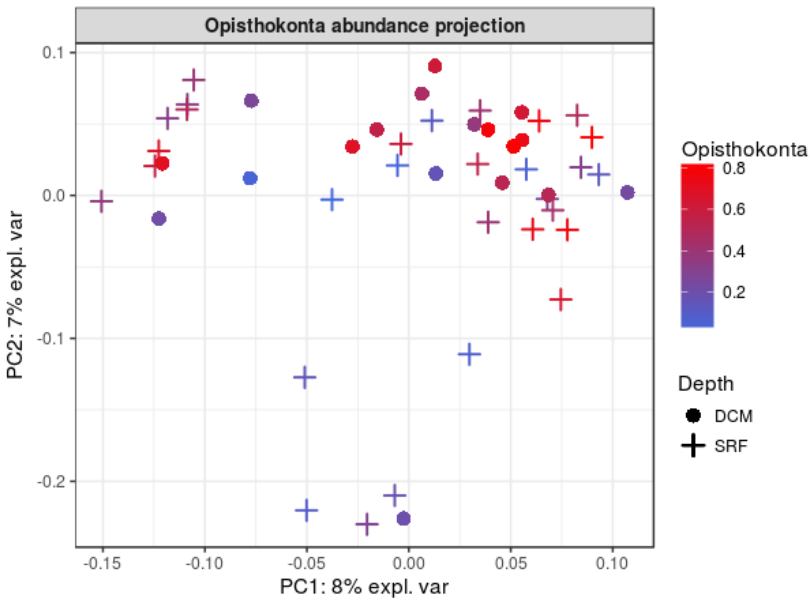
Supplementary Figure S13: Projection of the observations on the first two KPCA axes. Colors represent the relative abundance of *alveolata* organisms in the nanoplanktonic community: blue for low values and red for high values.



Supplementary Figure S14: Projection of the observations on the first two KPCA axes. Colors represent the longitude: blue for low values and red for high values.



Supplementary Figure S15: Projection of the observations on the first two KPCA axes. Colors represent the relative abundance of *rhizaria* organisms in the mesoplanktonic community: blue for low values and red for high values.



Supplementary Figure S16: Projection of the observations on the first two KPCA axes. Colors represent the relative abundance of *opisthokonta* organisms in the nanoplanktonic community: blue for low values and red for high values.

References

[Bardou et al., 2014] Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive venn diagram viewer. *BMC bioinformatics*, 15(1):293.

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

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

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

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

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