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, phylogenetic 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 *TARA* Oceans selected datasets and 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. The analysis, presented in Section 4.1, was performed on the whole available material but only with samples for which all the prokaryotic, eukaryotic and viral information was available: in [de Vargas et al., 2015], 334 size-fractionated samples were analysed 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] analysed respectively 43 and 89 viral-fractioned samples, collected from 45 stations from the SRF, the DCM and the MES layers. As shown in Supplementary Figure S1, this resulted in 48 common sampling locations which included two depth layers (SRF and MES) and 31 stations. From these selected samples, 8 dissimilarities 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.*, temperature, salinity. This dataset was extracted from Table W8, available on the companion website of [Sunagawa et al., 2015][1]. Missing values were previously imputed using a k-nearest neighbour

[1] `http://ocean-microbiome.embl.de/companion.html`

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 [Sunagawa et al., 2015][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 phylogenetic tree was built from these data using fasttree [Price et al., 2010]. The weighted Unifrac distance was then computed using the R package **phyloseq** [McMurdie and Holmes, 2013]: $d_{wUF}(x_i, x_j) = \frac{\sum_e l_e |p_e - q_e|}{\sum_e (p_e + q_e)}$, 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

$$d_{BC}(x_i, x_j) = \frac{\sum_s |n_{is} - n_{js}|}{\sum_s (n_{is} + n_{js})}, \tag{1}$$

  computed on the gene abundances of 39,246 bacterial genes. In Equation (1), 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 catalogue[2] and summarized at eggNOG gene families (genes annotated by eggNOG version 3 database: [Powell et al., 2012]). The gene abundance table is freely available from the companion website of [Sunagawa et al., 2015][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 (1), is computed on the abundance table of $\sim 150,000$ eukaryotic plankton OTUs. The dataset can be downloaded from the companion website of [de Vargas et al., 2015] [2];

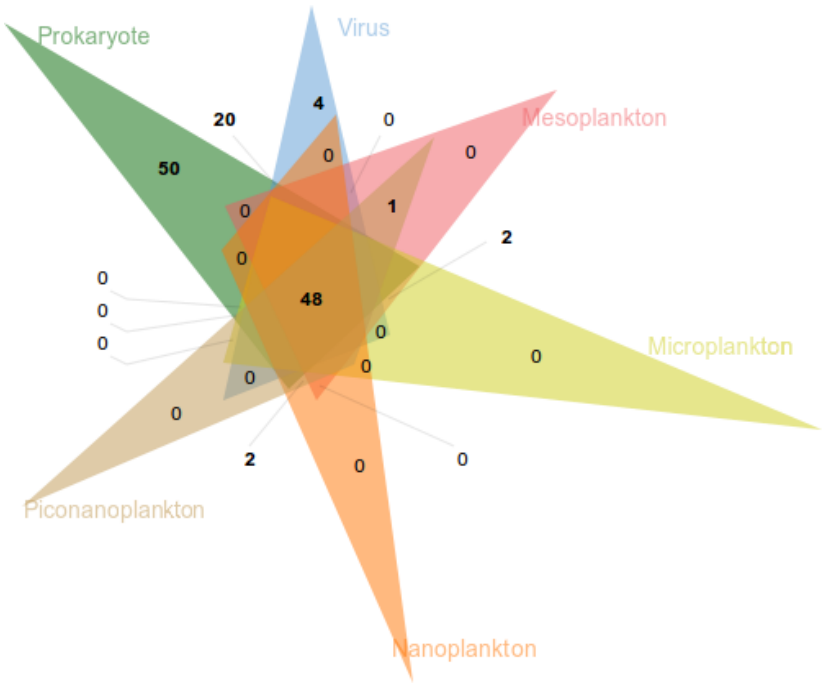
- The **vir.VCs** dissimilarity measures ocean viral communities and was computed using the Bray-Curtis dissimilarity, defined in Equation (1), on the abundance table of 867 Viral Clusters (VCs) available from the supplementary materials of [Roux et al., 2016].

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 [Lee and Verleysen, 2007]: $K_{ij} = -\frac{1}{2}\left(d(x_i, x_j) - \frac{1}{N}\sum_{k=1}^{N}(d(x_i, x_k) + d(x_k, x_j)) + \frac{1}{N^2}\sum_{k,k'=1}^{N} d(x_k, x_{k'})\right)$, 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 [Ben-Hur and Weston, 2010]: $\tilde{K}_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$.

3 Proof of concept with a restricted number of *TARA* Oceans datasets

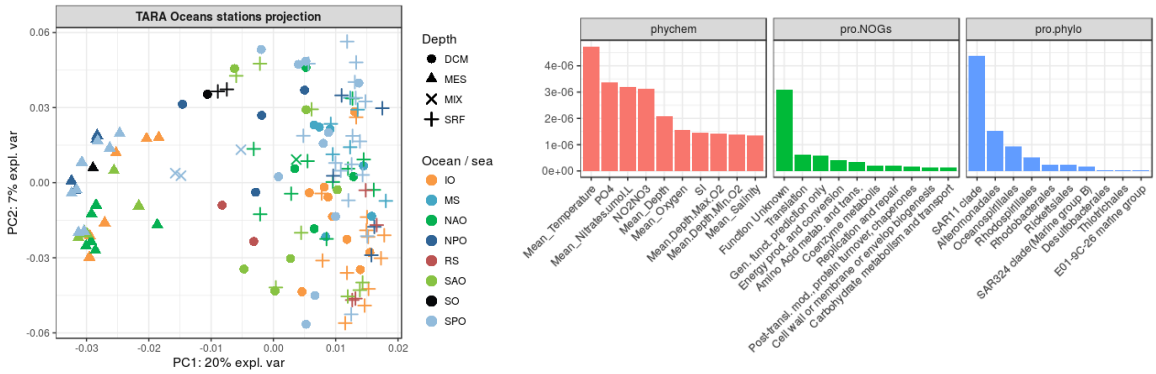
In the present section, only the datasets analysed in [Sunagawa et al., 2015] are explored. The results described in this paper are used as a ground truth to validate the relevance of our strategy. Kernels used are the environmental kernel, **phychem**, and the two prokaryotic kernels, **pro.phylo**

[2] `http://taraoceans.sb-roscoff.fr/EukDiv/`

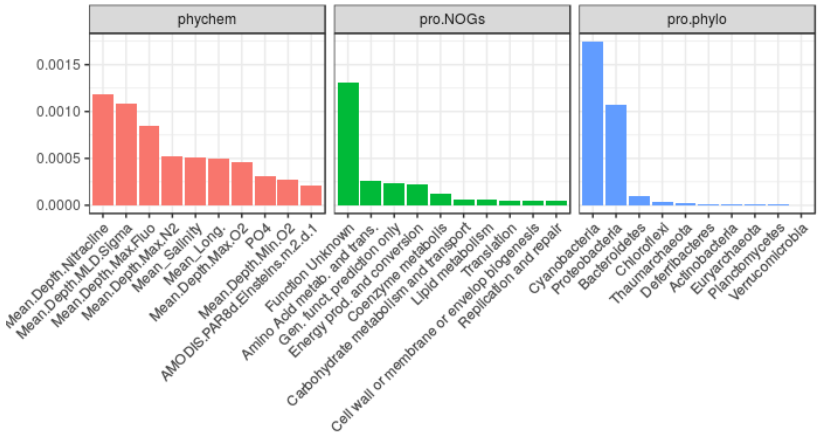


Supplementary Figure S1: Common sampling locations among prokaryotic, eukaryotic and viral samples. Figure was obtained using jvenn [Bardou et al., 2014].

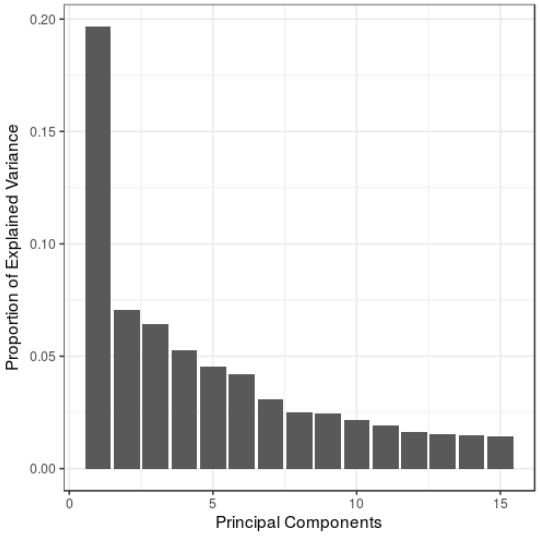
and **pro.NOGs** described in Supplementary Section S2. Kernels were computed on the 139 prokaryotic samples analysed in [Sunagawa et al., 2015], which were collected from 68 stations and spread across three depth layers: the surface (SRF), the deep chlorophyll maximum (DCM) layer and the mesopelagic (MES) zones. Supplementary Figure S2 (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 Supplementary Figure S2 (first axis) and in Supplementary Figure S3 (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 S4. Using the R package **mixKernel** on a 1 cpu computer with 16GB memory, the computational cost to combine the three kernels is $\sim$5 seconds. Permutations to assess the **pro.phylo** kernel important variables were performed in $\sim$9 hours without parallelization. This computational cost is due to the weighted Unifrac distance computation, which is high compared to others β-diversity measures such as the Bray-Curtis dissimilarity.



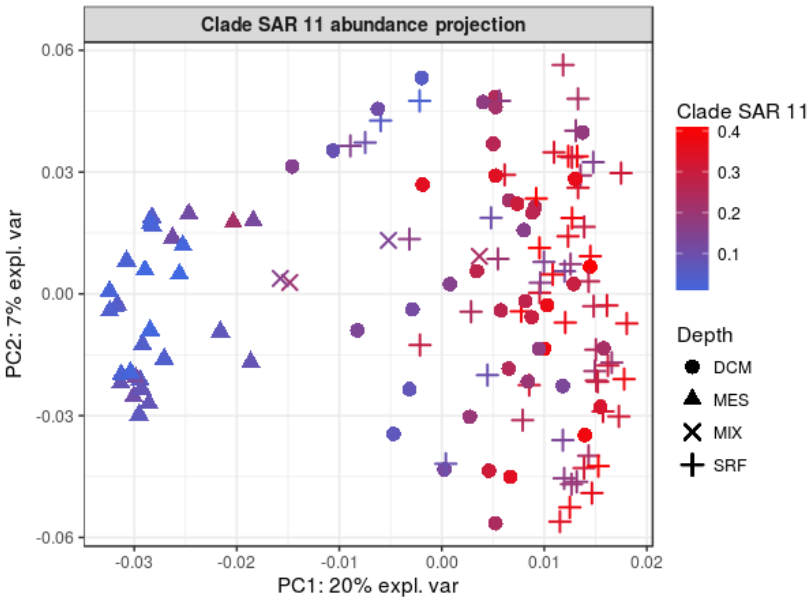
Supplementary Figure S2: **Only datasets of [Sunagawa et al., 2015]**. Left: Projection of the observations on the first two KPCA axes. Colours represent the oceanic regions and shapes the depth layers. Right: The 10 most important variables for the first KPCA axis, ranked by decreasing Crone-Crosby distance.



Supplementary Figure S3: **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 S4: **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.



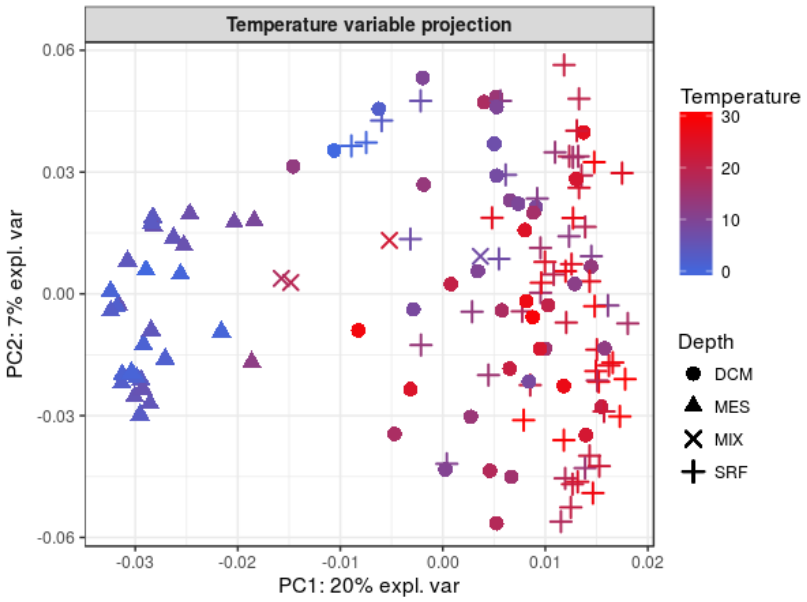
Supplementary Figure S5: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colours represent the relative abundance of *clade SAR11*: blue for low values and red for high values.

First, note that Supplementary Figure S2 shows very similar results to the ones returned by the PCA performed on community composition dissimilarities (Bray-Curtis) presented in [Sunagawa et al., 2015]: samples are separated by their depth layer of origin, *i.e.*, SRF, DCM or MES, with stronger differences for MES samples.

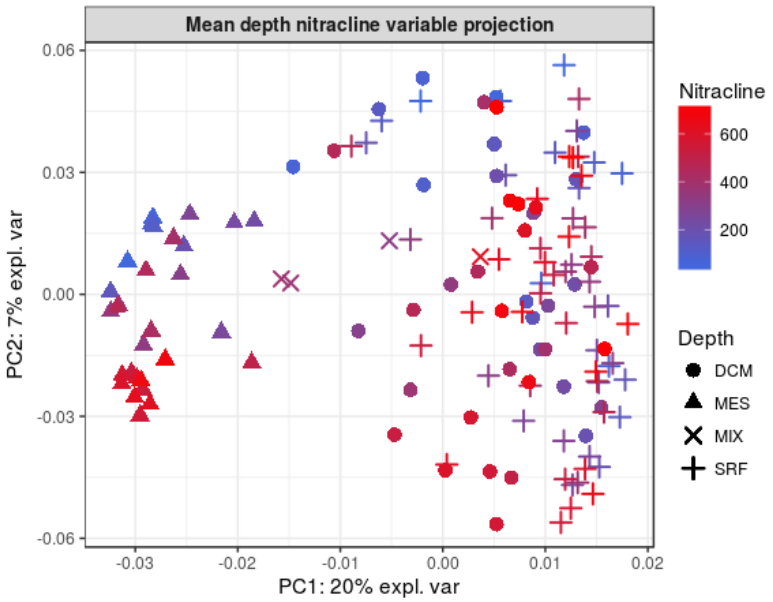
Supplementary Figure S2 exhibits that both the abundance of *clade SAR11* and the temperature lead to the largest Crone-Crosby distances, meaning that they contribute the most to the first KPCA axis definition. This result is validated by displaying the values of this variable on the KPCA projection (see Supplementary Figure S5 and S6). 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 [Sunagawa et al., 2015]: 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 S3 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 S7) shows a gradient on the second KPCA axis. Supplementary Figure S8, 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 S9). Those results are consistent with findings of [Sunagawa et al., 2015]: *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 [Sunagawa et al., 2015]: 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 [Sunagawa et al., 2015] to state the main conclusions.



Supplementary Figure S6: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colours represent the temperature: blue for cold waters and red for warm waters.



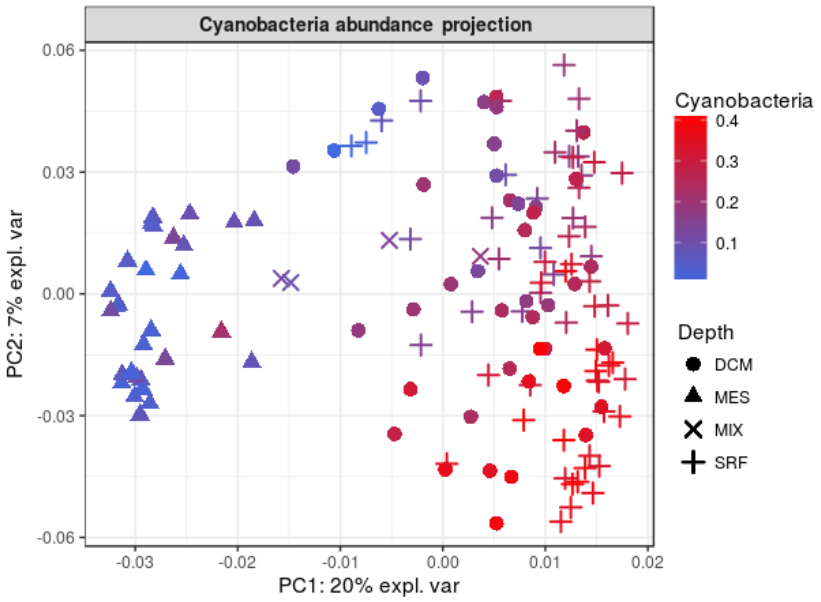
Supplementary Figure S7: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colours represent the nitracline mean depth: blue for low values and red for high values.

4 Similarities between kernels

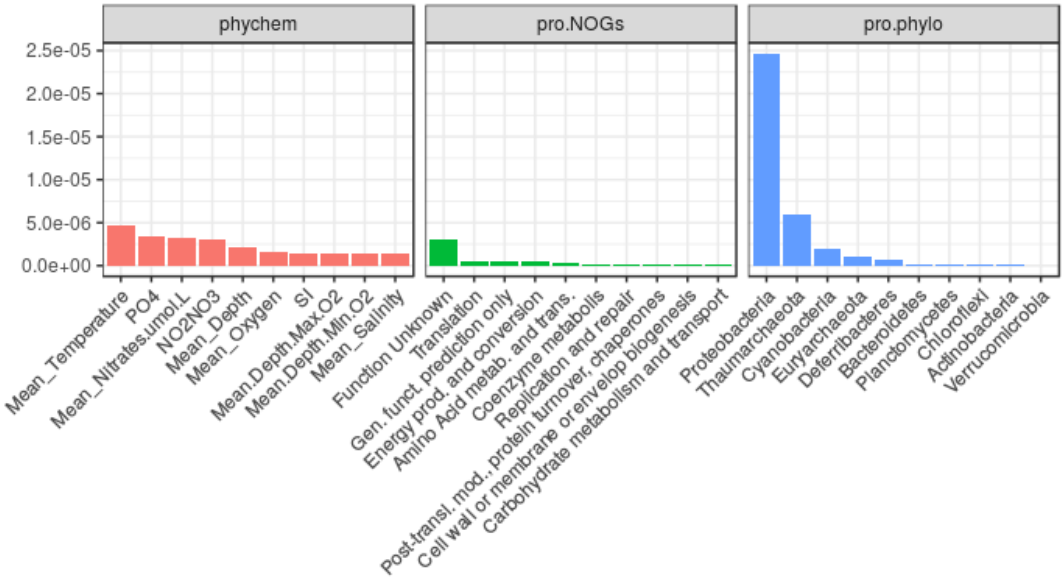
4.1 Similarities between *TARA* Oceans kernels

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 S10.

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



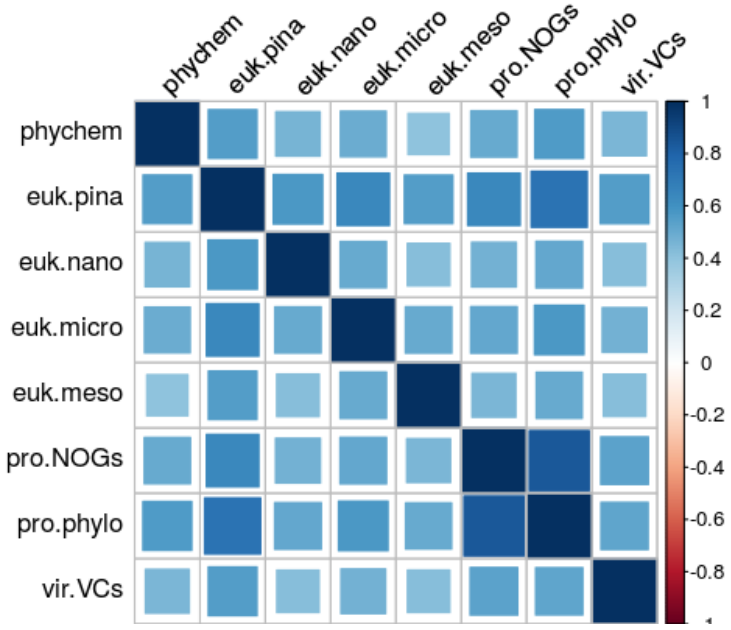
Supplementary Figure S8: **Only datasets of [Sunagawa et al., 2015]**. Projection of the observations on the first two KPCA axes. Colours represent the relative abundance of *cyanobacteria*: blue for low values and red for high values.



Supplementary Figure S9: **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.

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]:



Supplementary Figure S10: Similarities between *TARA* Oceans kernels computed using the STATIS-UMKL approach.

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 S5: STATIS-UMKL allows to have an overview on the different datasets and should be used when the integrated analysis focuses on correlated informations.

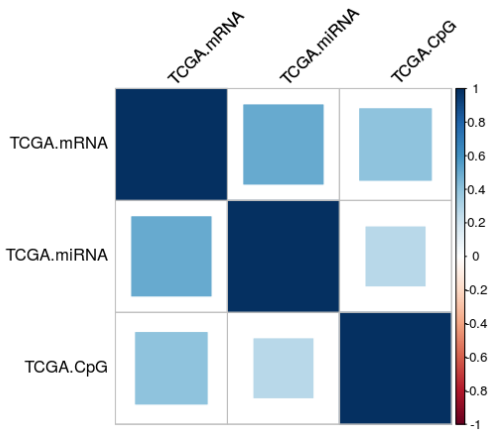
4.2 Similarities between breast cancer kernels

As performed in the previous section, similarities between breast cancer kernels are computed and presented in Supplementary Figure S11. Results show that the strongest correlations are obtained with **TCGA.mRNA**, which is expected as levels of mRNA expression is the main signature to classify breast tumours into subtypes.

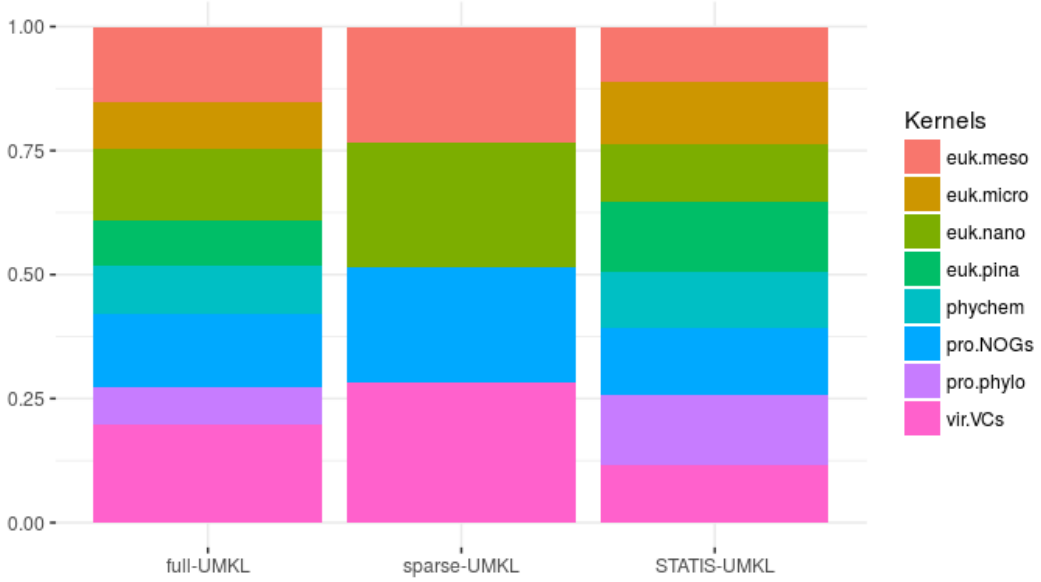
5 Comparison of the different integration options

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 S12.

First, note that, Supplementary Figure S12 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.



Supplementary Figure S11: Similarities between breast cancer kernels computed using the STATIS-UMKL approach.



Supplementary Figure S12: Kernels weights obtained for the three proposed approaches: full-UMKL, sparse-UMKL and STATIS-UMKL. Colours represent the different kernels.

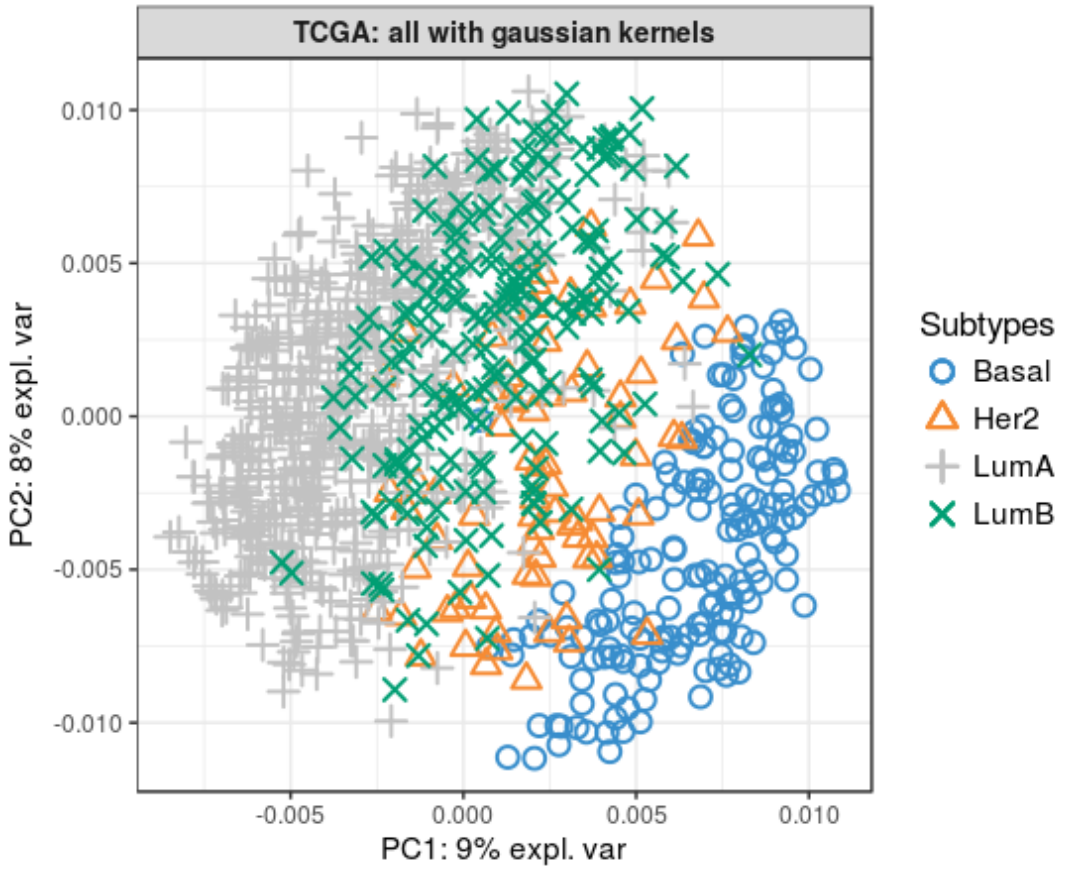
Results show that the three proposed methods are complementary and can be used depending on the research question and the analysis step. The STATIS-UMKL approach allows to have an overview on the correlation between the different datasets to analyse and to integrate them in a consensual way. sparse-UMKL 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.

6 KPCA analysis of TCGA datasets

In addition to KSOM, the meta-kernel obtained from the three TCGA datasets is used as an input for KPCA. Supplementary Figures S13 and 14 provide, respectively, the projection of the samples on the first two axes of the KPCA and the 10 variables found as the most important to explain this axis. The organization of the different samples on the first two axes of the KPCA (and especially on the

first one) is found to be consistent with the cancer subtype classification and typology, as described by the result of KSOM. Additional analyses (not shown for the sake of paper length), performing a KPCA on each single omic datasets, show that our multi-omics approach improves tumours subtype discrimination on the first KPCA axis.

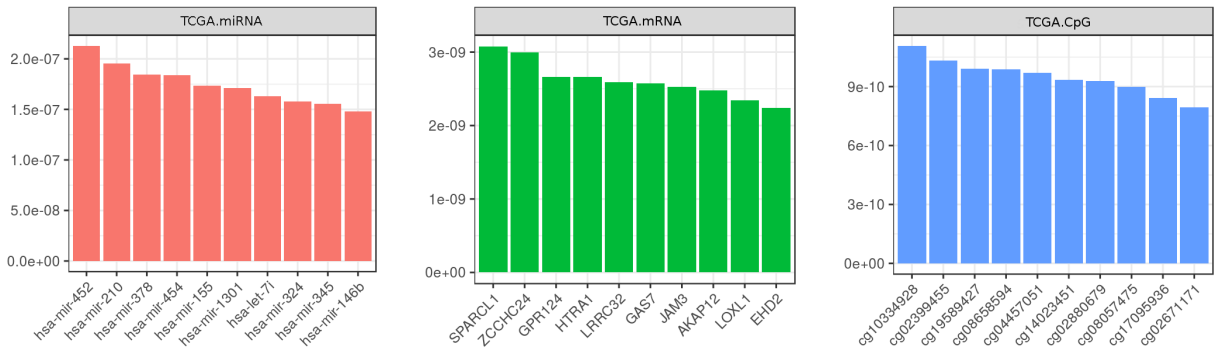
Variables found as important to explain the first axis are potential biomarkers that can discriminate between the cancer subtypes. Their expression profiles are presented in Supplementary Figure S15. First, note that the samples organization observed on the first axis of the KPCA is consistent with the 10 most important mRNAs expression since a gradient can be observed from samples of subtype *LumA* to samples of subtype *Basal* samples. Second, selected variables for the **TCGA.CpG** kernel reveal a specific pattern for *Basal* samples, this suggests that the identified CpGs may also be useful to identify this breast cancer subtype.



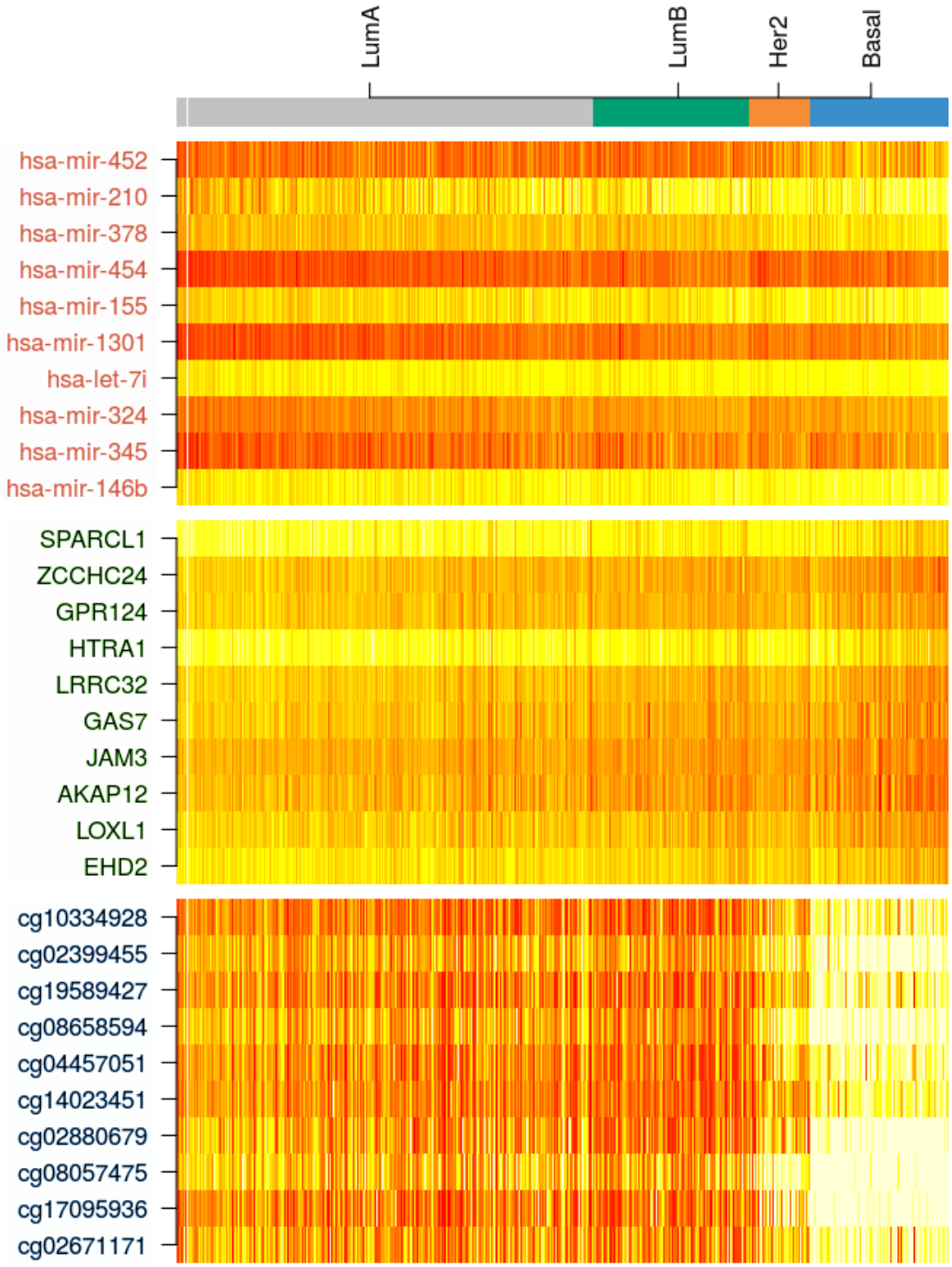
Supplementary Figure S13: Projection of the samples on the first two KPCA axes. Colours represent the cancer subtype.

In addition, the 10 mRNA selected as the most important were also submitted to an GO analysis (performed with **biomaRt** and **topGO**) and compared to the reference list of the 2,000 mRNA originally included in the data. Four biological processes were found to be significantly enriched in this list (at risk 10% with a Fisher test): extracellular matrix organization, extracellular structure organization, anatomical structure development and regulation of immune system process, that are all known as biological processes involved in cancer development.

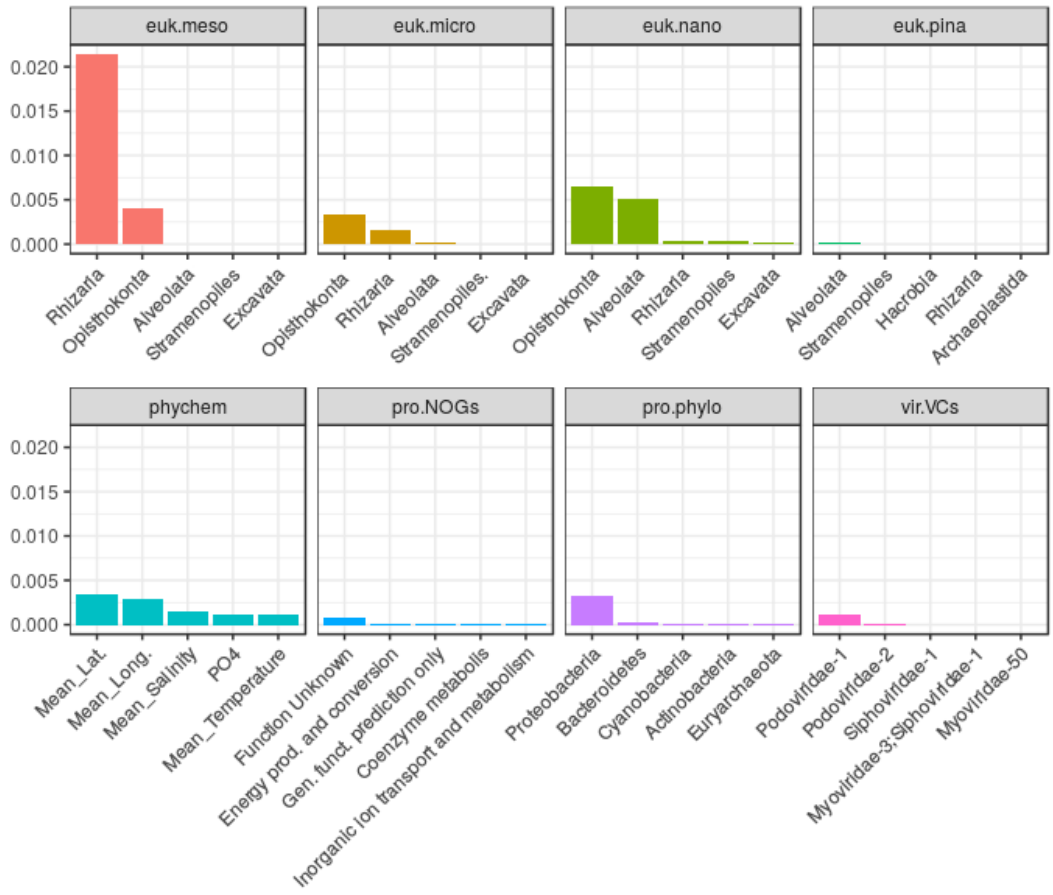
7 Supplementary figures



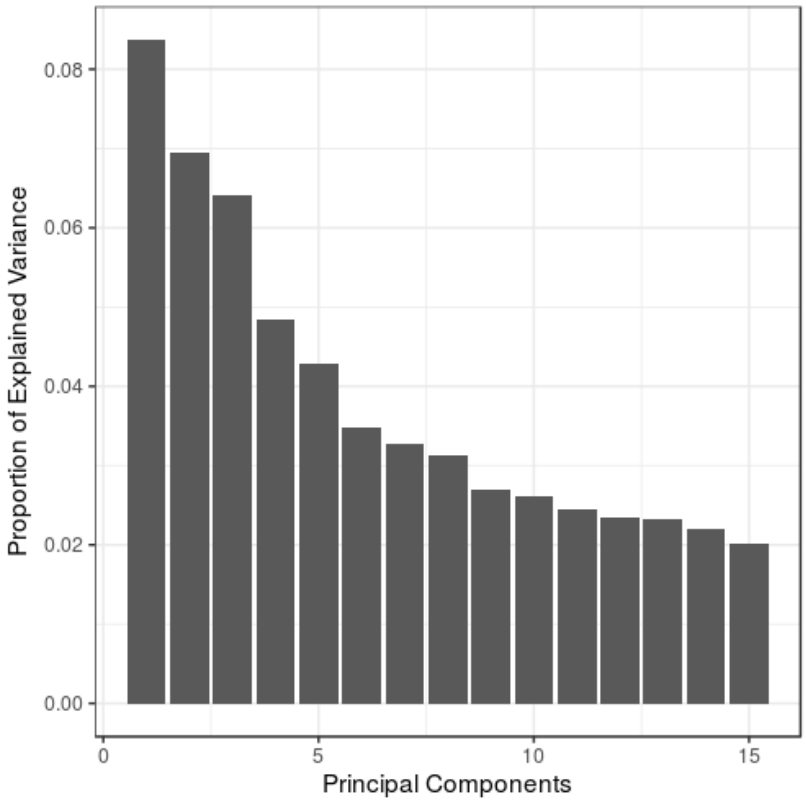
Supplementary Figure S14: The 10 most important variables for the first axis of the KPCA and for each of the 3 datasets, ranked by decreasing Crone-Crosby distance.



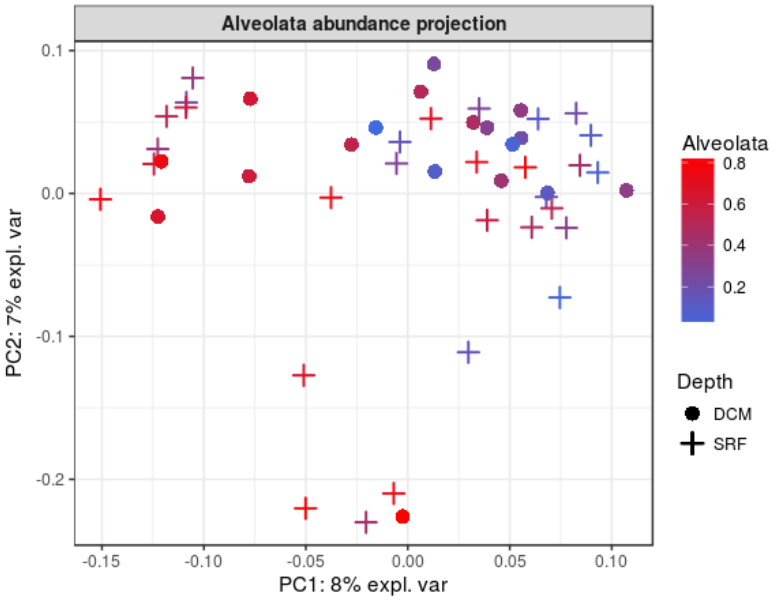
Supplementary Figure S15: Expression profiles of the 10 most important variables for each datasets. Samples were ranked by cancer subtypes and from the smallest to the highest coordinates on the first axis of the KPCA. Colours represent the expression levels from high (white) to low values (red).



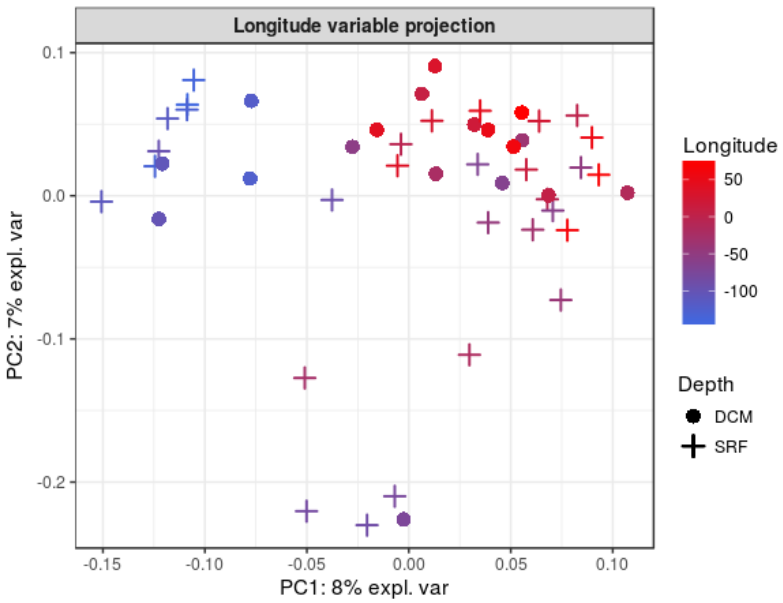
Supplementary Figure S16: The 5 most important variables for the second axis of the KPCA and for each of the 8 datasets, ranked by decreasing Crone-Crosby distance.



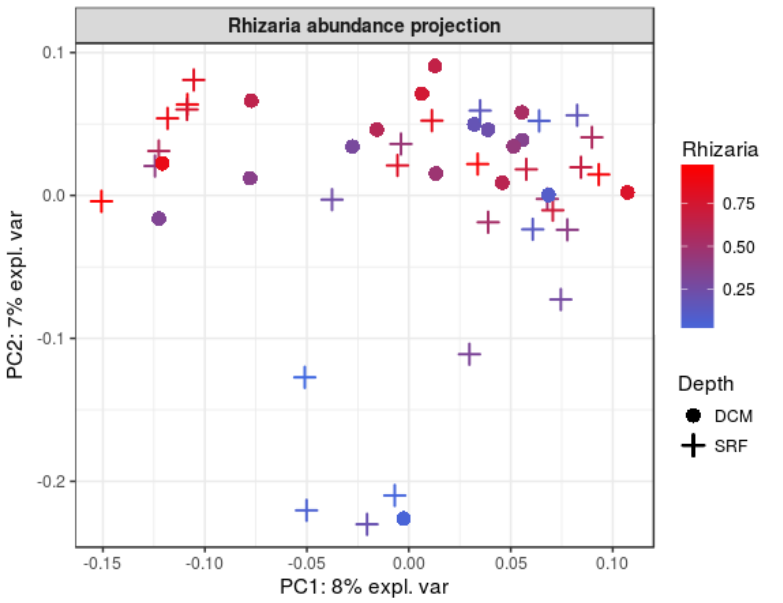
Supplementary Figure S17: Entropy preserved by the 15 first axes of the KPCA performed on the meta-kernel obtained using the full-UMKL approach and environmental, prokaryotic, eukaryotic and viral datasets.



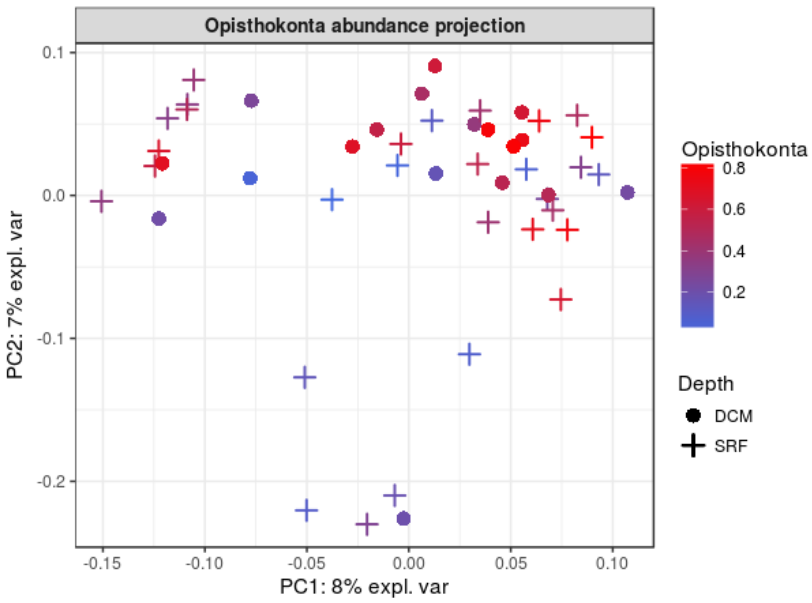
Supplementary Figure S18: Projection of the observations on the first two KPCA axes. Colours represent the relative abundance of *alveolata* organisms in the nanoplanktonic community: blue for low values and red for high values.



Supplementary Figure S19: Projection of the observations on the first two KPCA axes. Colours represent the longitude: blue for low values and red for high values.



Supplementary Figure S20: Projection of the observations on the first two KPCA axes. Colours represent the relative abundance of *rhizaria* organisms in the mesoplanktonic community: blue for low values and red for high values.



Supplementary Figure S21: Projection of the observations on the first two KPCA axes. Colours 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.

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

[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).

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

[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.

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

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

[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).