Integrating *Tara* Oceans data sets using Multiple Kernel Learning

Jérôme Mariette* and Nathalie Villa-Vialaneix

NETBIO 2017

MIA TOULOUSE

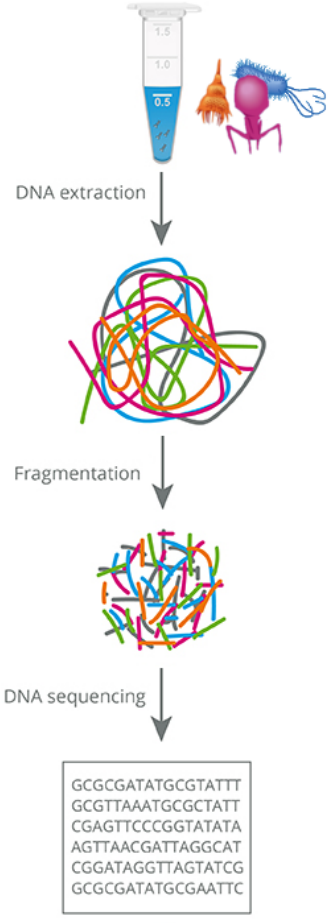
INRA SCIENCE & IMPACT

genotoul bioinfo

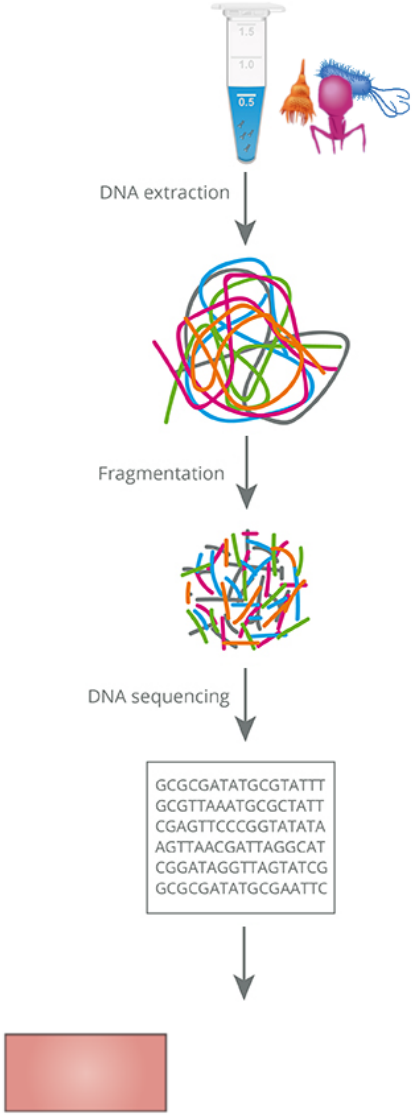
Contents

Contents

What are metagenomic data?



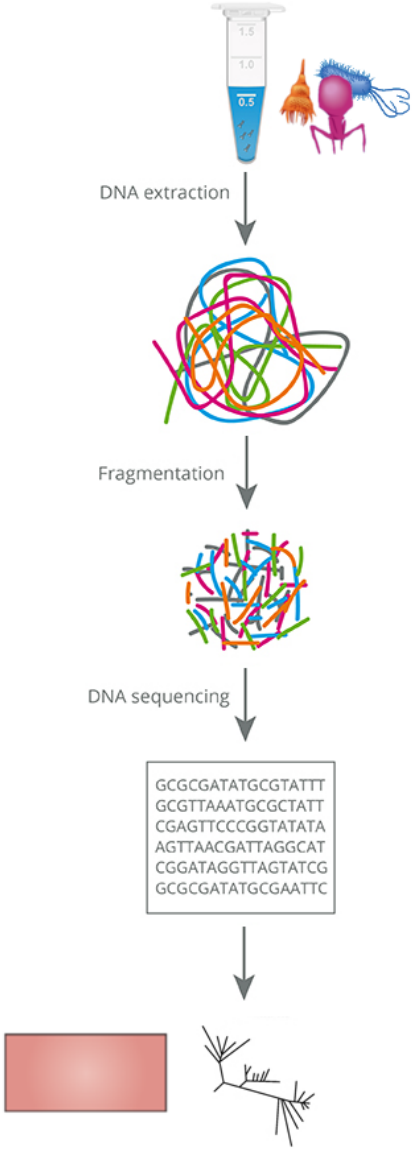
What are metagenomic data?



Heterogeneous types

- Abundance data: sparse $n \times p$-matrices with count data of samples in rows and descriptors (species, OTUs, KEGG groups, k-mer, ...) in columns.

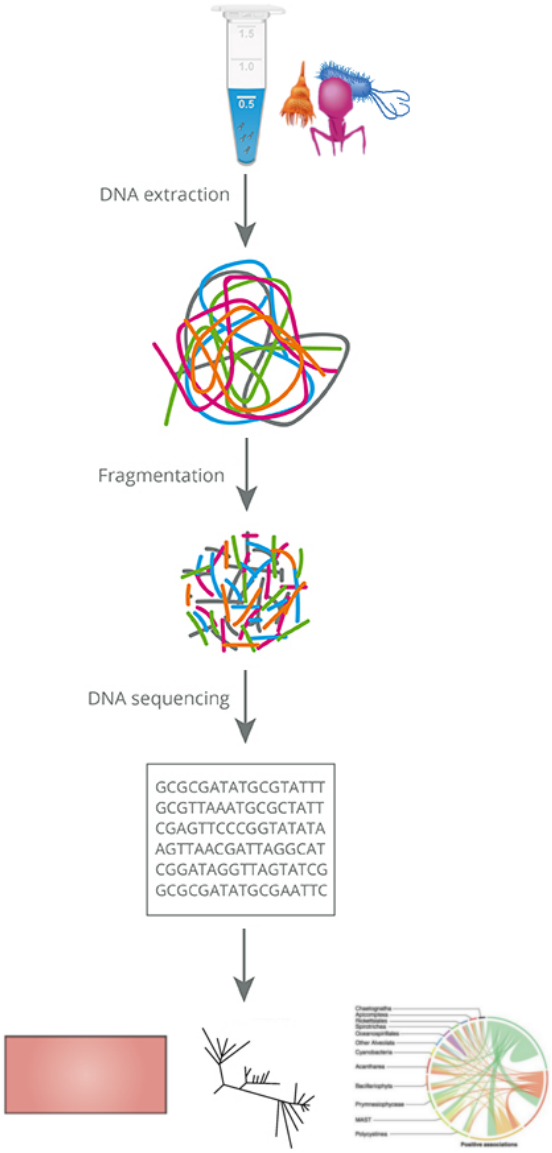
What are metagenomic data?



Heterogeneous types

- Abundance data: sparse $n \times p$-matrices with count data of samples in rows and descriptors (species, OTUs, KEGG groups, k-mer, ...) in columns.
- Phylogenetic tree: one tree with p leaves built from the sequences collected in the n samples.

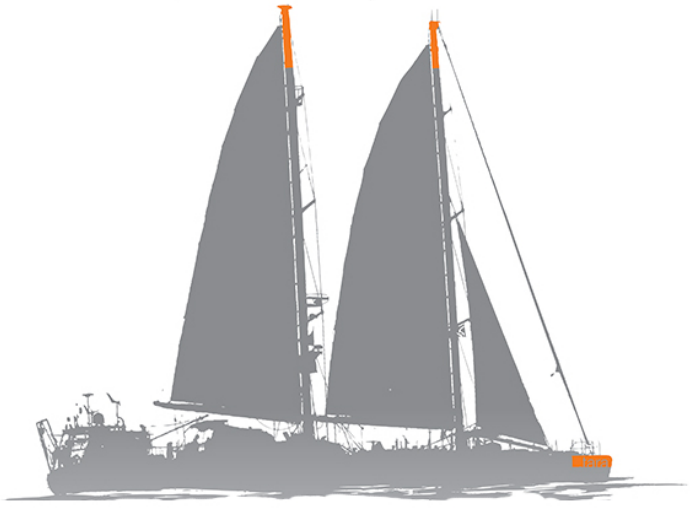
What are metagenomic data?



Heterogeneous types

- Abundance data: sparse $n \times p$-matrices with count data of samples in rows and descriptors (species, OTUs, KEGG groups, k-mer, ...) in columns.
- Phylogenetic tree: one tree with p leaves built from the sequences collected in the n samples.
- Co-occurence graph: a p nodes graph.
- ...

TARA Oceans expedition



TARA OCEANS

EXPEDITIONS
www.taraexpeditions.org

The 2009-2013 expedition

- Co-directed by Étienne Bourgois and Éric Karsenti.
- 7,012 datasets collected from 35,000 samples of plankton and water (11,535 Gb of data).
- Study of plankton: bacteria, protists, metazoans and viruses representing more than 90% of the biomass in the ocean.

TARA Oceans expedition

Science (May 2015) - Studies on:

- eukaryotic plankton diversity [de Vargas et al., 2015],
- ocean viral communities [Brum et al., 2015],
- global plankton interactome [Lima-Mendez et al., 2015],
- global ocean microbiome [Sunagawa et al., 2015],
- ....

→ datasets from different types and different sources analyzed separately.

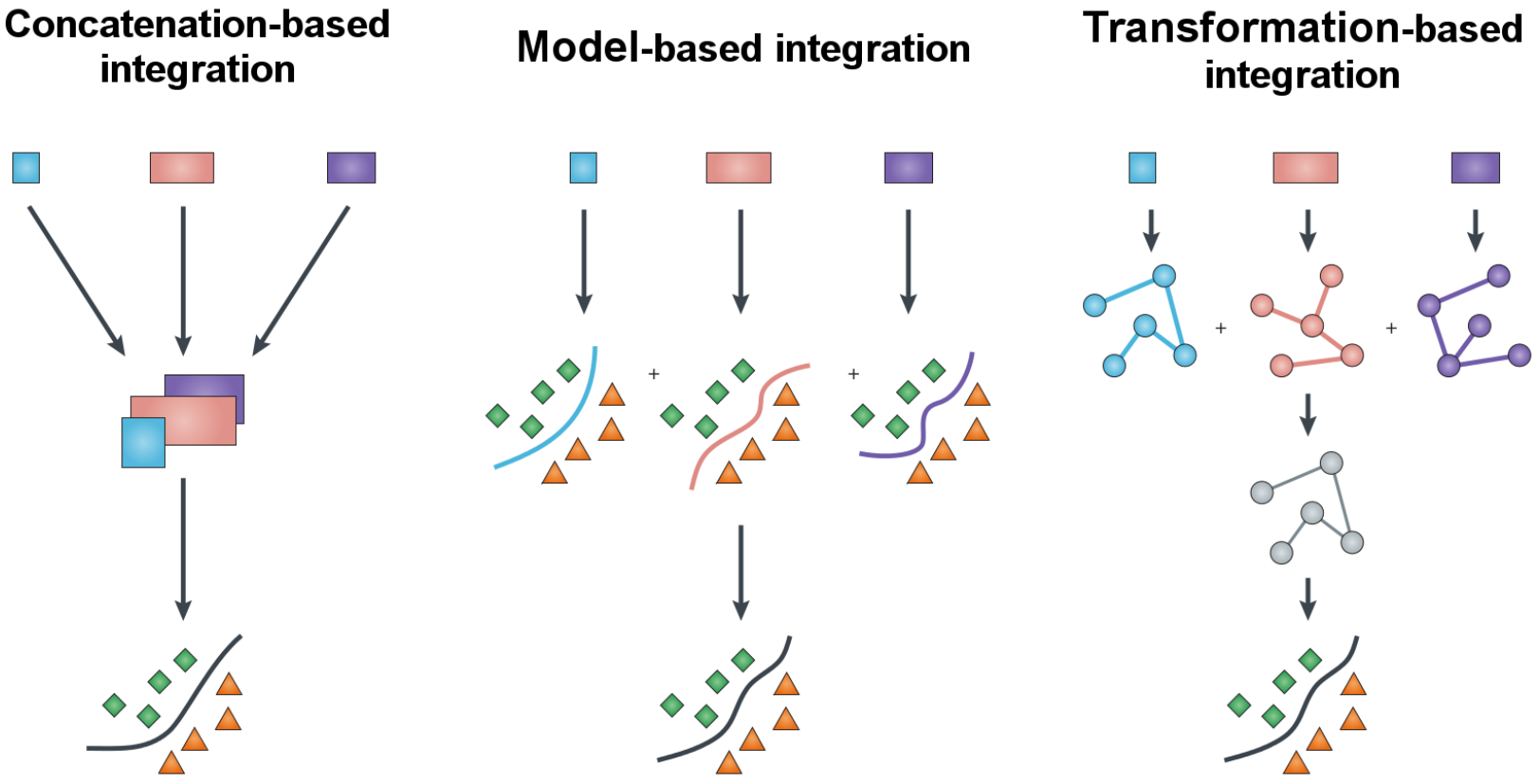
Background of this talk

Objectives

- Until now: many papers using many methods. No integrated analysis performed.
- What do the datasets reveal if integrated in a single analysis?
- Our purpose: develop a generic method to integrate phylogenetic, taxonomic and functional community composition to environmental factors.

Contents

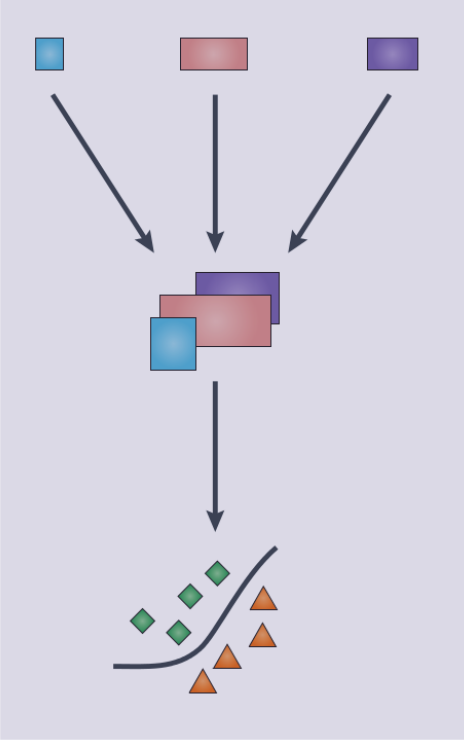
Integrating 'omics data



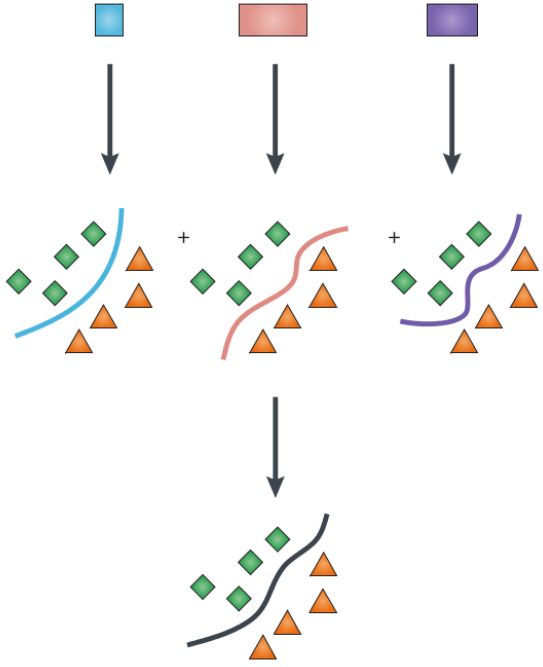
[Ritchie et al., 2015]

Integrating ’omics data

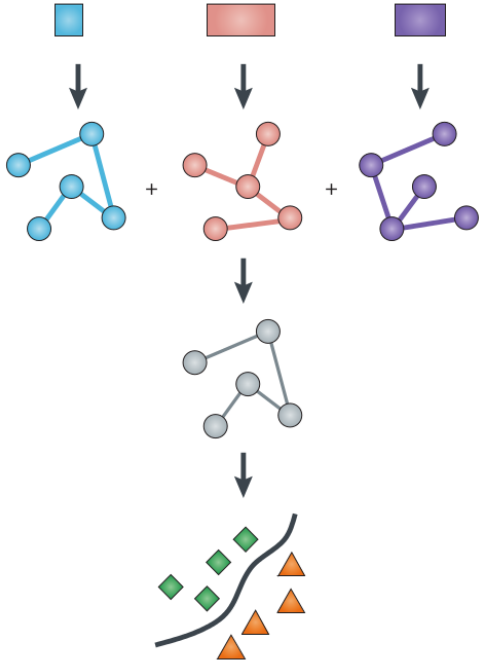
Concatenation-based integration



Model-based integration



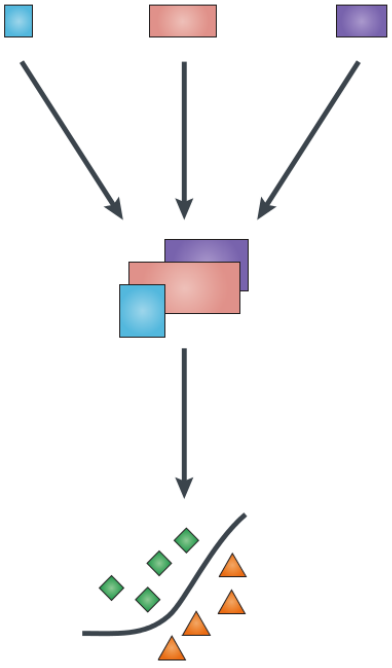
Transformation-based integration



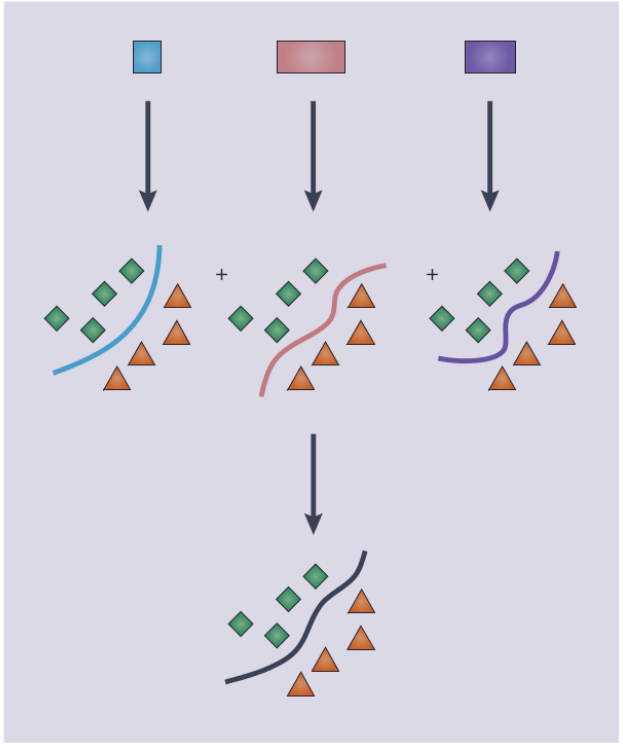
[Ritchie et al., 2015]

Integrating ’omics data

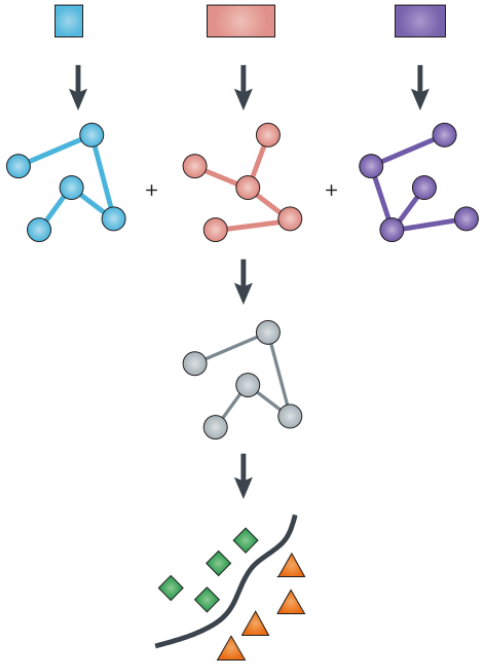
Concatenation-based integration



Model-based integration



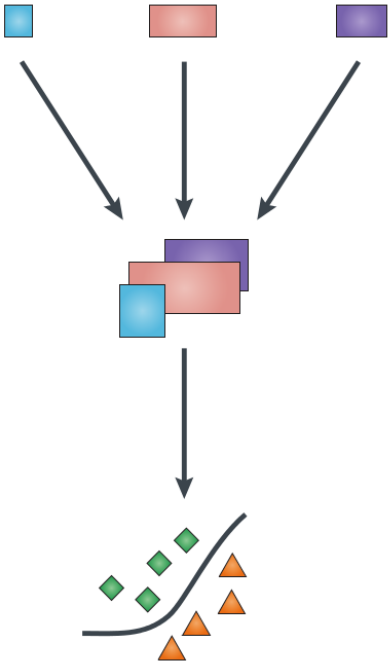
Transformation-based integration



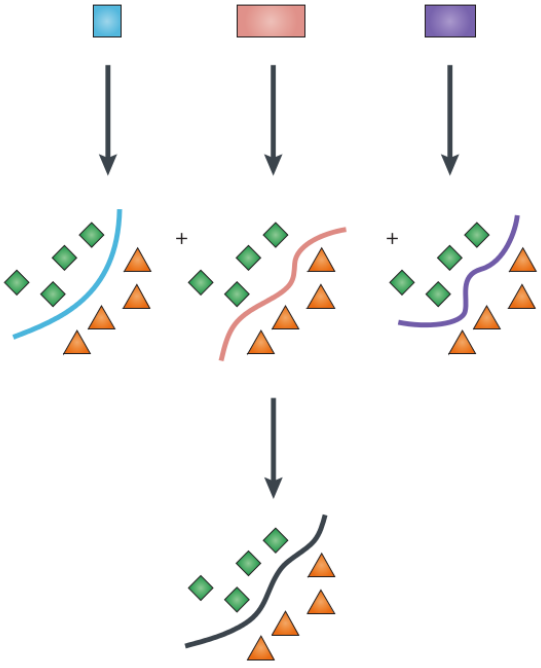
[Ritchie et al., 2015]

Integrating 'omics data

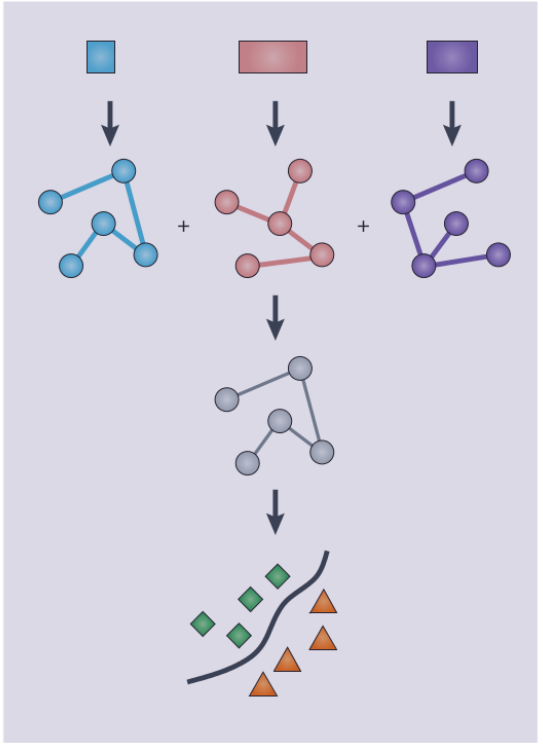
Concatenation-based integration



Model-based integration

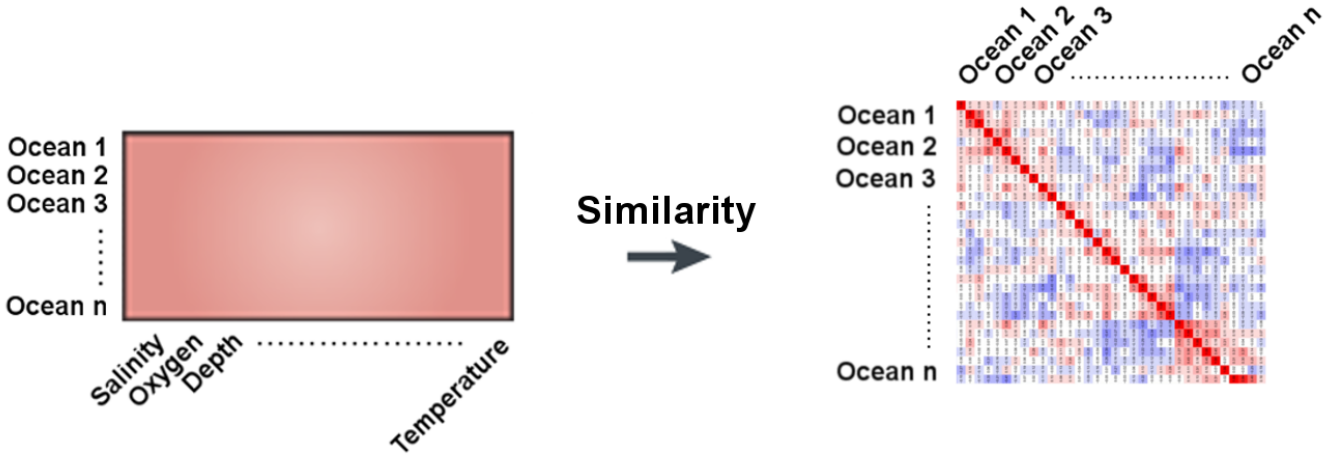


Transformation-based integration

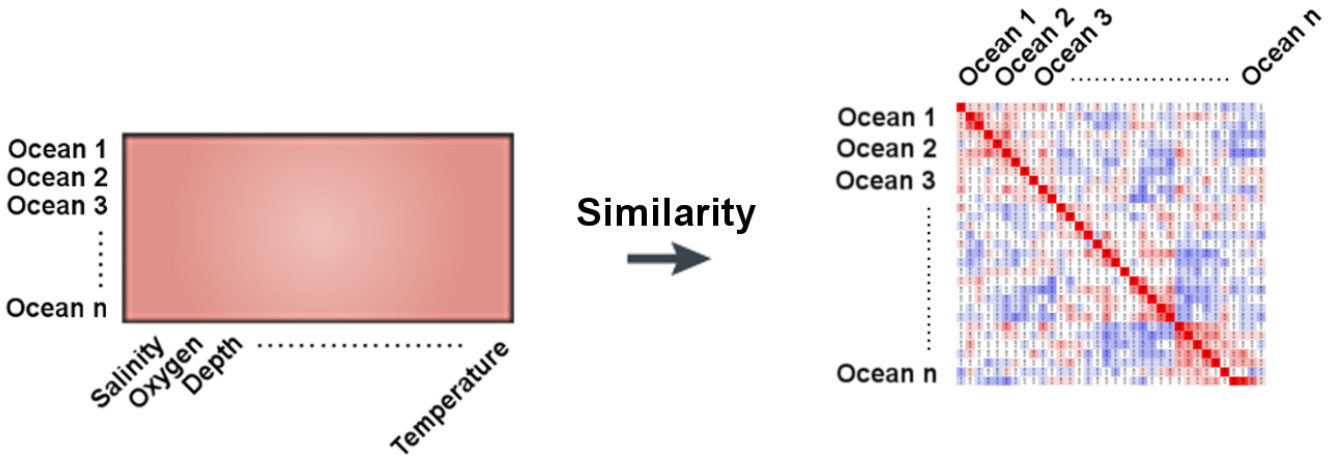


[Ritchie et al., 2015]

Integrating ’omics data using kernels



Integrating 'omics data using kernels



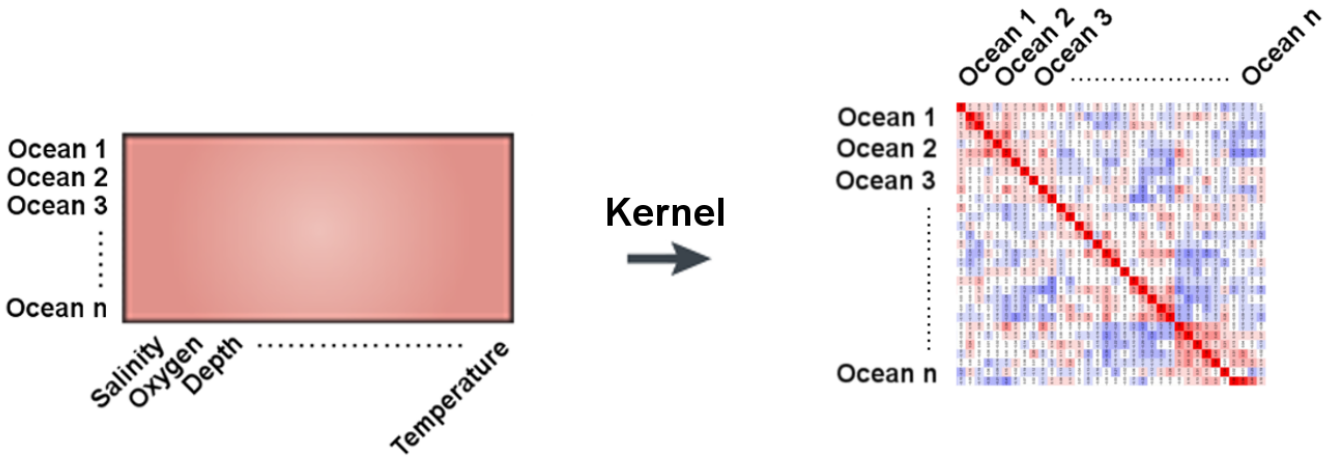
Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st: $K(x_i, x_j) = K(x_j, x_i)$ and $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$
In this case:

$$\exists (\mathcal{H}, \langle ., . \rangle), \phi : \mathcal{G} \to \mathcal{H} \text{ st: } K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

K can be viewed as a dot product for $\mathcal{G}$.

Integrating 'omics data using kernels



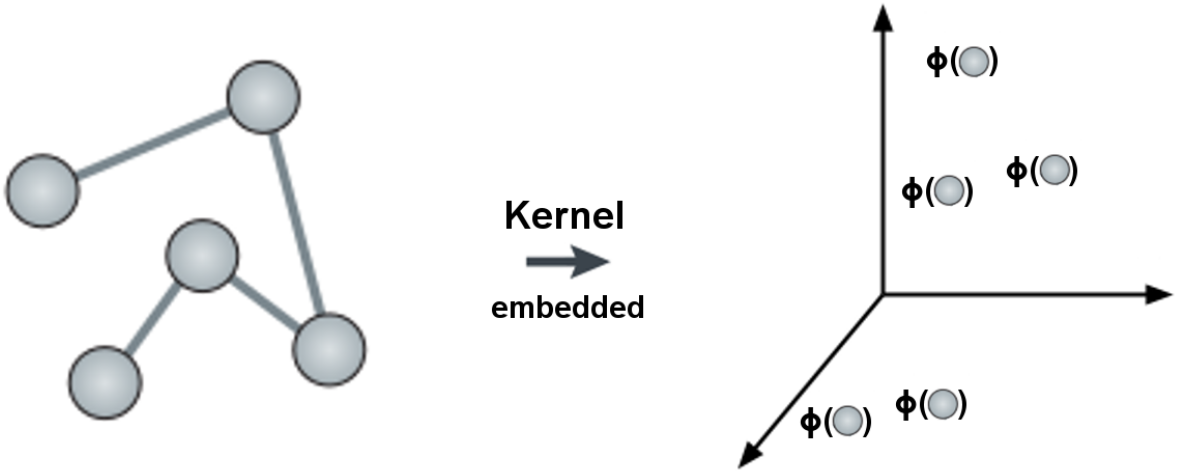
Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st: $K(x_i, x_j) = K(x_j, x_i)$ and $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$
In this case:

$$\exists (\mathcal{H}, \langle ., . \rangle), \phi : \mathcal{G} \to \mathcal{H} \text{ st: } K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

K can be viewed as a dot product for $\mathcal{G}$.

Integrating 'omics data using kernels



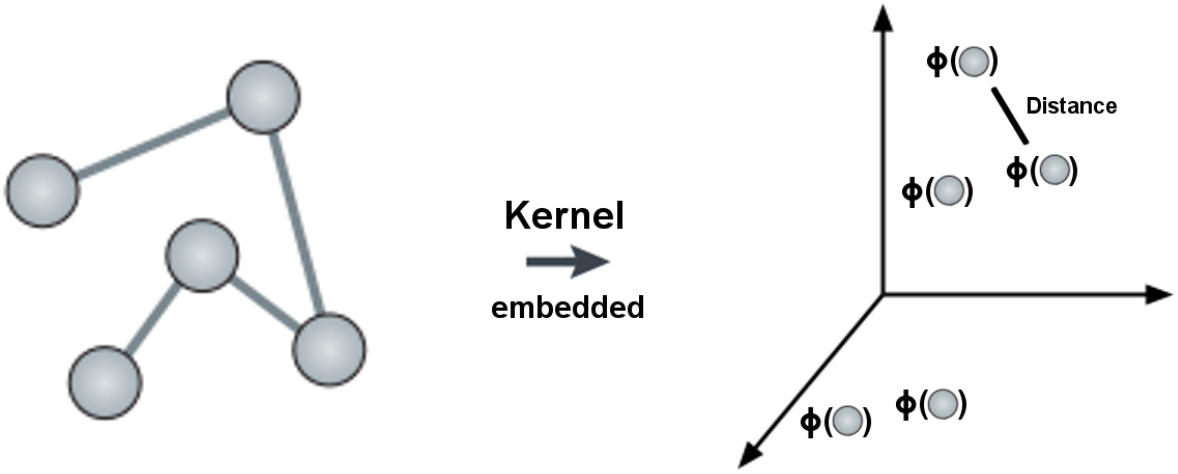
Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ st: $K(x_i, x_j) = K(x_j, x_i)$ and $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$
In this case:

$$\exists (\mathcal{H}, \langle ., . \rangle), \phi : \mathcal{G} \to \mathcal{H} \text{ st: } K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

K can be viewed as a dot product for $\mathcal{G}$.

Integrating 'omics data using kernels



Desired mathematical properties for the similarity

Function $K : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}$ st: $K(x_i, x_j) = K(x_j, x_i)$ and $\forall m \in \mathbb{N}$, $\forall x_1, ..., x_m \in \mathcal{G}$, $\forall \alpha_1, ..., \alpha_m \in \mathbb{R}$, $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$
In this case:

$$\exists (\mathcal{H}, \langle ., . \rangle), \phi : \mathcal{G} \rightarrow \mathcal{H} \text{ st: } K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

K can be viewed as a dot product for $\mathcal{G}$.

From multiple kernels to an integrated kernel

How to combine M kernels?

- naive approach: $K^* = \frac{1}{M} \sum_m^M K^m$

From multiple kernels to an integrated kernel

How to combine *M* kernels?

- naive approach: $K^* = \frac{1}{M} \sum_m^M K^m$
- supervised framework: $K^* = \sum_m^M \beta_m K^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$ with β_m chosen so as to minimize the prediction error [Gönen and Alpaydin, 2011]

From multiple kernels to an integrated kernel

How to combine M kernels?

- naive approach: $K^* = \frac{1}{M} \sum_m^M K^m$
- supervised framework: $K^* = \sum_m^M \beta_m K^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$ with β_m chosen so as to minimize the prediction error [Gönen and Alpaydin, 2011]
- unsupervised framework but input space is $\mathbb{R}^d$ [Zhuang et al., 2011] $K^* = \sum_m^M \beta_m K^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$ with β_m chosen so as to
  - minimize the distortion between all training data $\sum_{ij} K^*(x_i, x_j) \|x_i - x_j\|^2$;
  - AND minimize the approximation of the original data by the kernel embedding $\sum_i \left\| x_i - \sum_j K^*(x_i, x_j) x_j \right\|^2$.

:

From multiple kernels to an integrated kernel

Our proposal

- 2 UMKL frameworks which do not require data to have values in $\mathbb{R}^d$.
  - maximizing the average similarity between kernels (STATIS)
  - minimizing the distorsion with the topology of the data

From multiple kernels to an integrated kernel

Maximizing the average similarity between kernels (STATIS)

- STATIS: exploratory tools for multi-block datasets.

From multiple kernels to an integrated kernel

Maximizing the average similarity between kernels (STATIS)

- STATIS: exploratory tools for multi-block datasets.
- first step: compute a similarity matrix between kernels.

$$C_{mm'} = \frac{\langle K^m, K^{m'} \rangle_F}{\|K^m\|_F \|K^{m'}\|_F}.$$

From multiple kernels to an integrated kernel

Maximizing the average similarity between kernels (STATIS)

- STATIS: exploratory tools for multi-block datasets.
- first step: compute a similarity matrix between kernels.

$$C_{mm'} = \frac{\langle K^m, K^{m'} \rangle_F}{\|K^m\|_F \|K^{m'}\|_F}.$$

- second step:

$$\arg\max_{\beta} \sum_{m=1}^{M} \left\langle K^*(\beta), \frac{K^m}{\|K^m\|} \right\rangle = \arg\max_{\beta} \sum_{m,m'=1}^{M} \beta_m \beta_{m'} C_{mm'}.$$

From multiple kernels to an integrated kernel

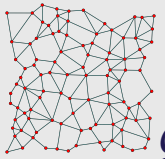
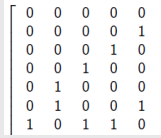
A kernel preserving the original topology of the data

- From an idea similar to that of [Lin et al., 2010], find a kernel such that the local geometry of the data in the feature space is similar to that of the original data.

From multiple kernels to an integrated kernel

A kernel preserving the original topology of the data

- From an idea similar to that of [Lin et al., 2010], find a kernel such that the local geometry of the data in the feature space is similar to that of the original data.

$K^m \longrightarrow$  $\mathcal{G}_k^m \longrightarrow$  A_k^m

$k-$nearest neighbors graph

adjacency matrix

$\Rightarrow W = \sum_m \mathbb{I}_{\{A_k^m > 0\}}$ or $W = \sum_m A_k^m$

From multiple kernels to an integrated kernel

A kernel preserving the original topology of the data

- From an idea similar to that of [Lin et al., 2010], find a kernel such that the local geometry of the data in the feature space is similar to that of the original data.

$$K^m \longrightarrow \underbrace{\mathcal{G}_k^m}_{k-\text{nearest neighbors graph}} \longrightarrow \underbrace{A_k^m}_{\text{adjacency matrix}}$$

$$\Rightarrow W = \sum_m \mathbb{I}_{\{A_k^m > 0\}} \text{ or } W = \sum_m A_k^m$$

-

$$\arg\min_{\beta} \sum_{i,j=1}^{n} W_{ij} \left\| \sum_{m=1}^{M} \beta_m \left(\begin{bmatrix} K_{i1}^m \\ \vdots \\ K_{in}^m \end{bmatrix} - \begin{bmatrix} K_{j1}^m \\ \vdots \\ K_{jn}^m \end{bmatrix} \right) \right\|^2 .$$

From multiple kernels to an integrated kernel

A kernel preserving the original topology of the data

- Sparse version with $\|\beta\|_1 = \sum_m \beta_m = 1 \Rightarrow$ standard QP problem with linear constrains (ex: package **quadprog** in R).

From multiple kernels to an integrated kernel

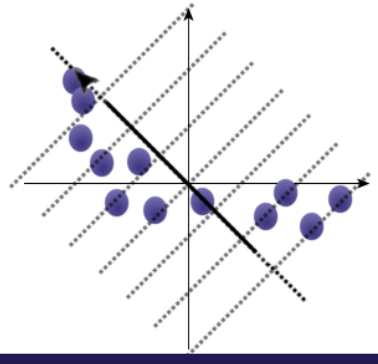
A kernel preserving the original topology of the data

- Sparse version with $\|\beta\|_1 = \sum_m \beta_m = 1 \Rightarrow$ standard QP problem with linear constrains (ex: package **quadprog** in R).
- Non sparse version with $\|\beta\|_2 = 1 \Rightarrow$ QPQC problem (hard to solve). Solved using Alternating Direction Method of Multipliers (ADMM [Boyd et al., 2011]).

Contents

K-PCA [Schölkopf et al., 1998]

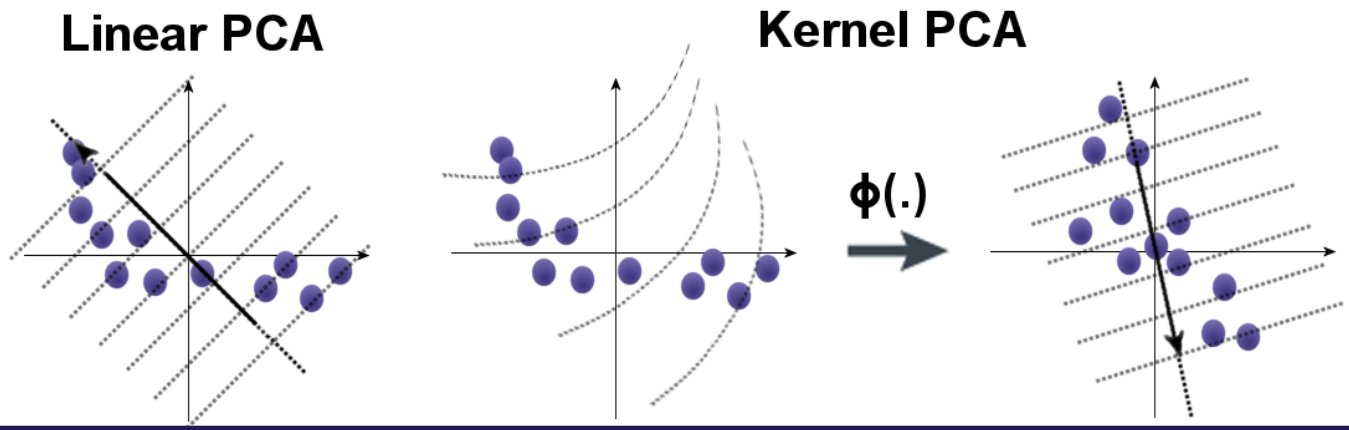
Linear PCA



Standard Principal Component Analysis (PCA)

- Projection of high dimensional dataset in a small dimensional space
- Designed so as to keep most of the data variability
- Axes interpretable from a variable and from an observation point of view (axes are linear combinations of the original variables)

K-PCA [Schölkopf et al., 1998]



Standard Principal Component Analysis (PCA)

- Projection of high dimensional dataset in a small dimensional space
- Designed so as to keep most of the data variability
- Axes interpretable from a variable and from an observation point of view (axes are linear combinations of the original variables)

Kernel Principal Component Analysis (K-PCA)

- PCA in the feature space (corresponds to a non linear projection of the original data in the original space)
- No representation for the variables

K-PCA [Schölkopf et al., 1998]

How to interprete the axes ?

- few attempts in the literature to help understand the relations of KPCA with the original measures.
- [Reverter et al., 2014] add a representation of the variables to the plot: visualizing their influence over the results from derivative computations (datasets take values in $\mathbb{R}^d$).

K-PCA [Schölkopf et al., 1998]

How to interprete the axes ?

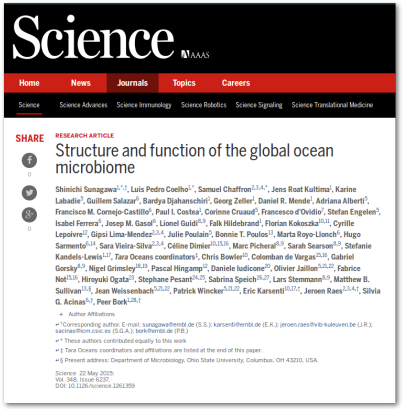
- few attempts in the literature to help understand the relations of KPCA with the original measures.
- [Reverter et al., 2014] add a representation of the variables to the plot: visualizing their influence over the results from derivative computations (datasets take values in $\mathbb{R}^d$).

Our proposal

- generic approach that assesses the influence of variables.
- randomize a dataset variable and build a new kernel $\tilde{K^*}$.
- compute the Crone and Crosby distance [Crone and Crosby, 1995] between the K^* and $\tilde{K^*}$ K-PCA sub-spaces.

Contents

TARA Oceans datasets



Science

Home | News | Journals | Topics | Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

RESEARCH ARTICLE

Structure and function of the global ocean microbiome

Shinichi Sunagawa[1,*,†], Luis Pedro Coelho[1,*], Samuel Chaffron[2,3,4,*], Jens Roat Kultima[1], Karine Labadie[5], Guillem Salazar[6], Bardya Djahanschiri[1], Georg Zeller[1], Daniel R. Mende[1], Adriana Alberti[5], Francisco M. Cornejo-Castillo[6], Paul I. Costea[1], Corinne Cruaud[5], Francesco d'Ovidio[7], Stefan Engelen[5], Isabel Ferrera[6], Josep M. Gasol[6], Lionel Guidi[8,9], Falk Hildebrand[1], Florian Kokoszka[10,11], Cyrille Lepoivre[12], Gipsi Lima-Mendez[2,3,4], Julie Poulain[5], Bonnie T. Poulos[13], Marta Royo-Llonch[6], Hugo Sarmento[6,14], Sara Vieira-Silva[2,3,4], Céline Dimier[15,15,16], Marc Picheral[8,9], Sarah Searson[8,9], Stefanie Kandels-Lewis[1,17], *Tara* Oceans coordinators[1], Chris Bowler[15], Colomban de Vargas[15,18], Gabriel Gorsky[8,9], Nigel Grimsley[19,20], Pascal Hingamp[12], Daniele Iudicone[10], Olivier Jaillon[5,21,22], Fabrice Not[15,16], Hiroyuki Ogata[23], Stephane Pesant[24,25], Sabrina Speich[26,27], Lars Stemmann[8,9], Matthew B. Sullivan[13,], Jean Weissenbach[5,21,22], Patrick Wincker[5,21,22], Eric Karsenti[10,17,‡], Jeroen Raes[2,3,4,‡], Silvia G. Acinas[6,‡], Peer Bork[1,28,‡]

+ Author Affiliations

↵† Corresponding author. E-mail: sunagawa@embl.de (S.S.); karsenti@embl.de (E.K.); jeroen.raes@vib-kuleuven.be (J.R.); sacinas@icm.csic.es (S.G.A.); bork@embl.de (P.B.)

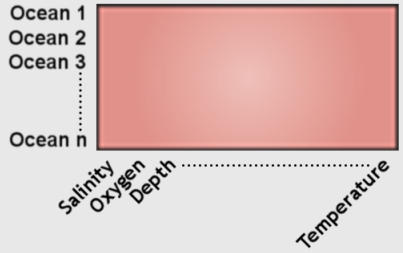
↵* These authors contributed equally to this work

↵‡ *Tara* Oceans coordinators and affiliations are listed at the end of this paper.

↵§ Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

Science 22 May 2015:
Vol. 348, Issue 6237.
DOI: 10.1126/science.1261359

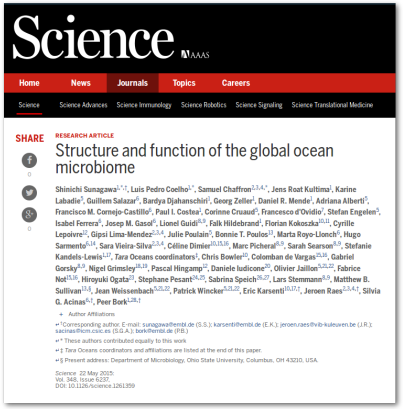
[Sunagawa et al., 2015]



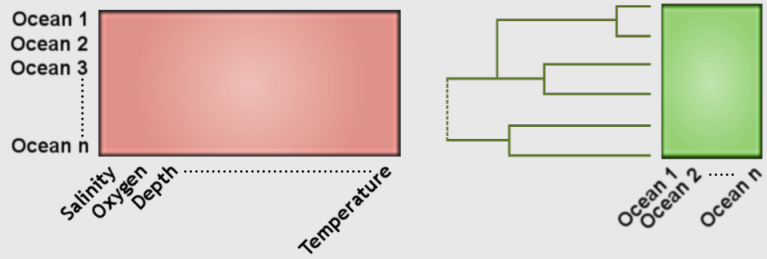
Datasets used

- environmental dataset: 22 numeric features (temperature, salinity, . . .).

TARA Oceans datasets



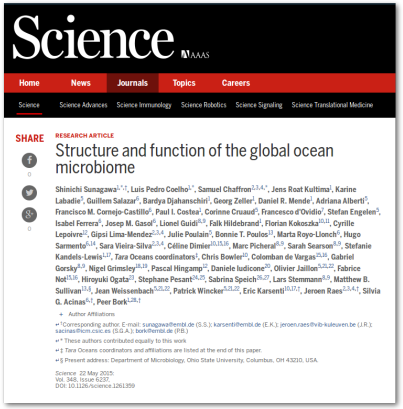
[Sunagawa et al., 2015]



Datasets used

- environmental dataset: 22 numeric features (temperature, salinity, . . .).
- bacteria phylogenomic tree: computed from $\sim$ 35,000 OTUs.

TARA Oceans datasets


Science

Home News Journals Topics Careers

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

SHARE

RESEARCH ARTICLE

Structure and function of the global ocean microbiome

Shinichi Sunagawa[1,*,†], Luis Pedro Coelho[1,*], Samuel Chaffron[2,3,4,*], Jens Roat Kultima[1], Karine Labadie[5], Guillem Salazar[6], Bardya Djahanschiri[1], Georg Zeller[1], Daniel R. Mende[1], Adriana Alberti[5], Francisco M. Cornejo-Castillo[6], Paul I. Costea[1], Corinne Cruaud[5], Francesco d'Ovidio[7], Stefan Engelen[5], Isabel Ferrera[6], Josep M. Gasol[6], Lionel Guidi[8,9], Falk Hildebrand[1], Florian Kokoszka[10,11], Cyrille Lepoivre[12], Gipsi Lima-Mendez[2,3,4], Julie Poulain[5], Bonnie T. Poulos[13], Marta Royo-Llonch[6], Hugo Sarmento[6,14], Sara Vieira-Silva[2,3,4], Céline Dimier[15,15,16], Marc Picheral[8,9], Sarah Searson[8,9], Stefanie Kandels-Lewis[1,17], *Tara* Oceans coordinators[1], Chris Bowler[15], Colomban de Vargas[15,18], Gabriel Gorsky[8,9], Nigel Grimsley[19,20], Pascal Hingamp[12], Daniele Iudicone[10], Olivier Jaillon[5,21,22], Fabrice Not[15,16], Hiroyuki Ogata[12], Stephane Pesant[23,24], Sabrina Speich[25,17], Lars Stemmann[8,9], Matthew B. Sullivan[13,3], Jean Weissenbach[5,21,22], Patrick Wincker[5,21,22], Eric Karsenti[10,17,*], Jeroen Raes[2,3,4,*], Silvia G. Acinas[6,*], Peer Bork[1,26,*]

+ Author Affiliations

*Corresponding author. E-mail: sunagawa@embl.de (S.S.); karsenti@embl.de (E.K.); jeroen.raes@vib-kuleuven.be (J.R.); sacinas@icm.csic.es (S.G.A.); bork@embl.de (P.B.)

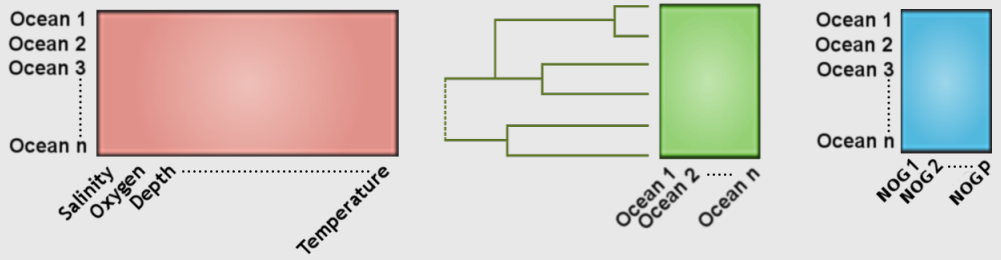
† These authors contributed equally to this work

‡ Tara Oceans coordinators and affiliations are listed at the end of this paper.

§ Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

Science 22 May 2015:
Vol. 348, Issue 6237.
DOI: 10.1126/science.1261359

[Sunagawa et al., 2015]

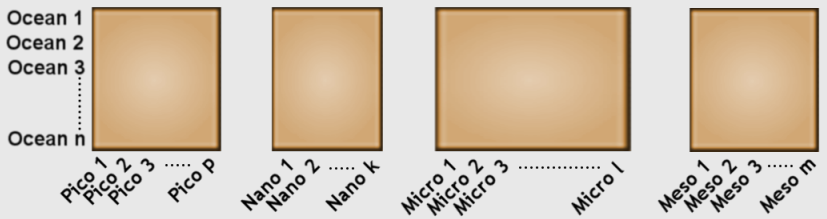


Datasets used

- environmental dataset: 22 numeric features (temperature, salinity, . . .).
- bacteria phylogenomic tree: computed from $\sim$ 35,000 OTUs.
- bacteria functional composition: $\sim$ 63,000 eggNOG gene families.

TARA Oceans datasets

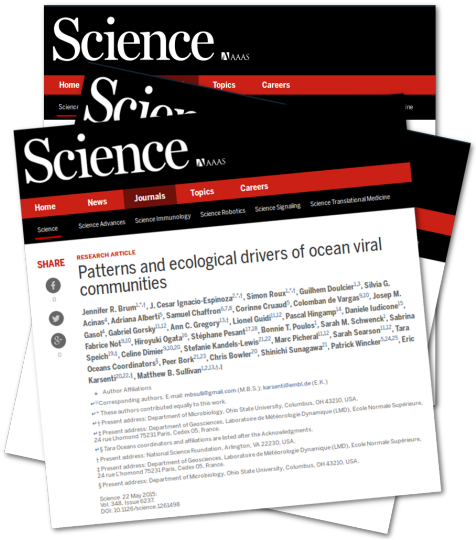
[de Vargas et al., 2015]



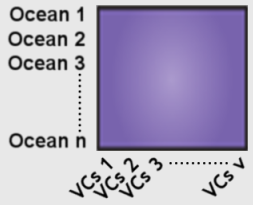
Datasets used

- environmental dataset: 22 numeric features (temperature, salinity, . . .).
- bacteria phylogenomic tree: computed from $\sim$ 35,000 OTUs.
- bacteria functional composition: $\sim$ 63,000 eggNOG gene families.
- eukaryotic plankton composition split into 4 groups: pico ($0.8 - 5\mu m$), nano ($5 - 20\mu m$), micro ($20 - 180\mu m$) and meso ($180 - 2000\mu m$).

TARA Oceans datasets



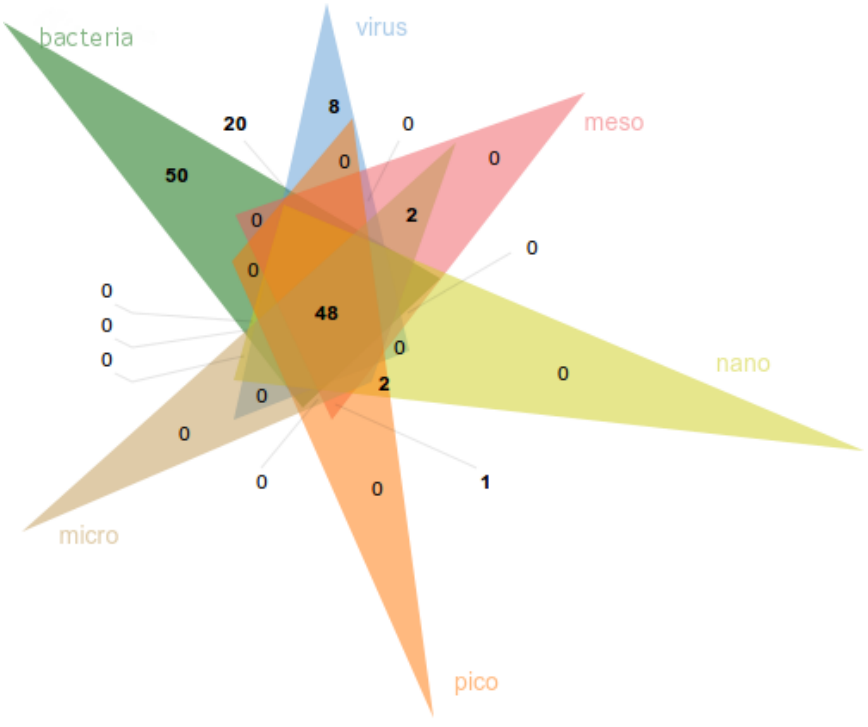
[Brum et al., 2015]



Datasets used

- environmental dataset: 22 numeric features (temperature, salinity, ...).
- bacteria phylogenomic tree: computed from $\sim$ 35,000 OTUs.
- bacteria functional composition: $\sim$ 63,000 eggNOG gene families.
- eukaryotic plankton composition split into 4 groups: pico ($0.8 - 5\mu m$), nano ($5 - 20\mu m$), micro ($20 - 180\mu m$) and meso ($180 - 2000\mu m$).
- virus composition: $\sim$ 867 virus clusters based on shared gene content.

TARA Oceans datasets



Common samples

- 48 samples,
- 2 depth layers: surface (SRF) and deep chlorophyll maximum (DCM),
- 31 different sampling stations.

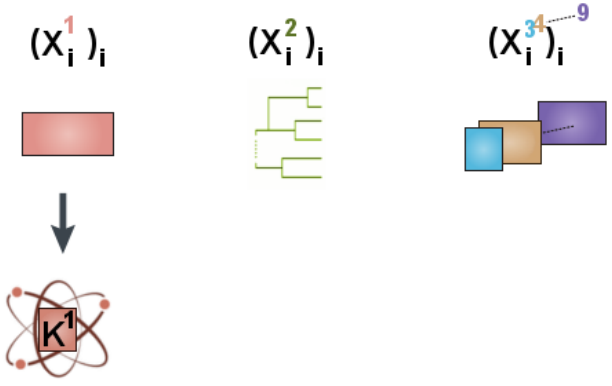
Integrating *TARA* Oceans datasets

M TARA Oceans datasets

$(x_i^m)_{i=1,\ldots,N,m=1,\ldots,M}$ measured on the same ocean samples $(1,\ldots,N)$ which take values in an arbitrary space $(\mathcal{X}^m)_m$:

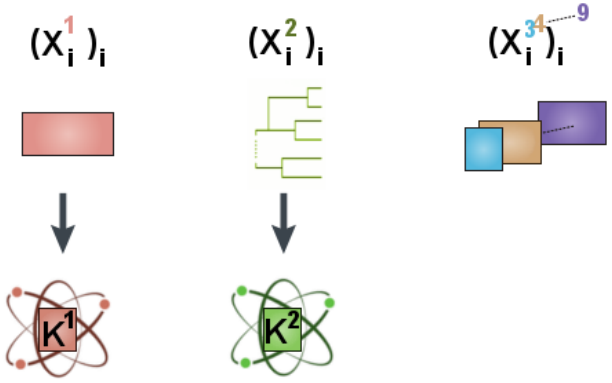
- phychem: environmental dataset,
- pro.phylo: prokaryote phylogenomic tree,
- pro.NOGs: prokaryote functional composition,
- euk.pina: eukaryote pico-nano-plankton composition,
- ...
- vir.VCs: virus composition.

Integrating *TARA* Oceans datasets



phychem (environmental dataset): standard euclidean distance, given by $K(x_i, x_j) = x_i^T x_j$.

Integrating *TARA* Oceans datasets

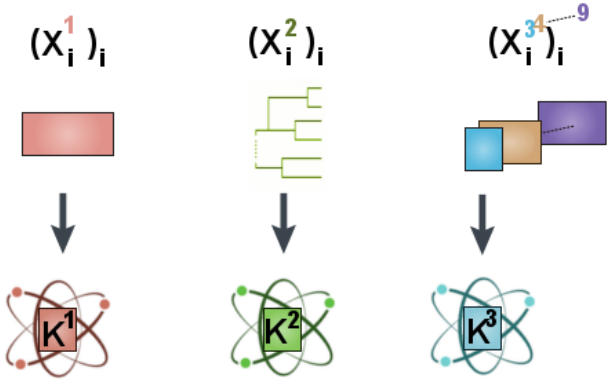


pro.phylo (prokarote phylogenomic tree): the weighted Unifrac distance, given by

$$d_{wUF}(A,B) = \frac{\sum_e l_e |p_e - q_e|}{\sum_e p_e + q_e},$$

l_e: length of branch e.
p_e: the fraction of community A below branch e.
q_e: the fraction of community B below branch e.

Integrating *TARA* Oceans datasets

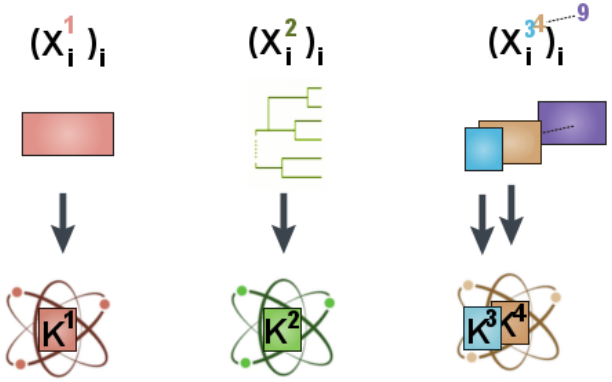


All composition based datasets: pro.NOGs (bacteria functional composition), eukaryote composition (euk.pina, euk.nano, euk.micro, euk.meso) and vir.VCs (virus composition) calculated using the Bray-Curtis dissimilarity,

$$d_{BC}(A,B) = \frac{\sum_g |n_g^A - n_g^B|}{\sum_g n_g^A + n_g^B},$$

n_g^A: gene g abundances summarized at eggNOG gene families level in community A.
n_g^B: same for community B.

Integrating *TARA* Oceans datasets

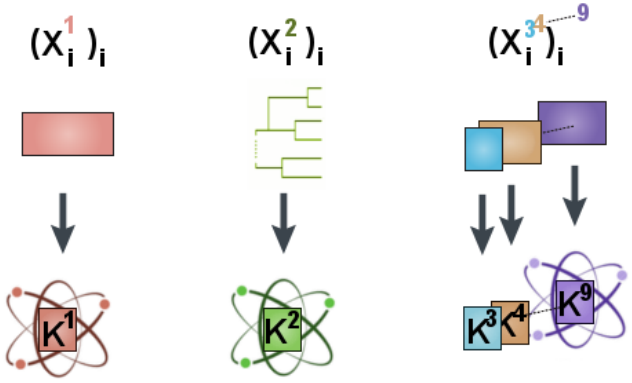


All composition based datasets: pro.NOGs (bacteria functional composition), eukaryote composition (euk.pina, euk.nano, euk.micro, euk.meso) and vir.VCs (virus composition) calculated using the Bray-Curtis dissimilarity,

$$d_{BC}(A,B) = \frac{\sum_g |n_g^A - n_g^B|}{\sum_g n_g^A + n_g^B},$$

n_g^A: gene g abundances summarized at eggNOG gene families level in community A.
n_g^B: same for community B.

Integrating *TARA* Oceans datasets

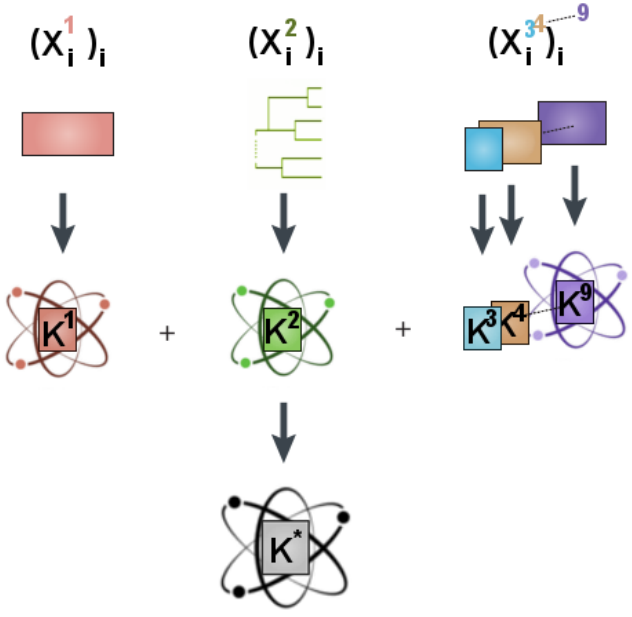


All composition based datasets: pro.NOGs (bacteria functional composition), eukaryote composition (euk.pina, euk.nano, euk.micro, euk.meso) and vir.VCs (virus composition) calculated using the Bray-Curtis dissimilarity,

$$d_{BC}(A,B) = \frac{\sum_g |n_g^A - n_g^B|}{\sum_g n_g^A + n_g^B},$$

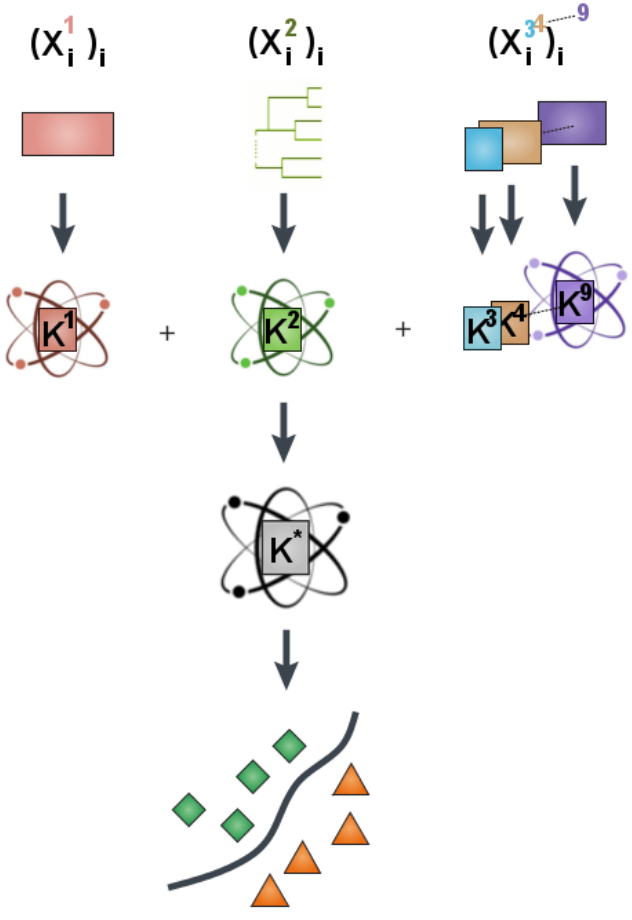
n_g^A: gene g abundances summarized at eggNOG gene families level in community A.
n_g^B: same for community B.

Integrating *TARA* Oceans datasets



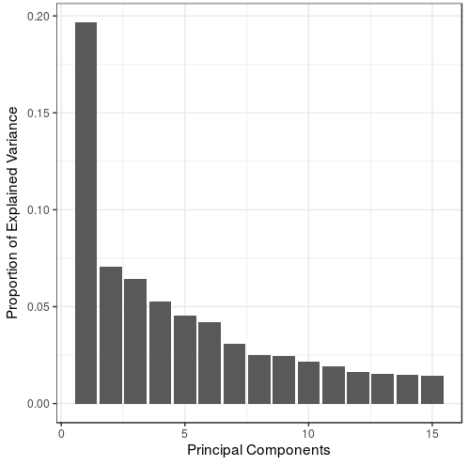
Combinaison of the M kernels to obtain K^*, a kernel preserving topology with L_2-norm constraint.

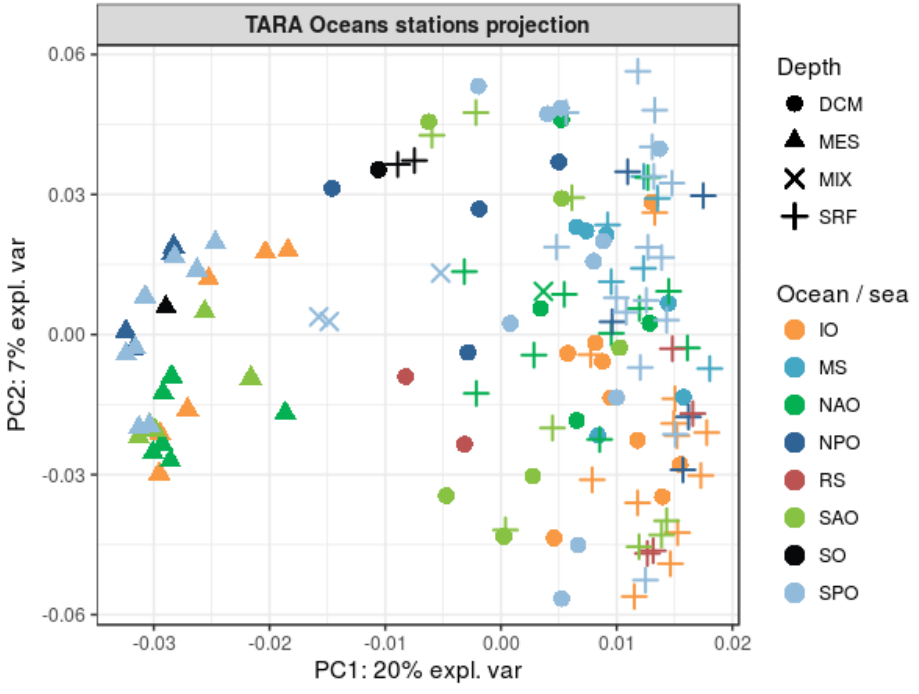
Integrating *TARA* Oceans datasets



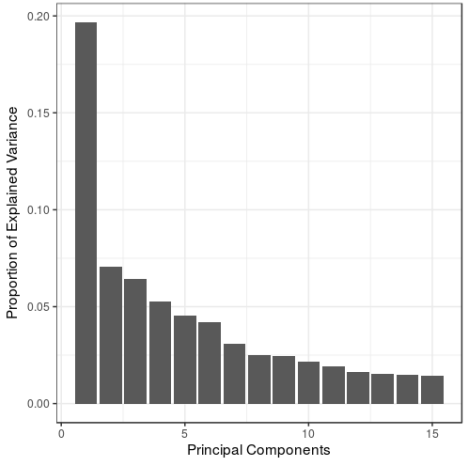
Apply KPCA (could have been clustering, linear model, . . . , in the feature space).

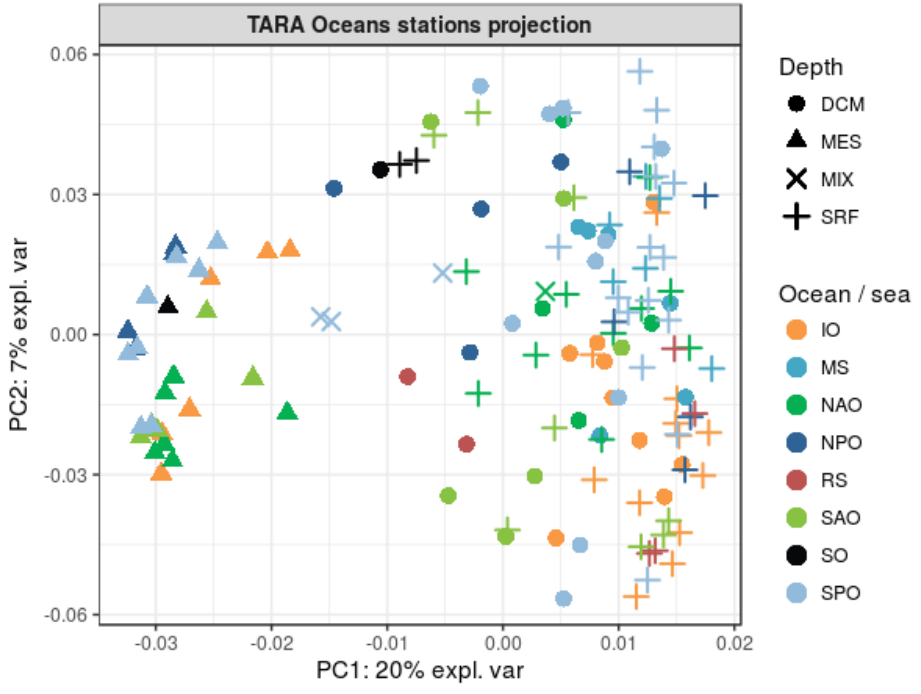
Proof of concept on [Sunagawa et al., 2015]





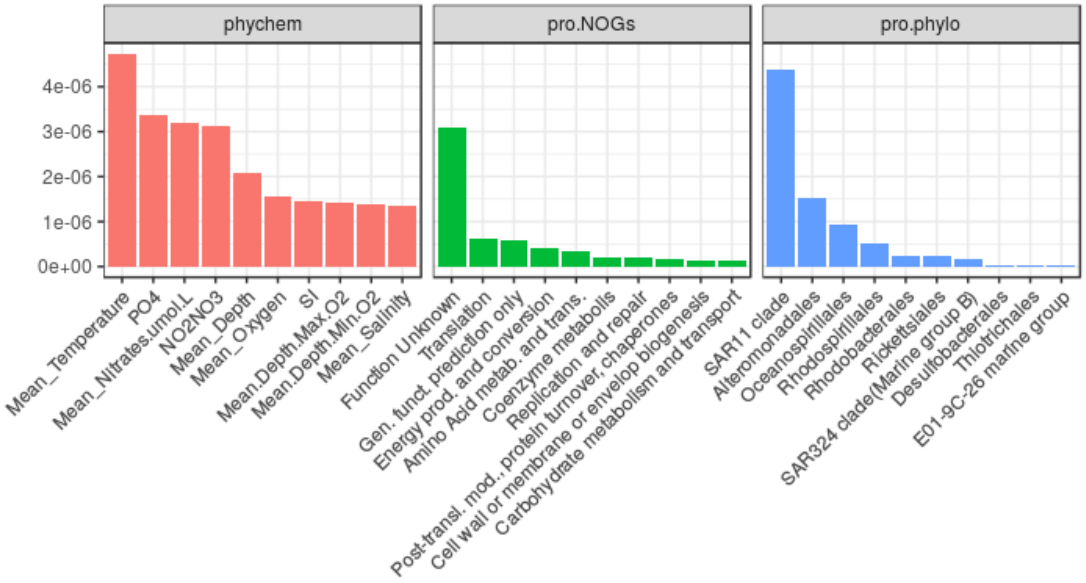
Proof of concept on [Sunagawa et al., 2015]



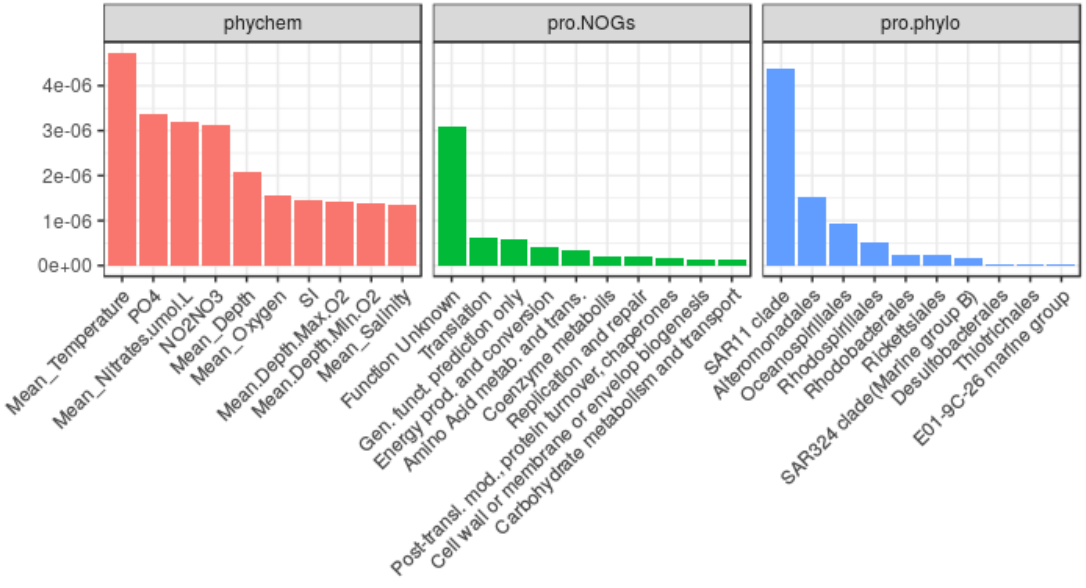


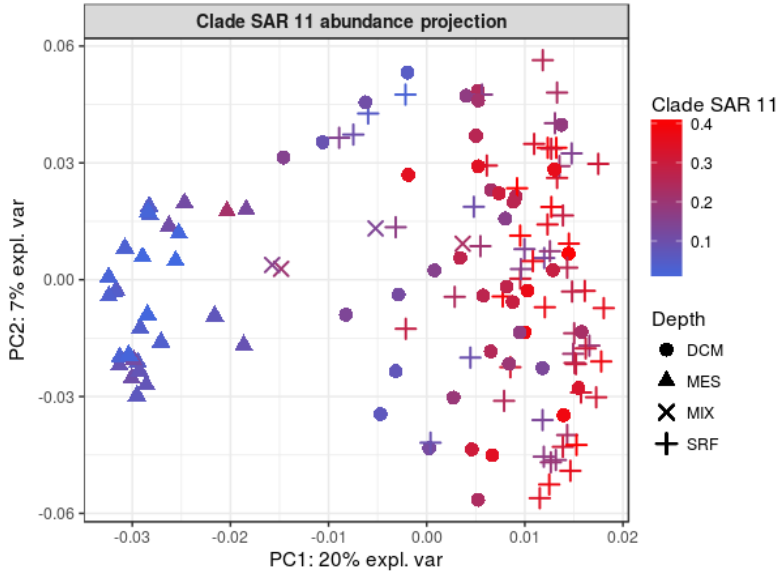
- samples are separated by their depth layer of origin, *i.e.*, SRF, DCM or MES, with stronger differences for MES samples.

Proof of concept on [Sunagawa et al., 2015]



Proof of concept on [Sunagawa et al., 2015]

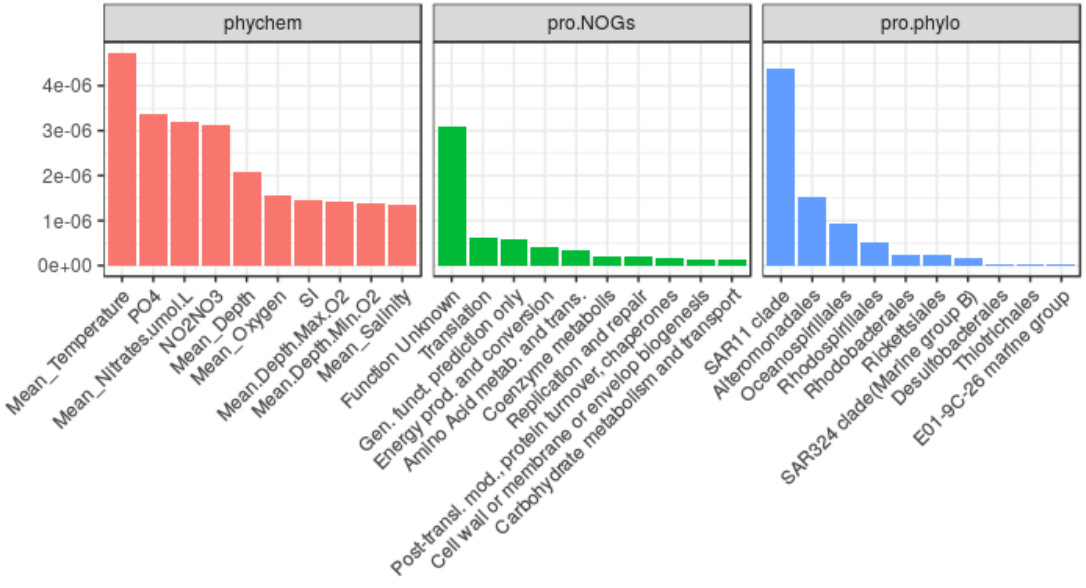


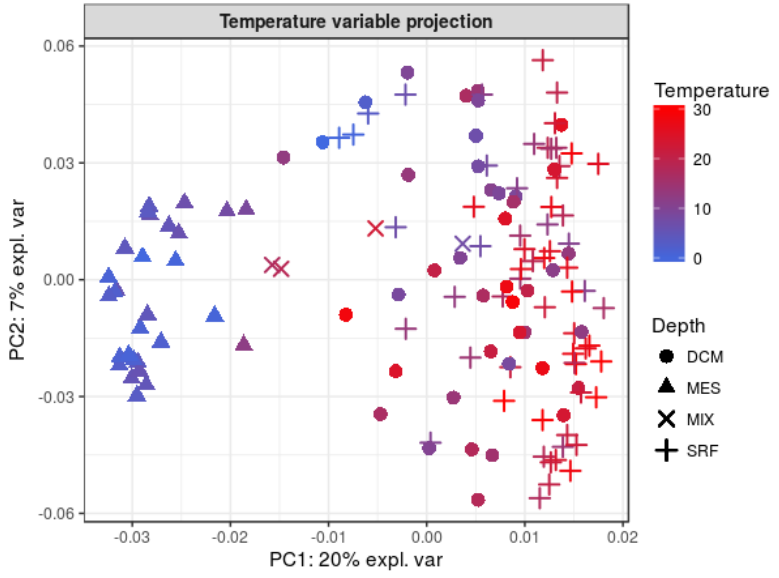


[Sunagawa et al., 2015]

- Proteobacteria (clade SAR11 (Alphaproteobacteria)) dominate the sampled areas.

Proof of concept on [Sunagawa et al., 2015]

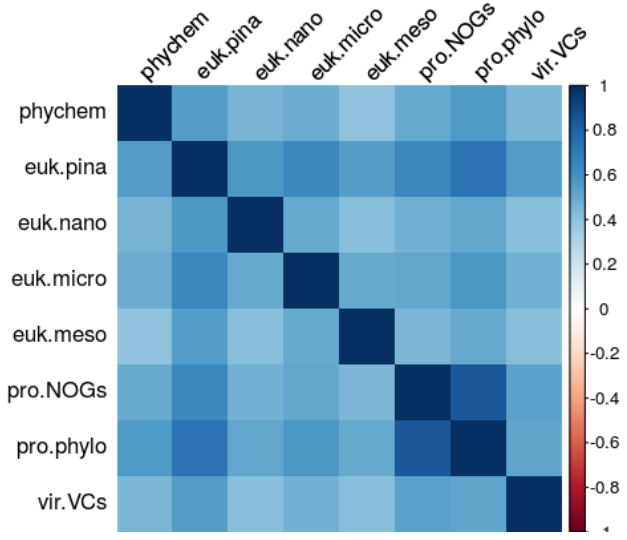




[Sunagawa et al., 2015]

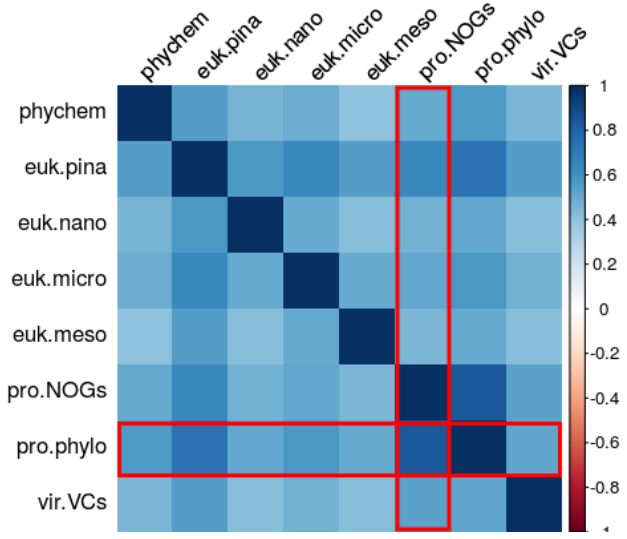
- Proteobacteria (clade SAR11 (Alphaproteobacteria)) dominate the sampled areas.
- Vertical stratification mostly driven by temperature.

Integrating all *Tara* Oceans data sets



Similarities between kernels (STATIS)

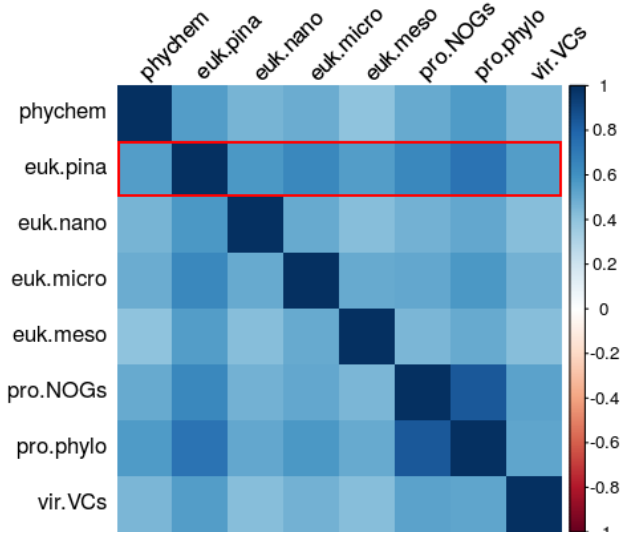
Integrating all *Tara* Oceans data sets



Similarities between kernels (STATIS)

- High similarities between prokaryote phylogenomic tree (pro.phylo) and prokaryote functional composition (pro.NOGs).

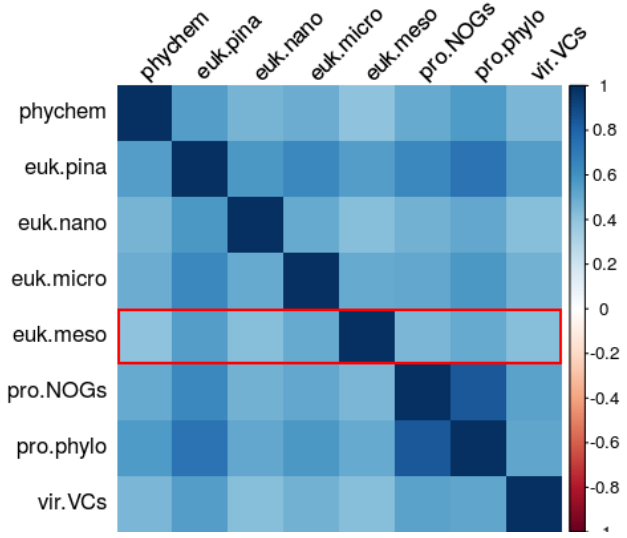
Integrating all *Tara* Oceans data sets



Similarities between kernels (STATIS)

- High similarities between prokaryote phylogenomic tree (pro.phylo) and prokaryote functional composition (pro.NOGs).
- High similarities between pico-nano-plankton (euk.pina) and other datasets: piconanoplankton communities are more homogeneous across the world's ocean ([de Vargas et al., 2015]).

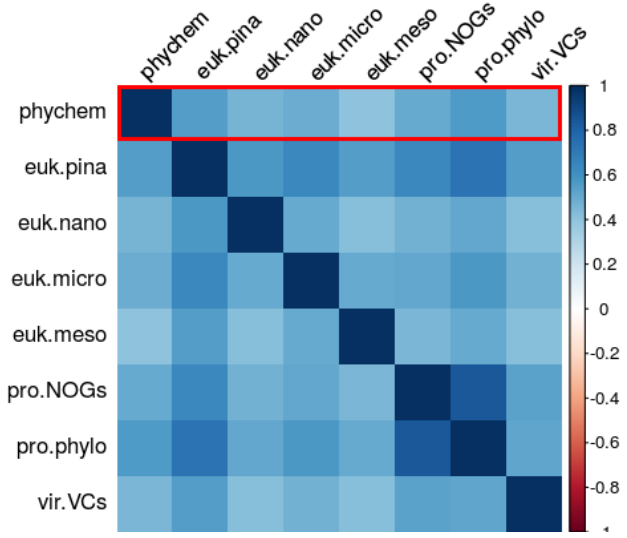
Integrating all *Tara* Oceans data sets



Similarities between kernels (STATIS)

- Low similarities between meso-plankton (euk.meso) and other datasets: strong geographical structure of mesoplanktonic communities ([de Vargas et al., 2015]).

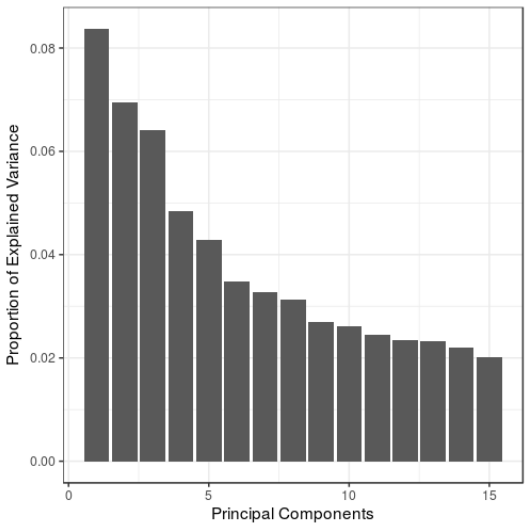
Integrating all *Tara* Oceans data sets

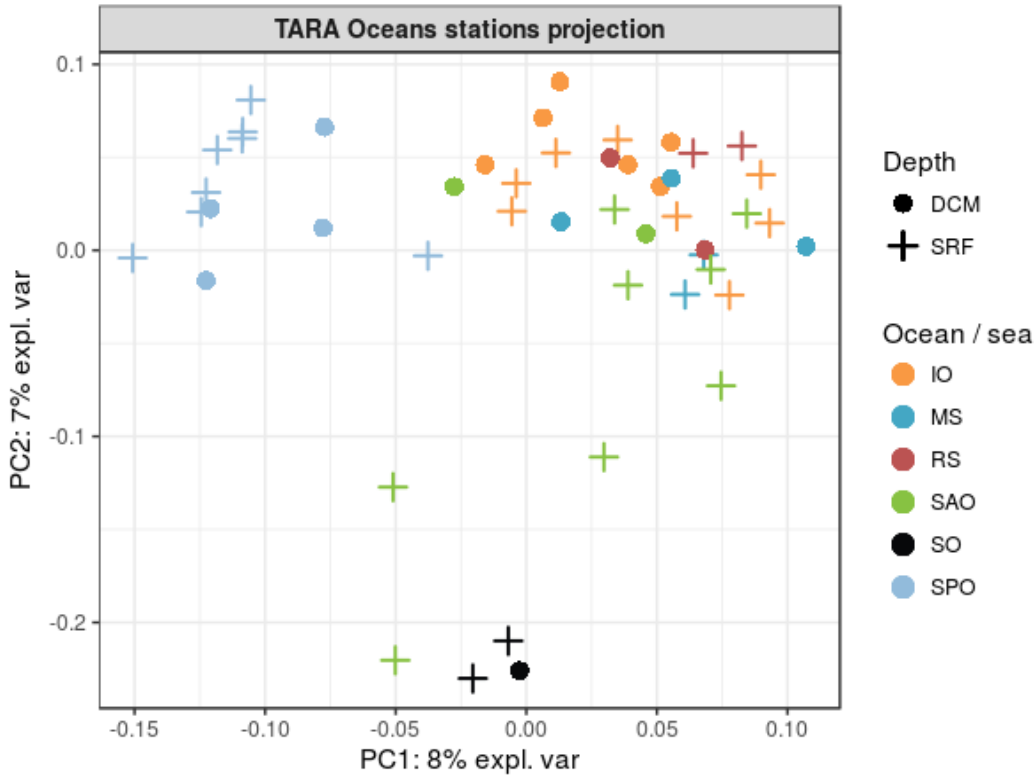


Similarities between kernels (STATIS)

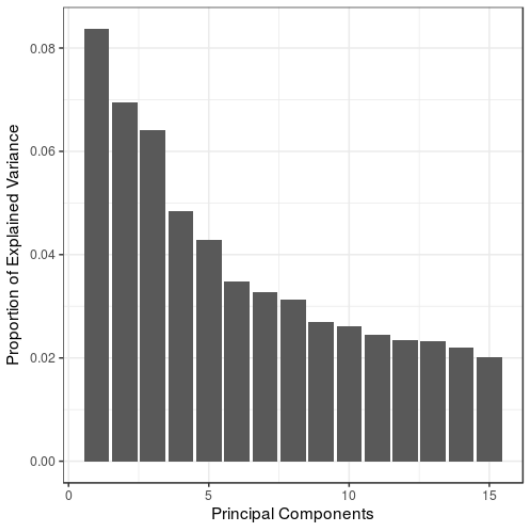
- Low similarities between meso-plankton (euk.meso) and other datasets: strong geographical structure of mesoplanktonic communities ([de Vargas et al., 2015]).
- Strongest similarities between environmental variables and small organisms than largest ones ([de Vargas et al., 2015] and [Sunagawa et al., 2015]).

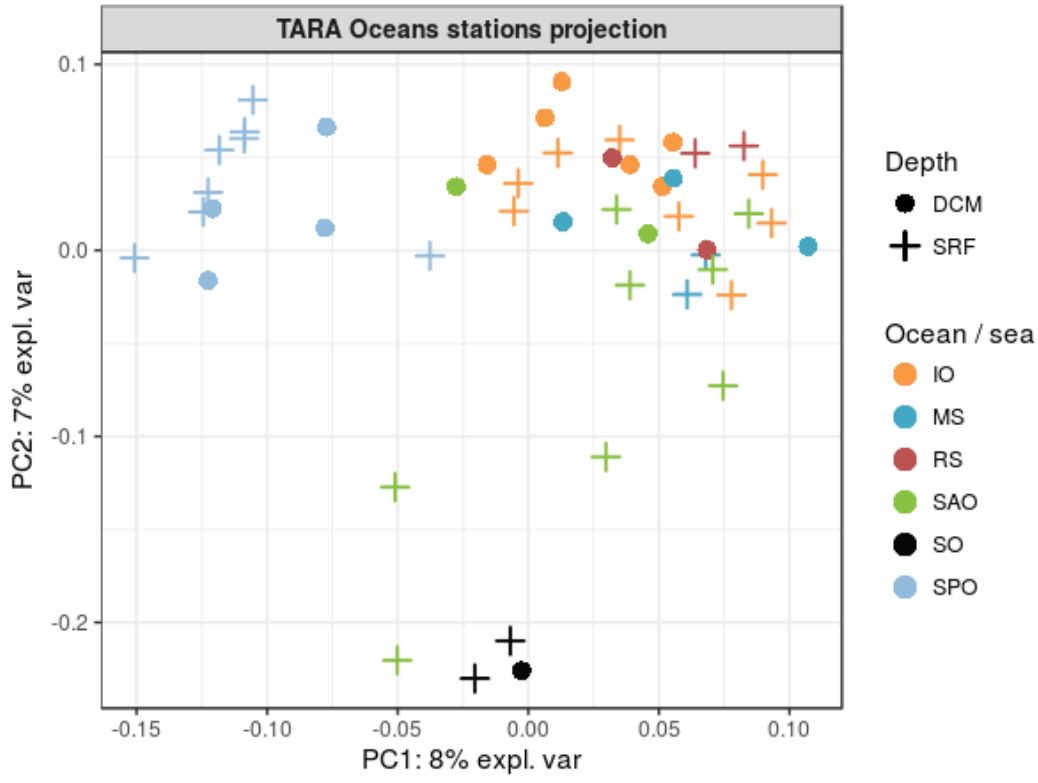
Integrating all *Tara* Oceans data sets





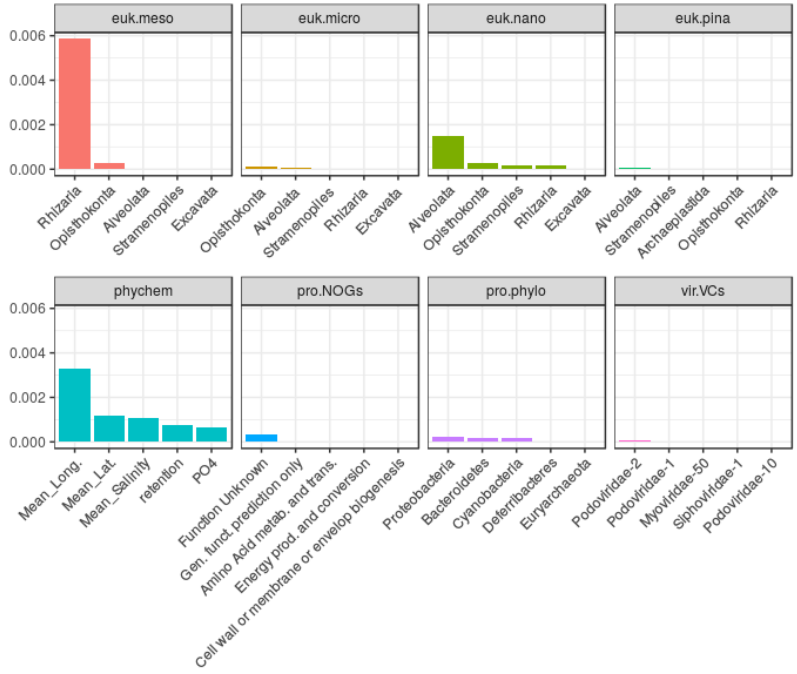
Integrating all *Tara* Oceans data sets



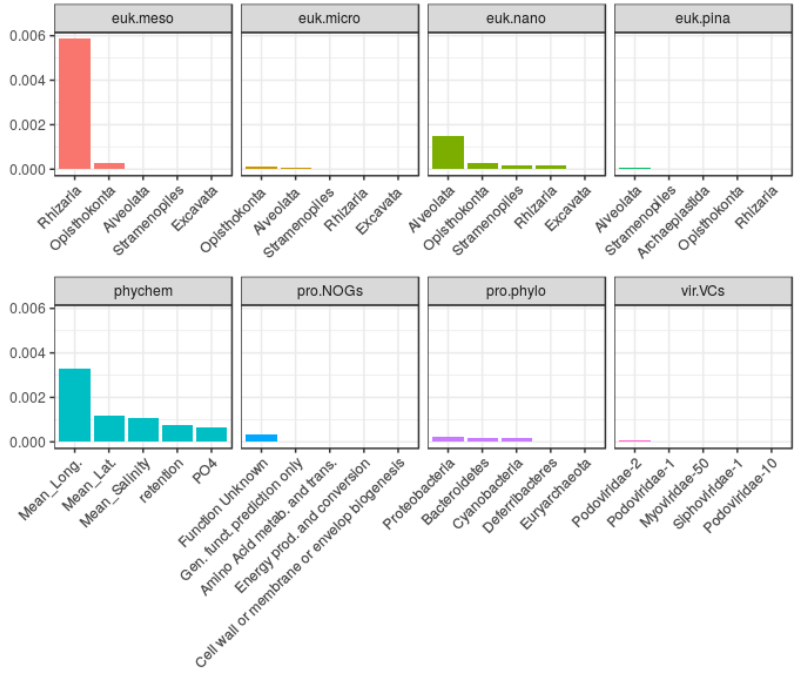


- no particular pattern in terms of depth layers but in terms of geography.

Integrating all *Tara* Oceans data sets

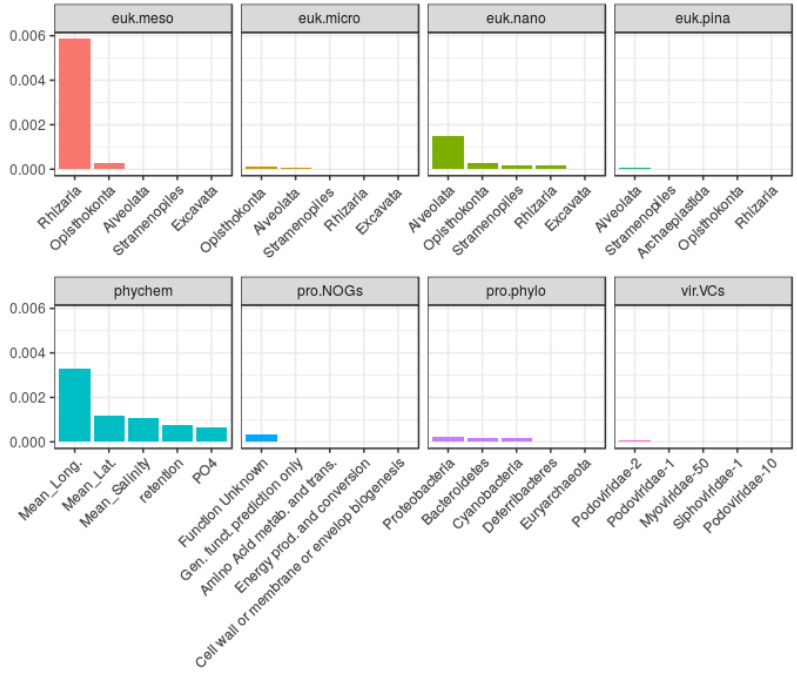


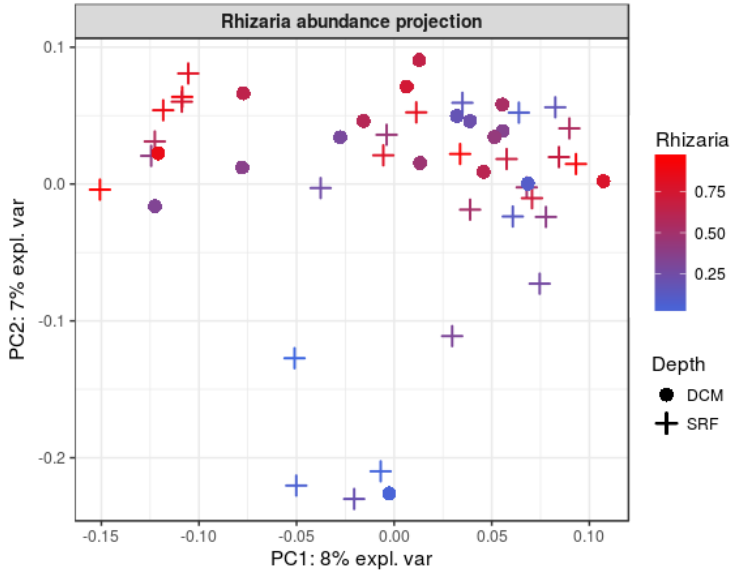
Integrating all *Tara* Oceans data sets



- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.

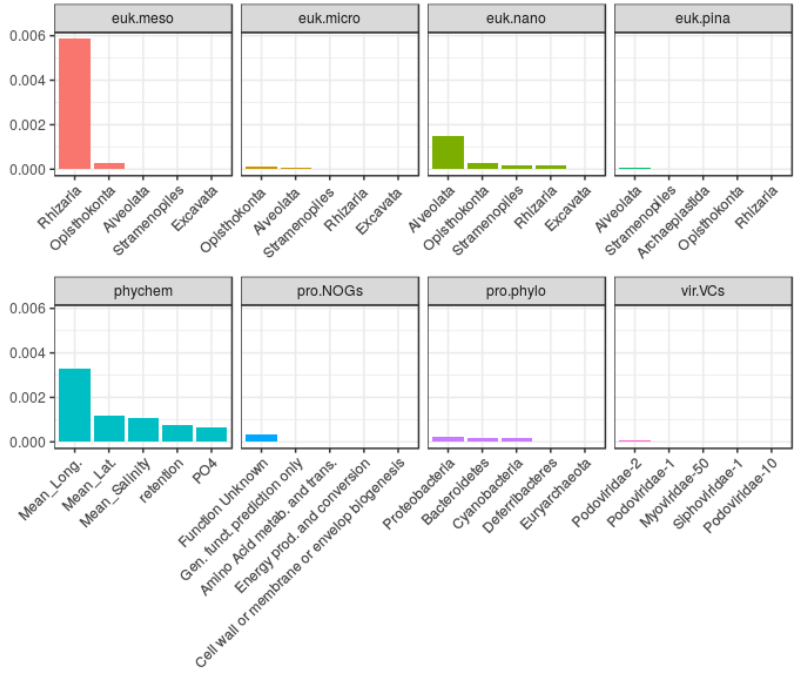
Integrating all *Tara* Oceans data sets

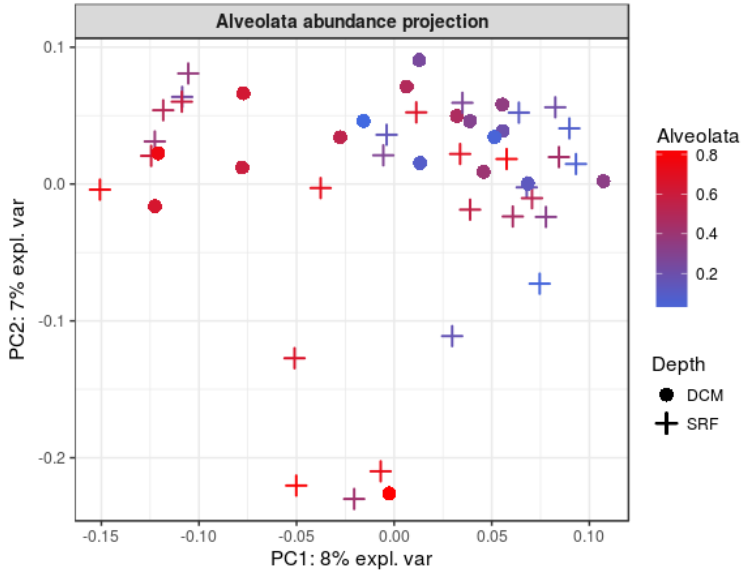




- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.
- SO and SPO epipelagic waters mainly differ in terms of Rhizarians abundances

Integrating all *Tara* Oceans data sets





- Large size organisms are the most important: *Rhizaria* and *Alveolata* phyla.
- SO and SPO epipelagic waters mainly differ in terms of Rhizarians abundances
- both of them differ from the other studied waters in terms of alveolata abundances.

Conclusions & Perspectives

What did we do?

- Integrate taxonomic, functional and community composition with environmental factors
- Use a K-PCA to visualize the datasets in an integrated way and improved its interpretability by assessing the influence of input variables in a generic way.
- Learn the kernels weights using MKL algorithms in order to understand their respective importance/contribution

$\Rightarrow$ Give access to a fast insight of the different datasets within a single analysis

Conclusions & Perspectives

What did we do?

- Integrate taxonomic, functional and community composition with environmental factors
- Use a K-PCA to visualize the datasets in an integrated way and improved its interpretability by assessing the influence of input variables in a generic way.
- Learn the kernels weights using MKL algorithms in order to understand their respective importance/contribution

$\Rightarrow$ Give access to a fast insight of the different datasets within a single analysis

Availability [Mariette and Villa-Vialaneix, 2017]

- Available in the R package **mixKernel**, released on CRAN.
- Fully compatible with the **mixOmics** package, coming with a tutorial describing the approach.

Thanks for your attention!

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

[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).

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

[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).

[Gönen and Alpaydin, 2011] Gönen, M. and Alpaydin, E. (2011).
Multiple kernel learning algorithms.
Journal of Machine Learning Research, 12:2211–2268.

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

[Lin et al., 2010] Lin, Y., Liu, T., and CS., F. (2010).
Multiple kernel learning for dimensionality reduction.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33:1147–1160.

[Mariette and Villa-Vialaneix, 2017] Mariette, J. and Villa-Vialaneix, N. (2017).
Unsupervised multiple kernel learning for heterogeneous data integration.
Bioinformatics, page btx682.

[Reverter et al., 2014] Reverter, F., Vegas, E., and Oller, J. (2014).
Kernel-pca data integration with enhanced interpretability.
BMC Systems Biology, 8.

[Ritchie et al., 2015] Ritchie, M., Holzinger, E., Li, R., S.A., P., and Kim, D. (2015).
Methods of integrating data to uncover genotype-phenotype interactions.
Nature Reviews Genetics.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K. (1998).
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319.

[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).

[Zhuang et al., 2011] Zhuang, J., Wang, J., Hoi, S., and Lan, X. (2011).
Unsupervised multiple kernel clustering.
Journal of Machine Learning Research: Workshop and Conference Proceedings, 20:129–144.