



Towards a modern analysis of omics data
of the Ocean

Mission Microbiome: CEODOS and AtlantEco expeditions

As genomics sciences meet data sciences and modeling

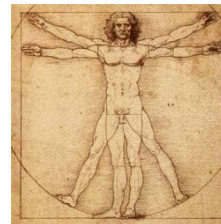
Olivier Jaillon
Genoscope – CEA

Valparaiso. May 16th 2023.

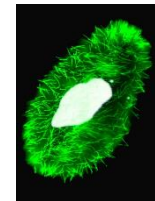
Genoscope overview

- French National Sequencing Center lead by Patrick Wincker, created in 1997 and part of the CEA since 2007.
- Provide high-throughput sequencing data to the Academic community, and carry out in-house genomic projects
- Focus on biodiversity : *de novo* sequencing and metagenomic projects (TaraOceans)

<http://www.genoscope.cns.fr>



Human



Paramecium Tetraurelia



Vitis vinifera
(grape vine)



Triticum sp
(wheat)



Quercus robur
(oak)



Musa acuminata
(banana)



Flickr/chaojikazu
Brassica napus
(seed rape)



Drosophila melanogaster



Anopheles gambiae



Arabidopsis thaliana



Tetraodon nigroviridis

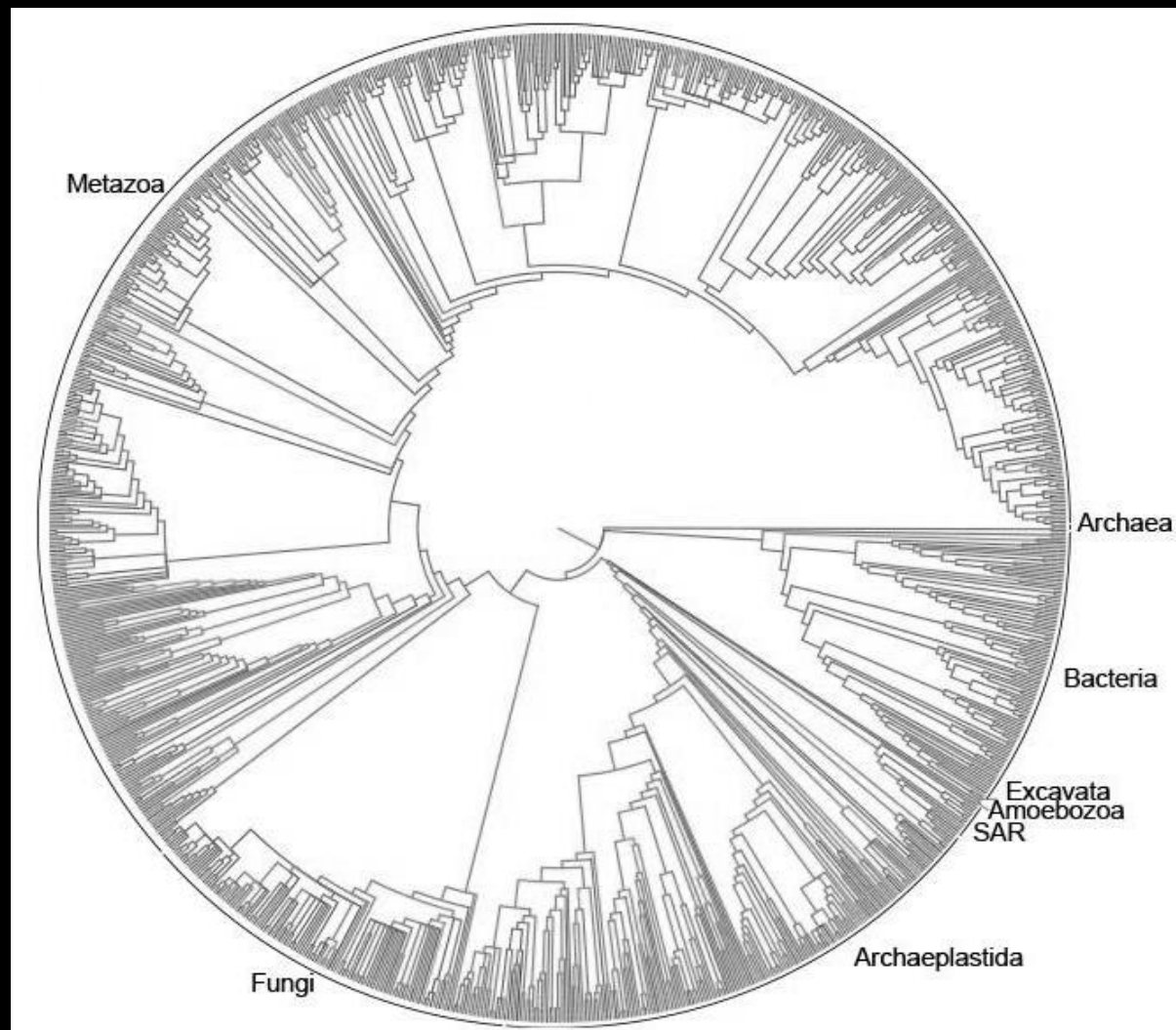


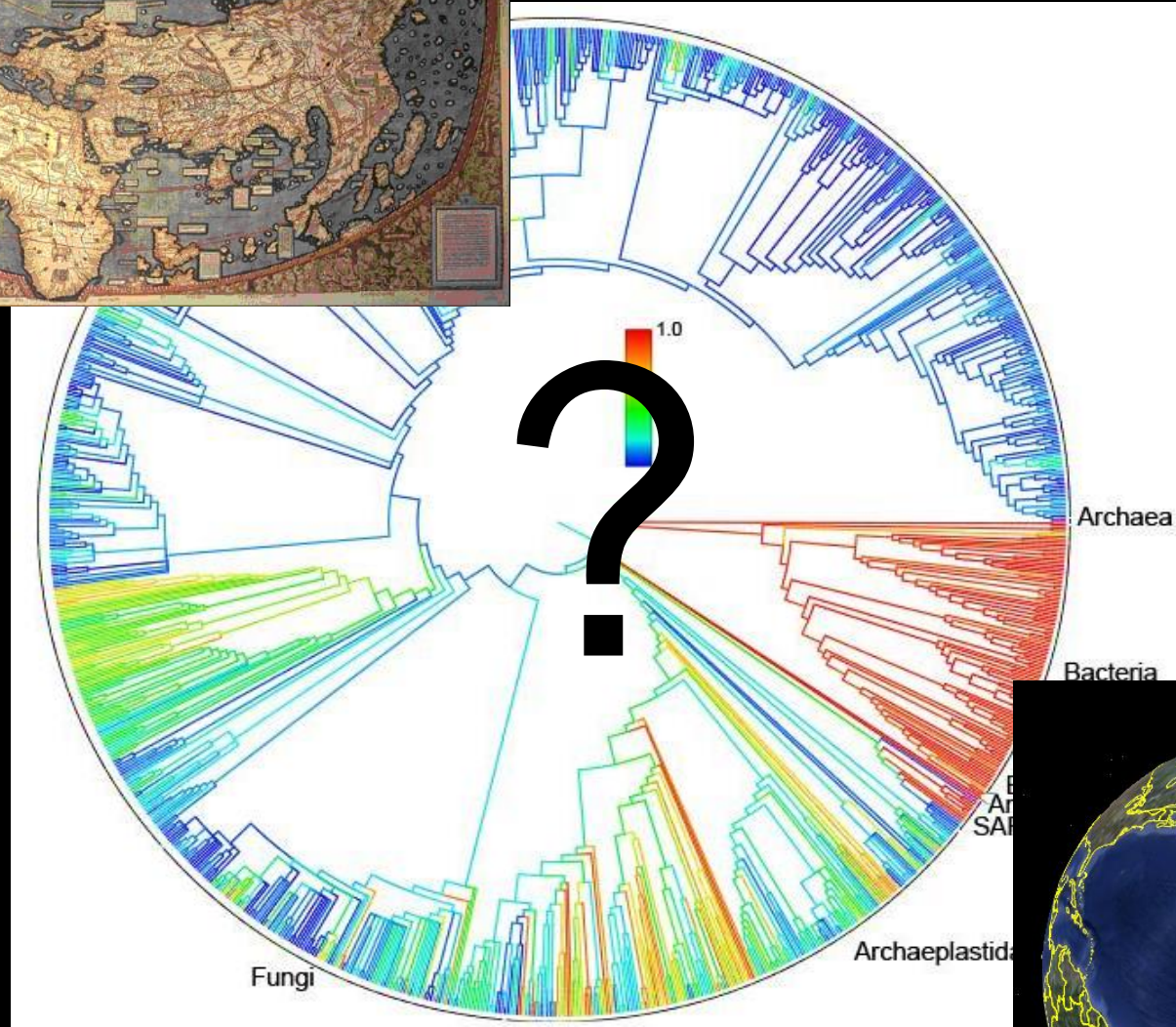
Adineta vaga
(rotifer)



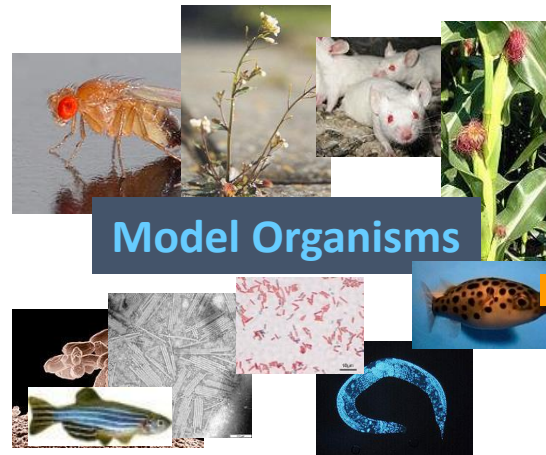
5 centuries



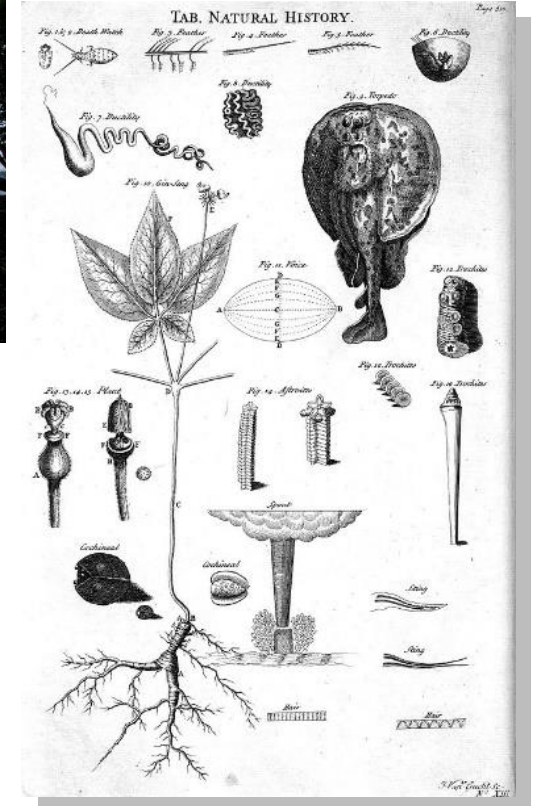




Biology may lead to 2 strategies :



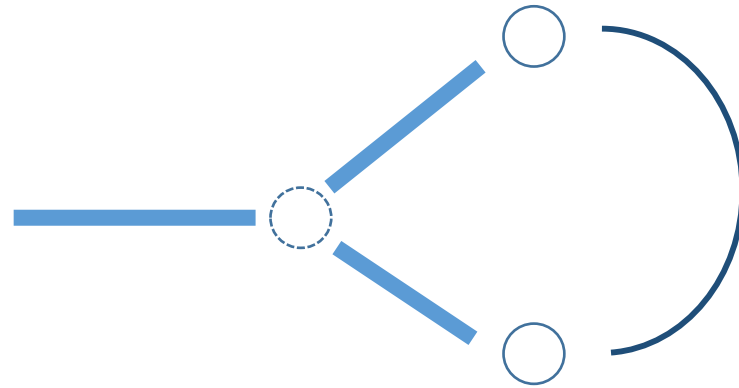
Knowledge
Technologies



- Different models for different goals
 - Evo/devo
 - Genetics
 - Genomics
 - ...
- Precise knowledge but restricted on examples.

- Catalog of organisms
- Biotic & Abiotic Connections
- Recent interest motivated by
 - global warming
 - Impact of human overpopulation
 - Screening of chemicals, enzymes of interest

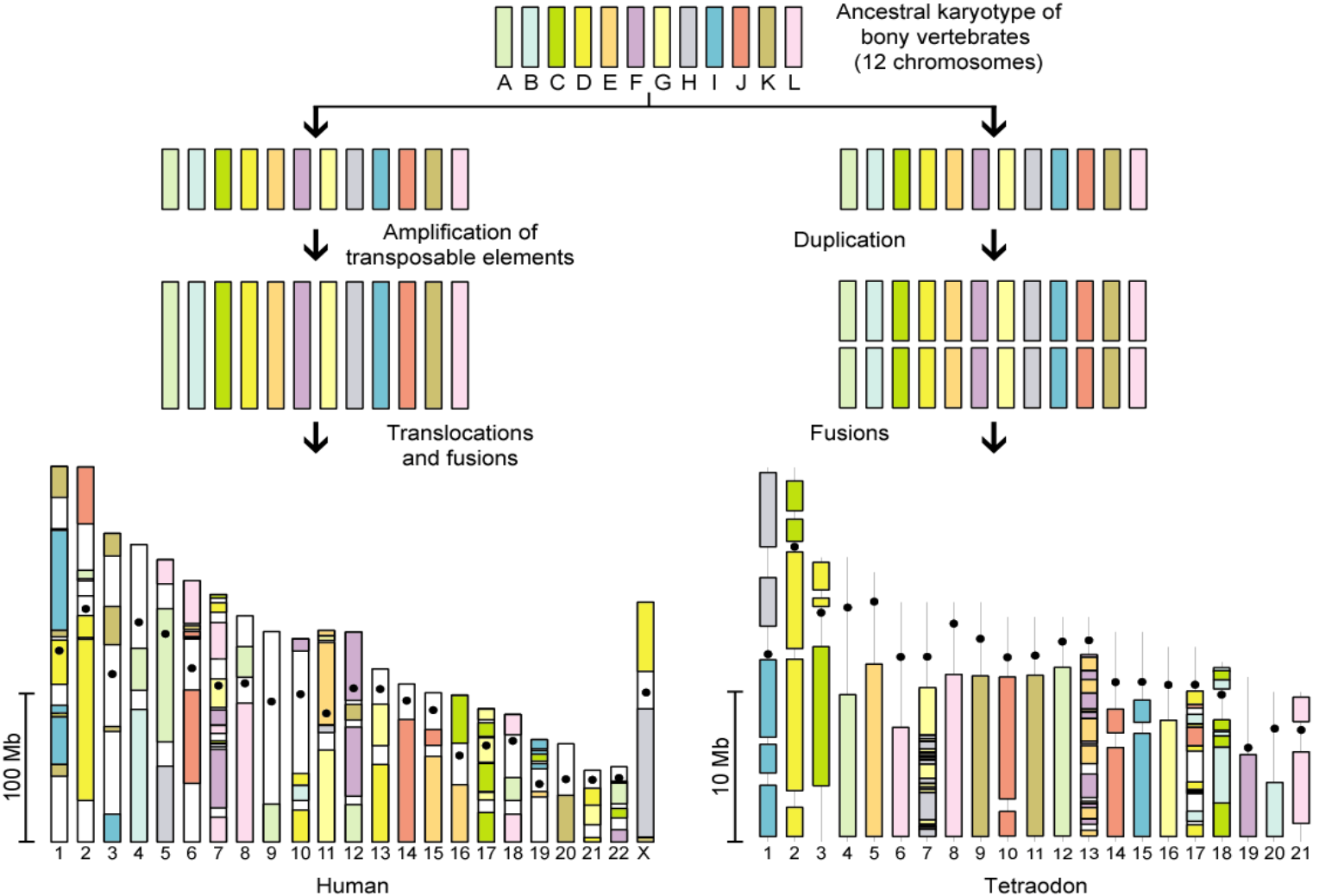
Sequence comparisons on Model Organisms



Comparing 2 sequences is always tracking back from last common ancestor.

Nothing in Biology Makes Sense Except in the Light of Evolution. Dhobzanski 1973.

GENOME SEQUENCES REVEAL MAJOR EVOLUTIONARY EVENTS

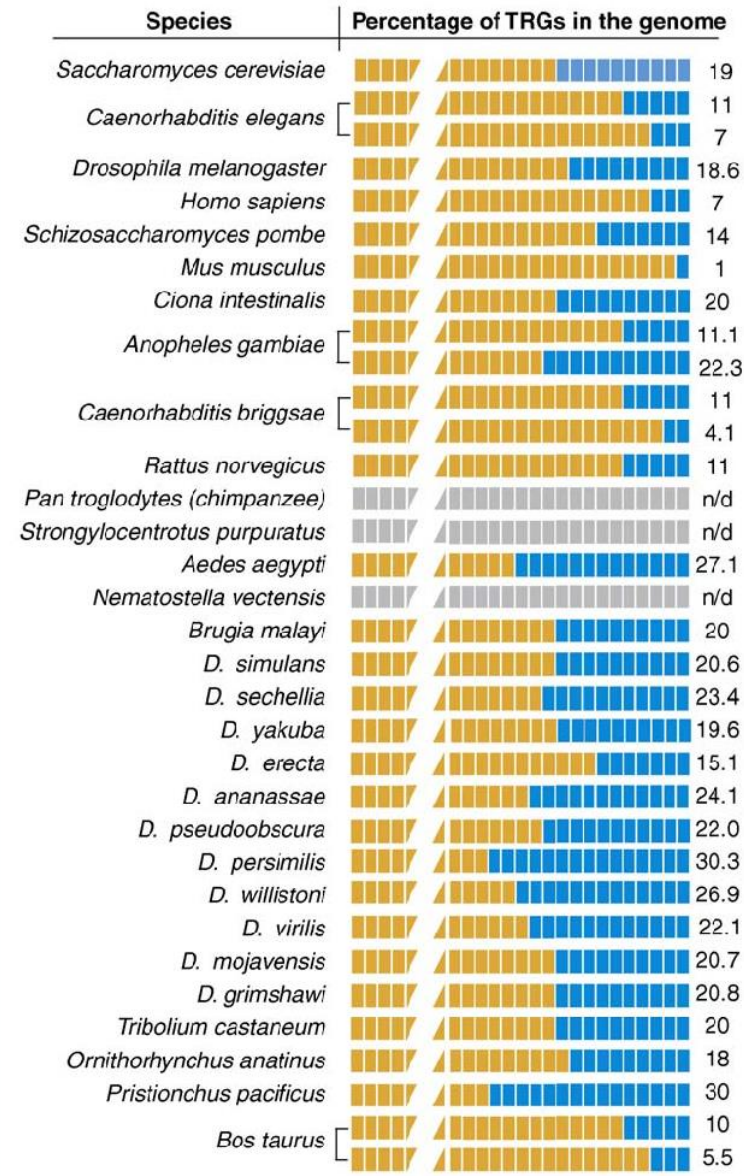


Jaillon. *Et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature. 2004



Unknown Genic diversity in Model Organisms: Taxonomically Restricted Genes

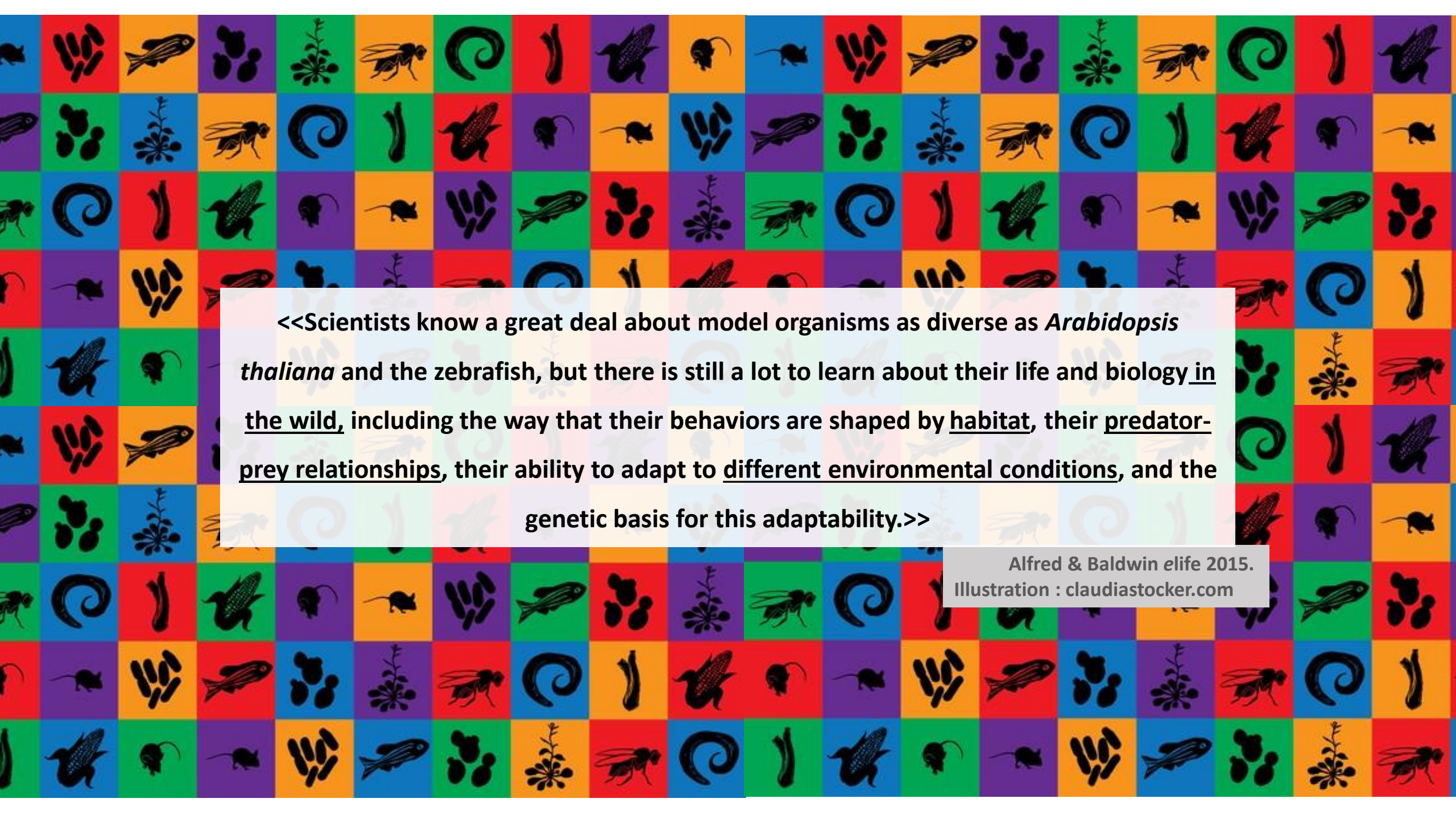
Sequenced genome



- Each block represents 2% of the predicted gene models
- Gene models without homology (putative TRGs)
- Species where the number of TRGs was not calculated



Are Model organisms representative?



<<Scientists know a great deal about model organisms as diverse as *Arabidopsis thaliana* and the zebrafish, but there is still a lot to learn about their life and biology in the wild, including the way that their behaviors are shaped by habitat, their predator-prey relationships, their ability to adapt to different environmental conditions, and the genetic basis for this adaptability.>>

Alfred & Baldwin elife 2015.
Illustration : claudiastocker.com

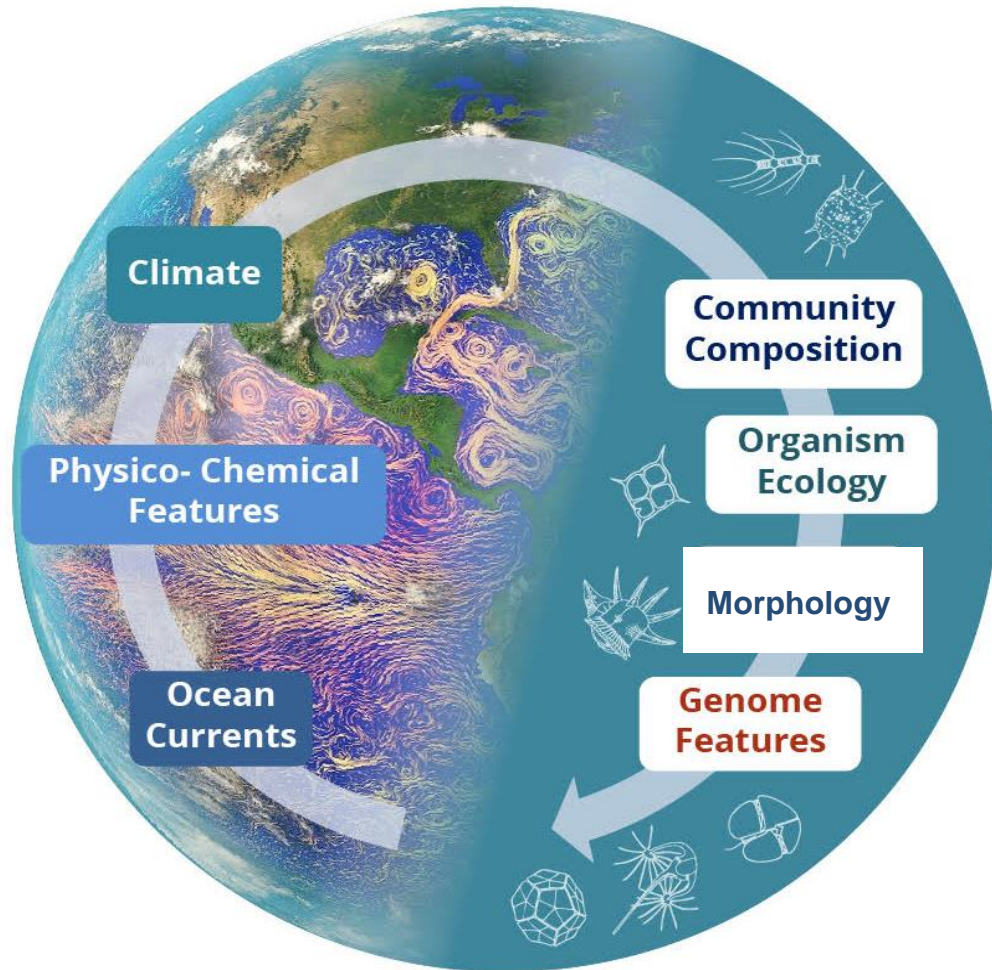
How to scale up from model organisms to a more global vision ?



Image: Thomas Pesquet



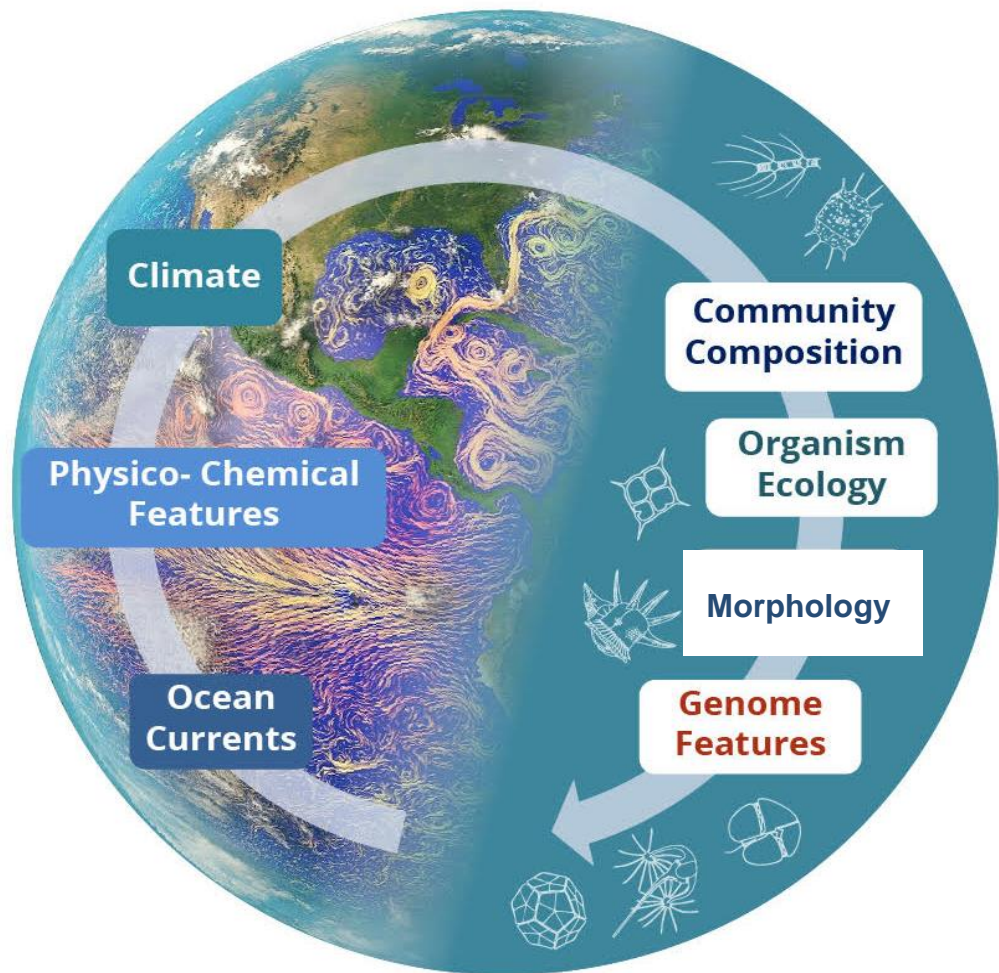
Interplay between biology and physico-chemical processes : « the seascape »



Dependencies among :

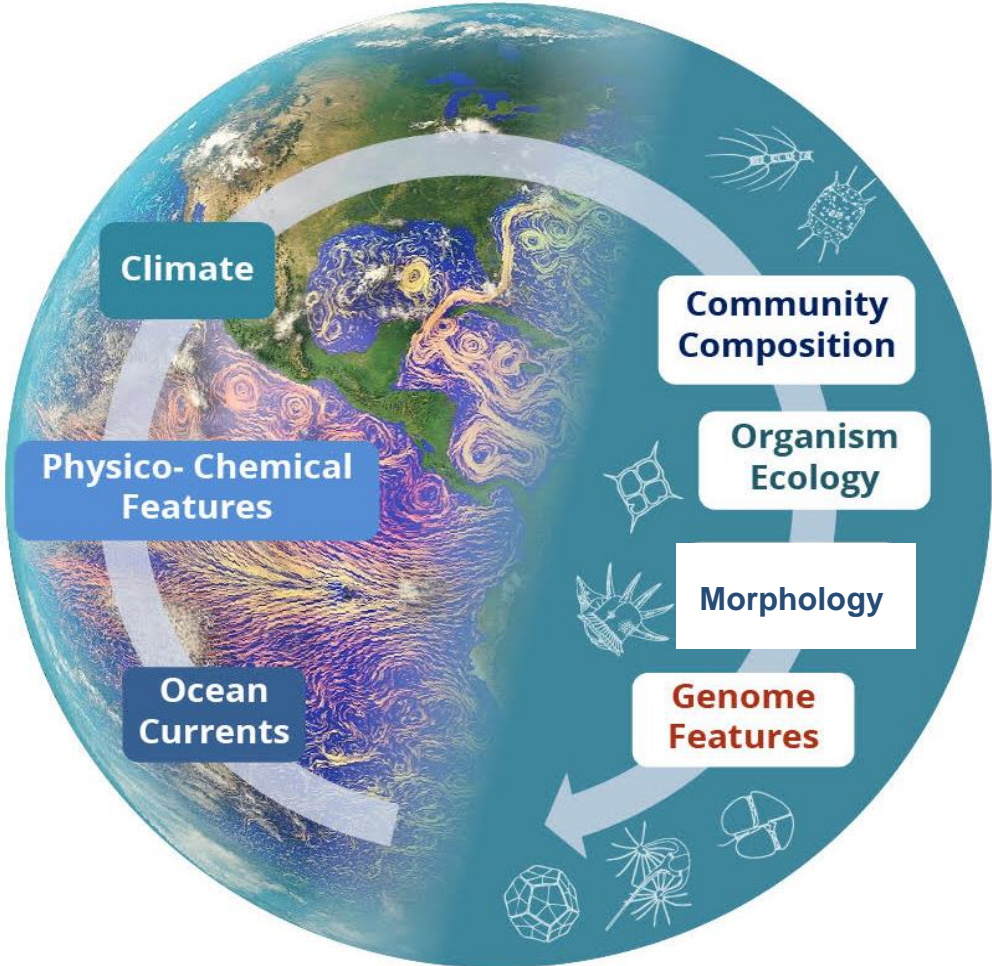
Environmental features at different scales Biological features at incremental scales

Interplay between biology and physico-chemical processes : « the seascape »



Dependencies among :
Environmental features at different scales Biological features at incremental scales

Interplay between biology and physico-chemical processes : « the seascape »



Dependencies among :
Environmental features at different scales Biological features at incremental scales

Good data, bad data and ugly data *Nat Microbiol* 4, 209 (2015)



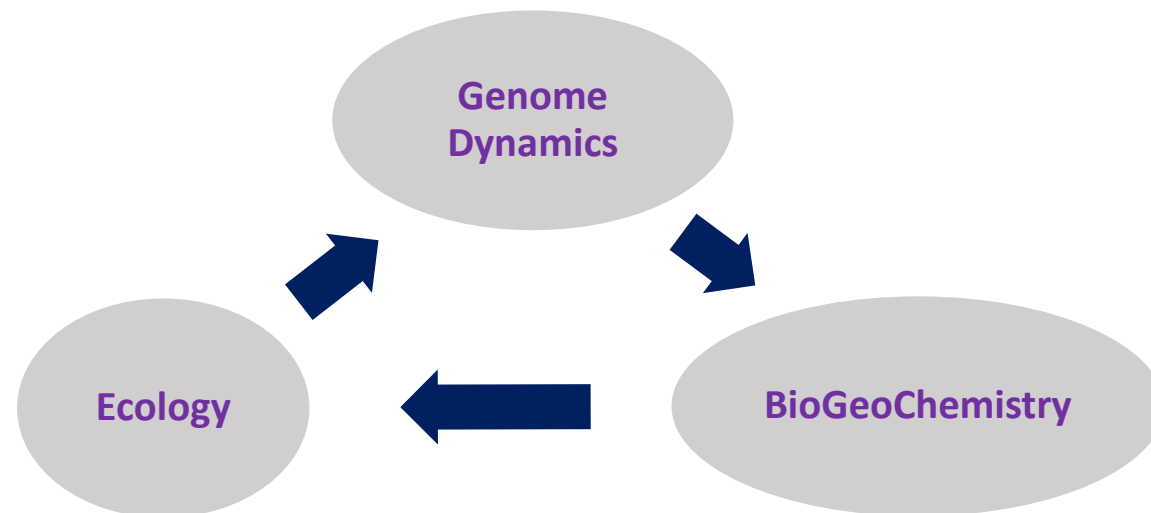
Seascape Genomics & Biogeography are linked

From Ramette A. & Tiedje J. M. Microbial Ecology 2006 :

In a broader sense, this discipline examines variation of microbial features (e.g., genetic, phenotypic, physiological) at different spatial scales, between distantly located sampling sites or along large environmental gradients.

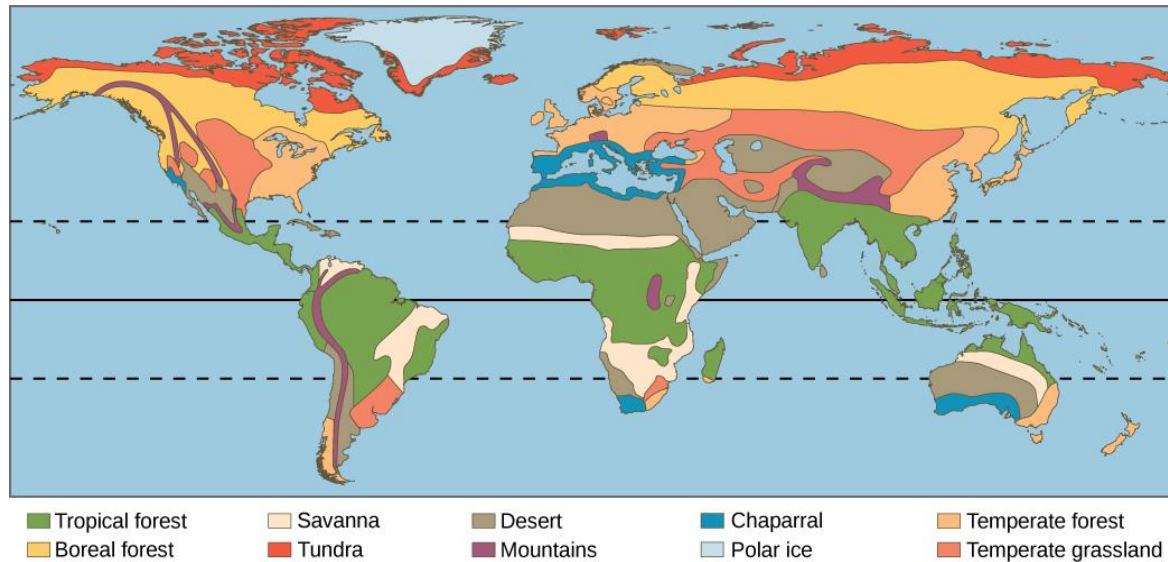
Its scope also encompasses the understanding of the processes generating and maintaining those distribution patterns.

The ultimate goals are to propose and evaluate theories regarding the creation and evolution of such diversity patterns in the environment.



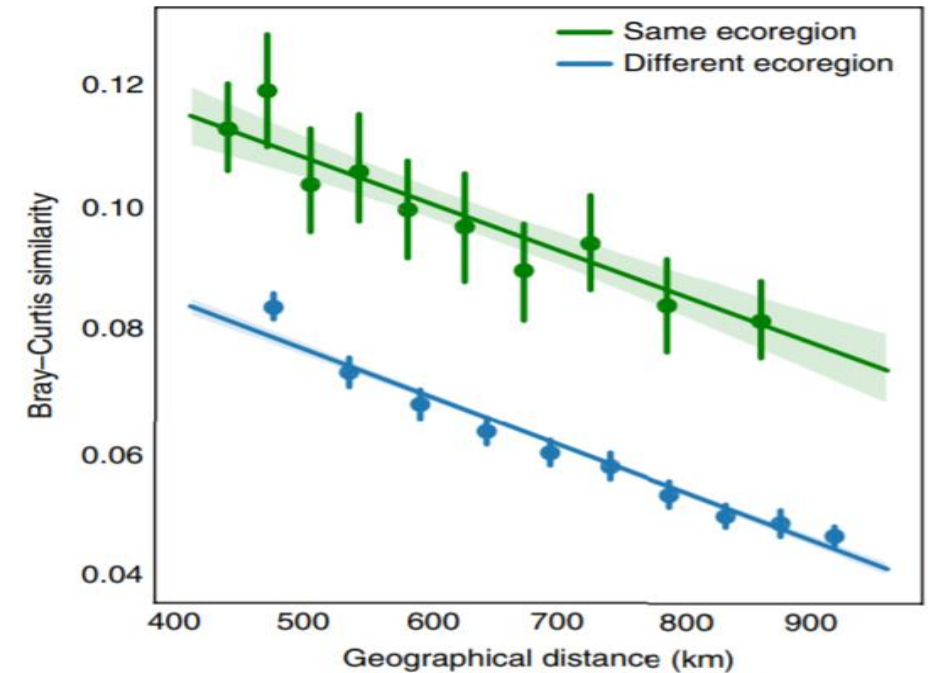
Terrestrial Biomes

Eight major terrestrial biomes defined by temperature and precipitation



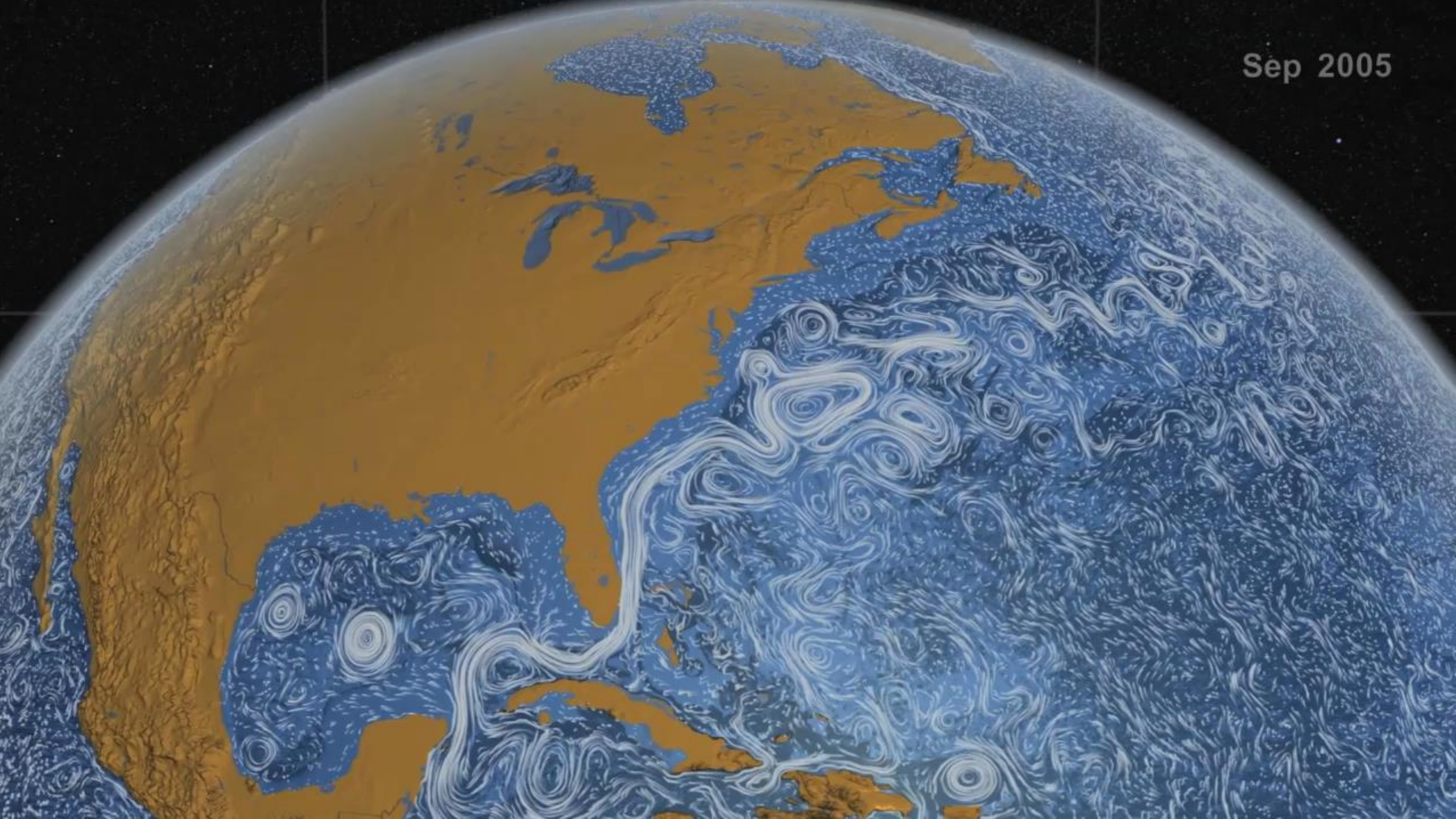
Credit: "Climate influence on terrestrial biome" by Navarras is in the Public Domain

β -diversity is proportional to Geographical distance

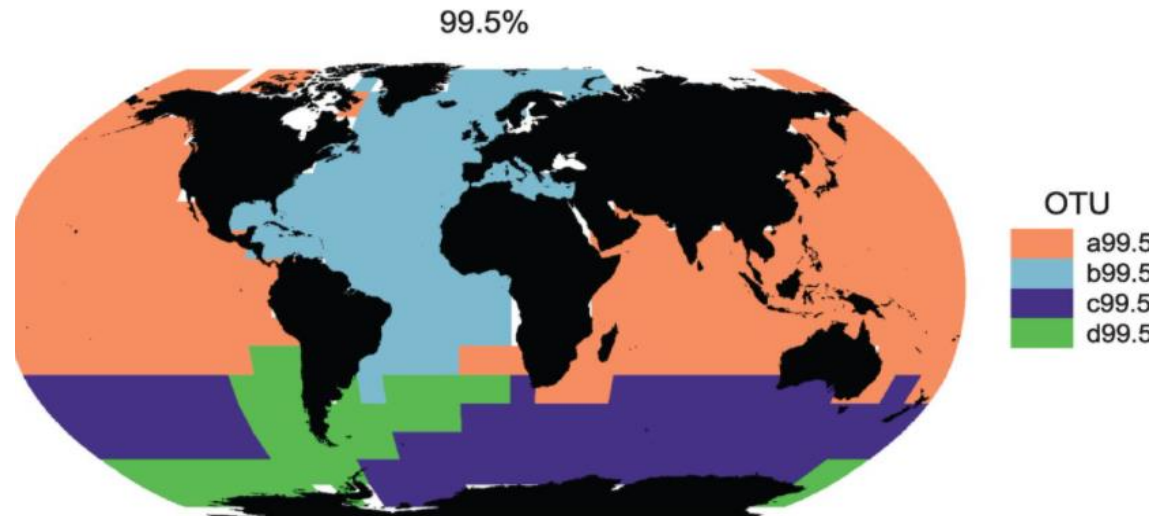
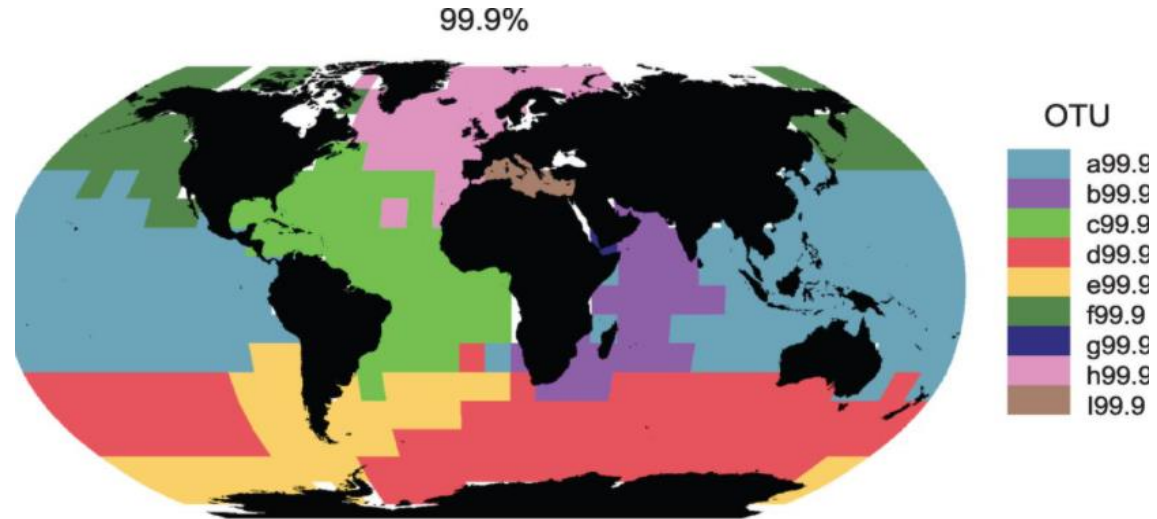


Smith et al. Nature Ecology & Evolution 2018

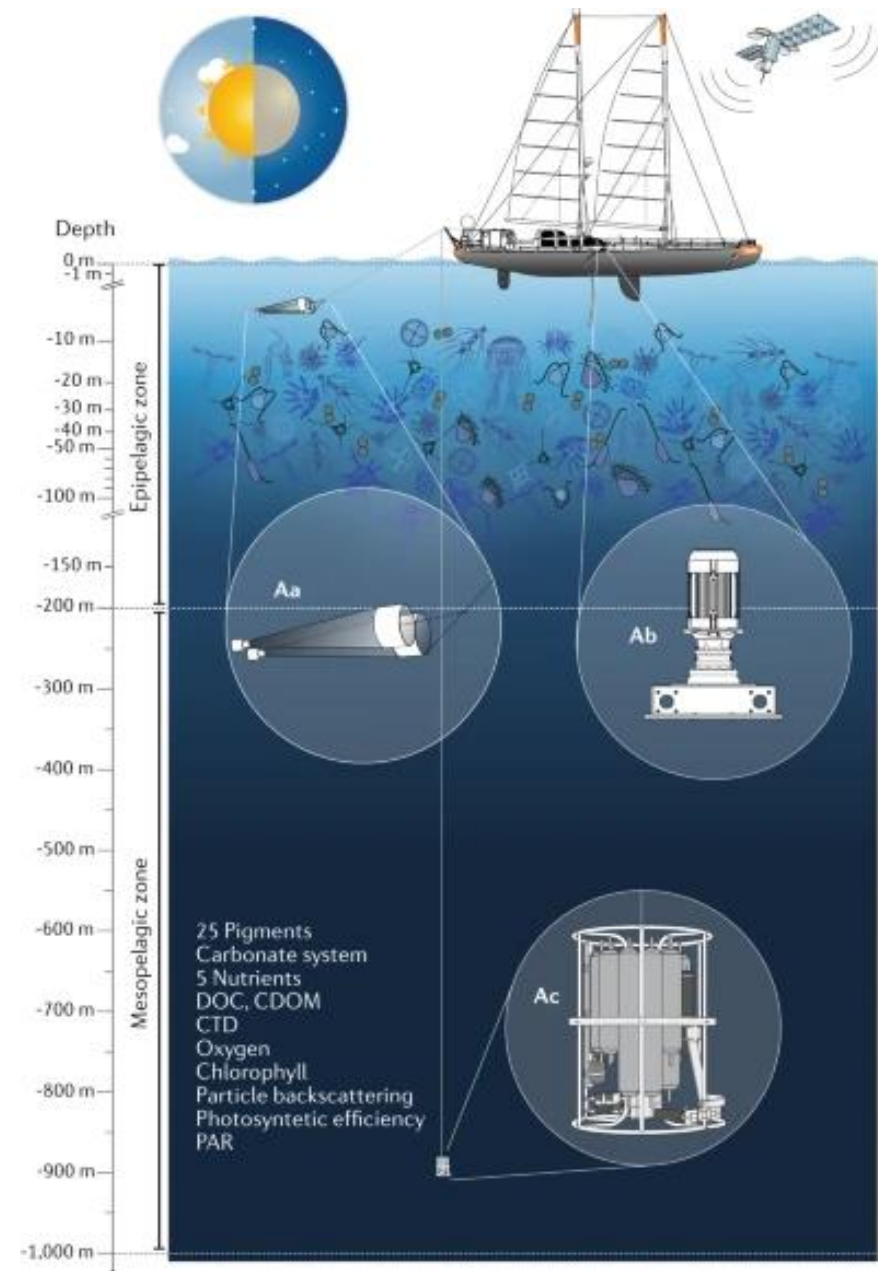
Sep 2005



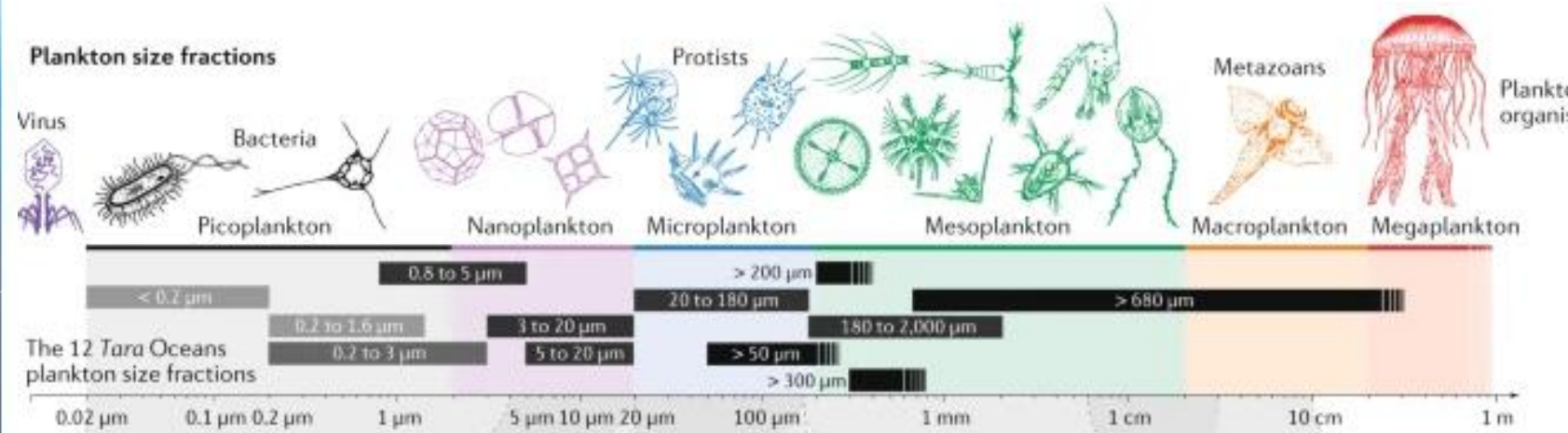
Combination of neutral evolution of DNA with Ocean Circulation



Ferdi L. Hellweger
Erik van Sebille,
Neil D. Fredrick
Science 2014

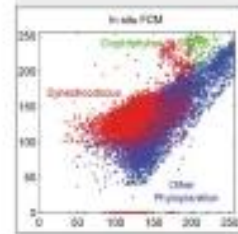


Plankton size fractions

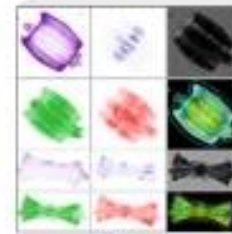


The 12 Tara Oceans plankton size fractions

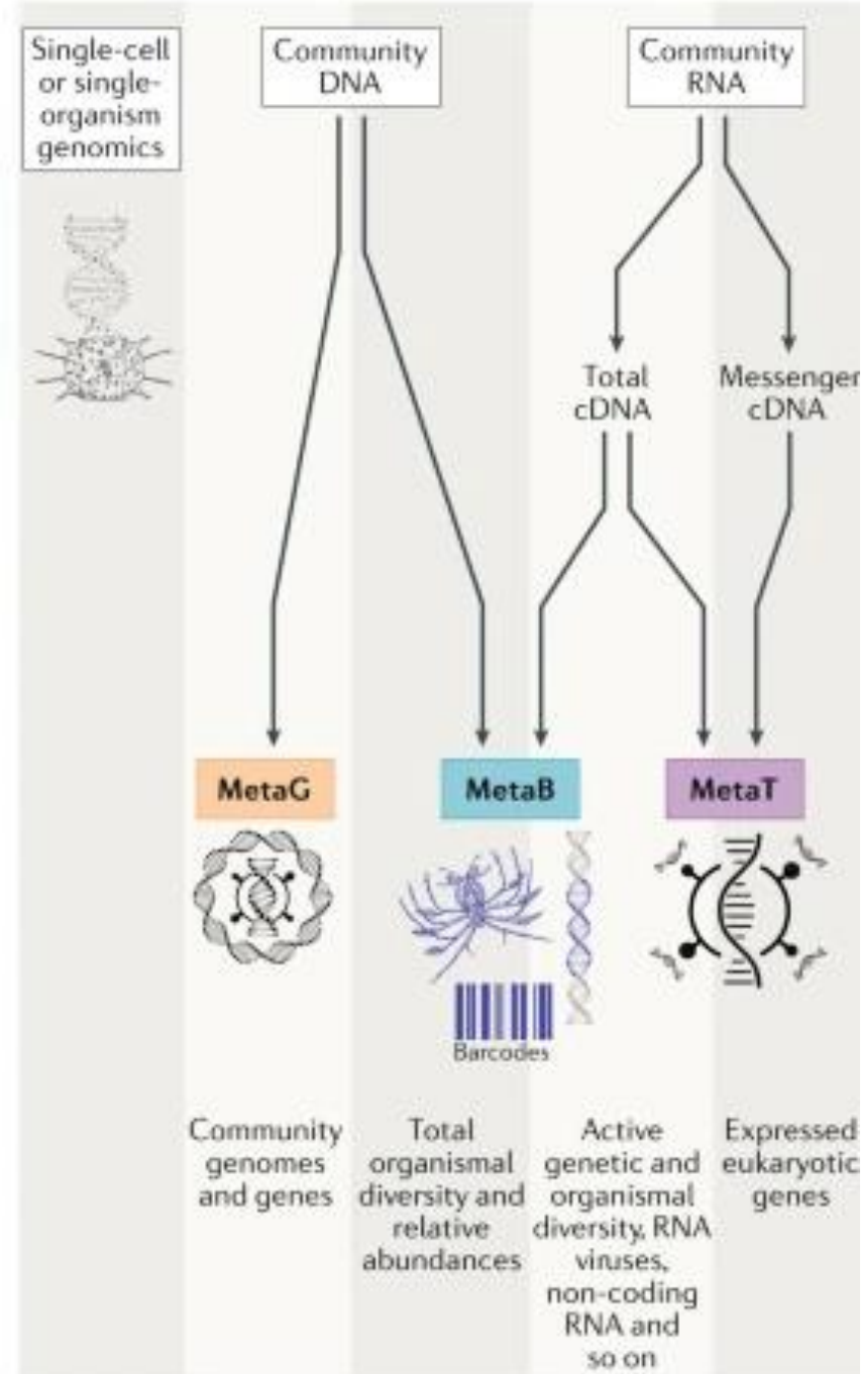
High-throughput imaging



Flow cytometry



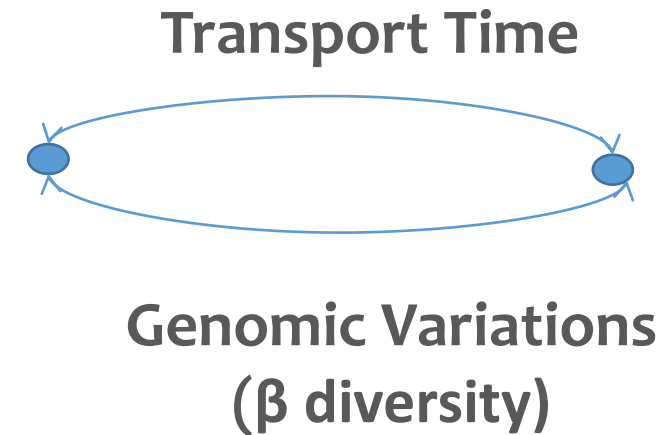
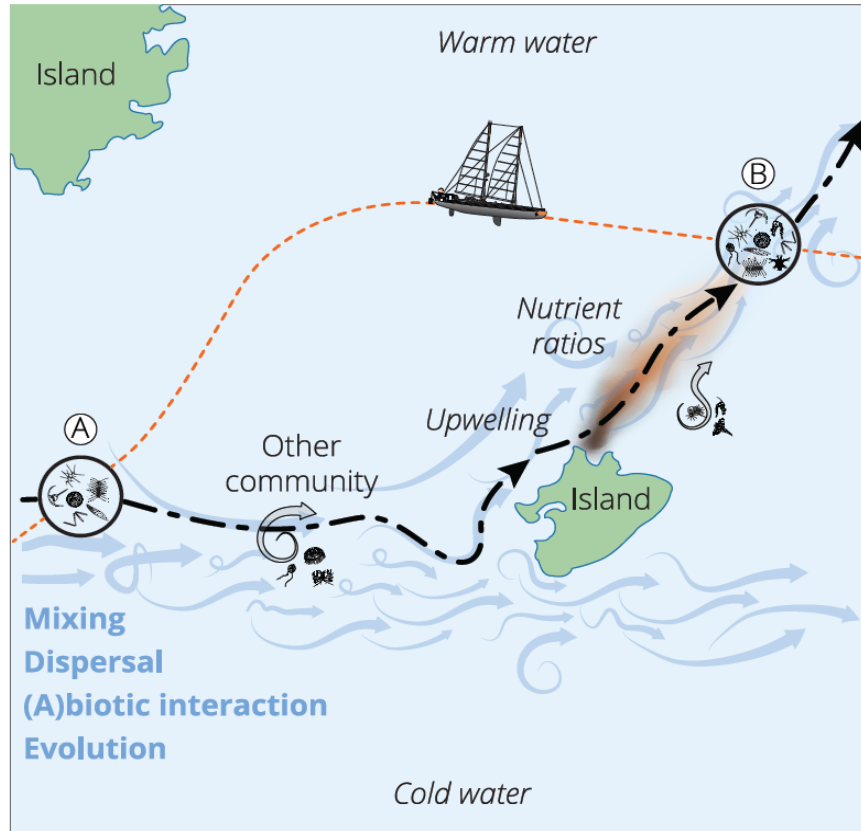
Different types of "omics" data : different goals



Sunagawa, S., *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* 18, 428–445 (2020).

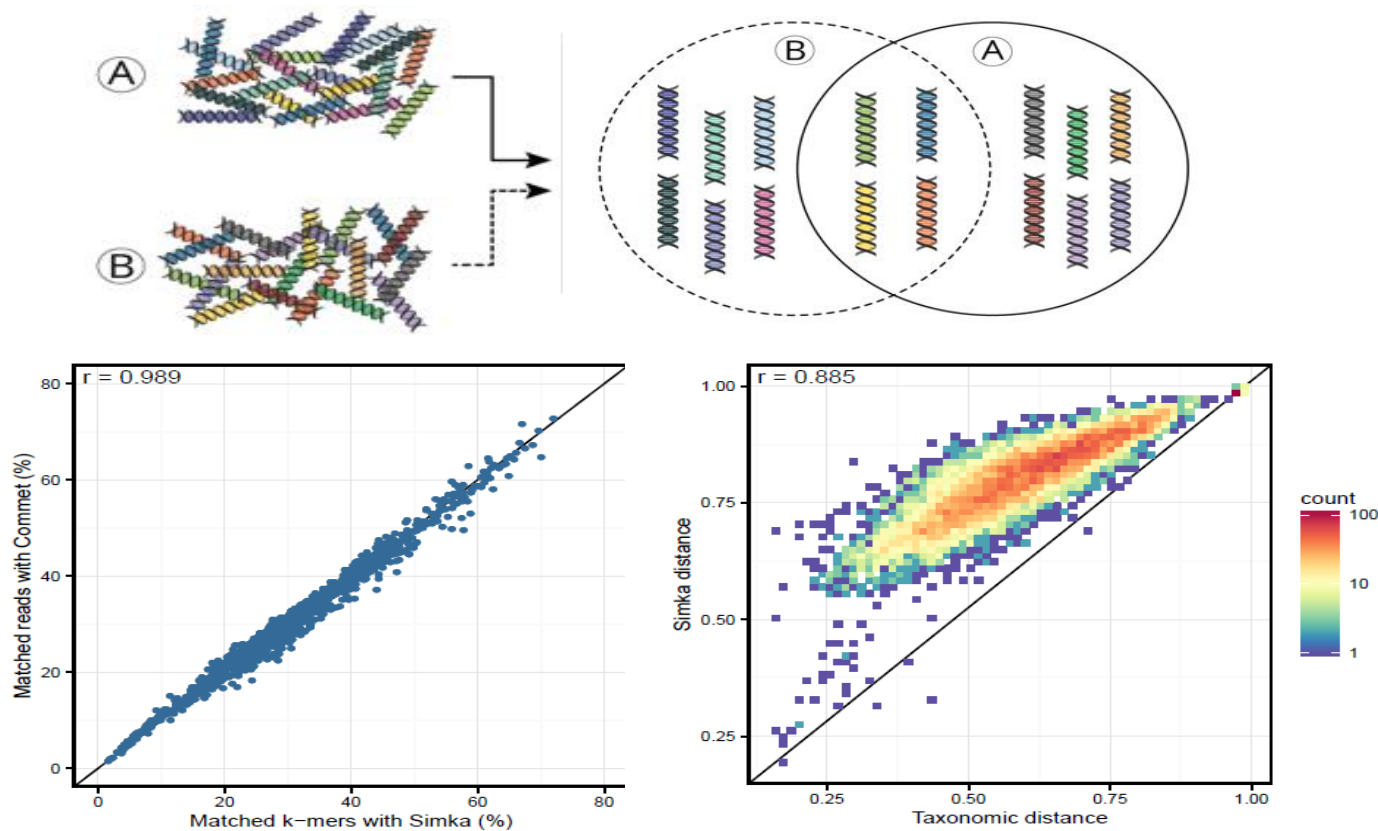
SEASCAPE GENOMICS WITH TARA OCEANS

Oceanographic variations and genomic variations are closely linked



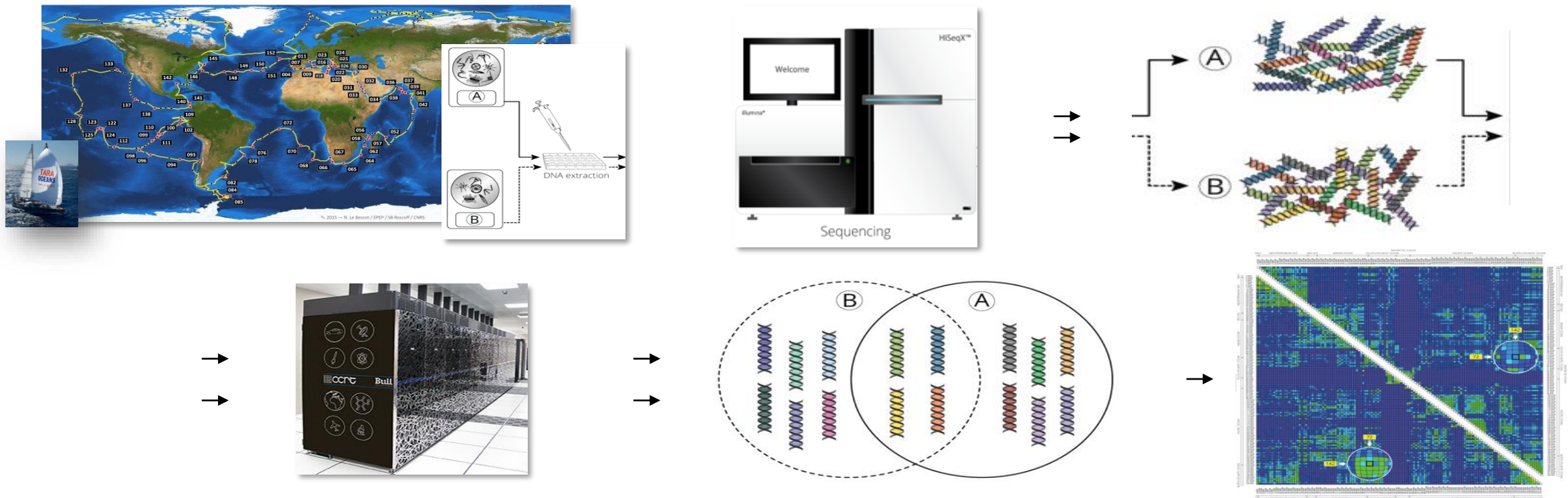
Rationale: Leverage Metagenomes in a Blind approach to reveal a structure of global ocean partitioning

- Use DNA kmers from metagenomes as biological tracer of organisms and to approximate betadiversity
- Compute beta-diversity between metagenomic samples using DNA kmers. etag : Simka tool (Benoit,G. Et al. PeerJ Computer Science, 2016; Benoit, G. ET al. *Bioinformatics*, 2020)



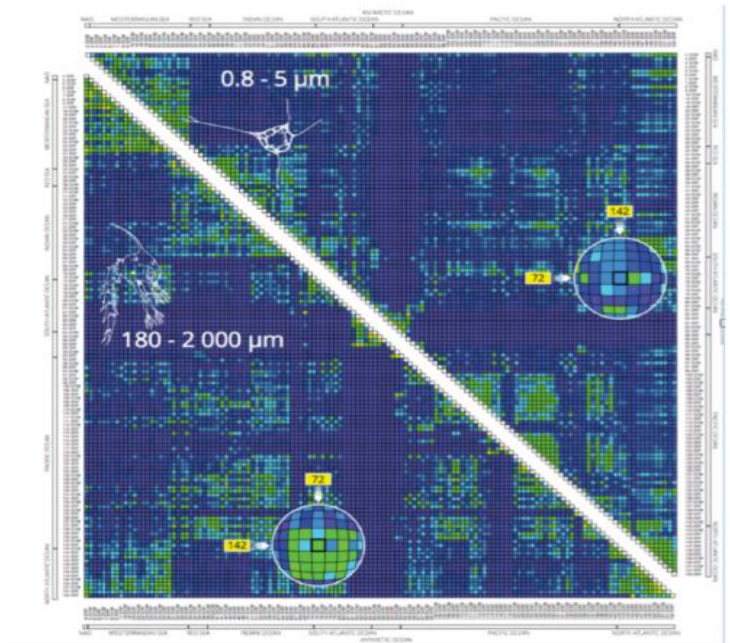
Deciphering a global Biogeography of plankton without reference

- Consider each set of metagenomic reads as representative of local plankton community.
- Compute the number of very similar reads between all pairs of samples => proxy for a “metagenomic distance” between communities



Deciphering a global Biogeography of plankton without reference

Simka Matrix

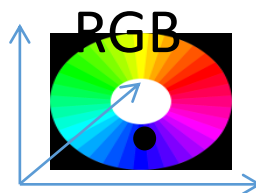
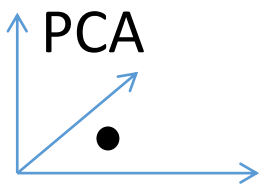


UPGMA clustering

Relational Trees of Samples



RGB Coloring of samples

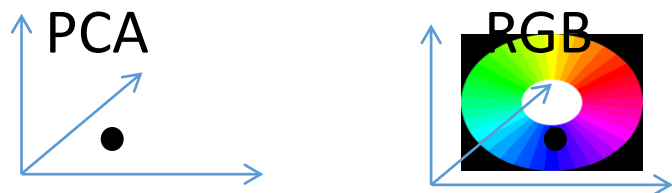


Samples colored in RGB values using 3 main axis of Principal Component analysis => differences in colors reflect beta-diversity among samples.

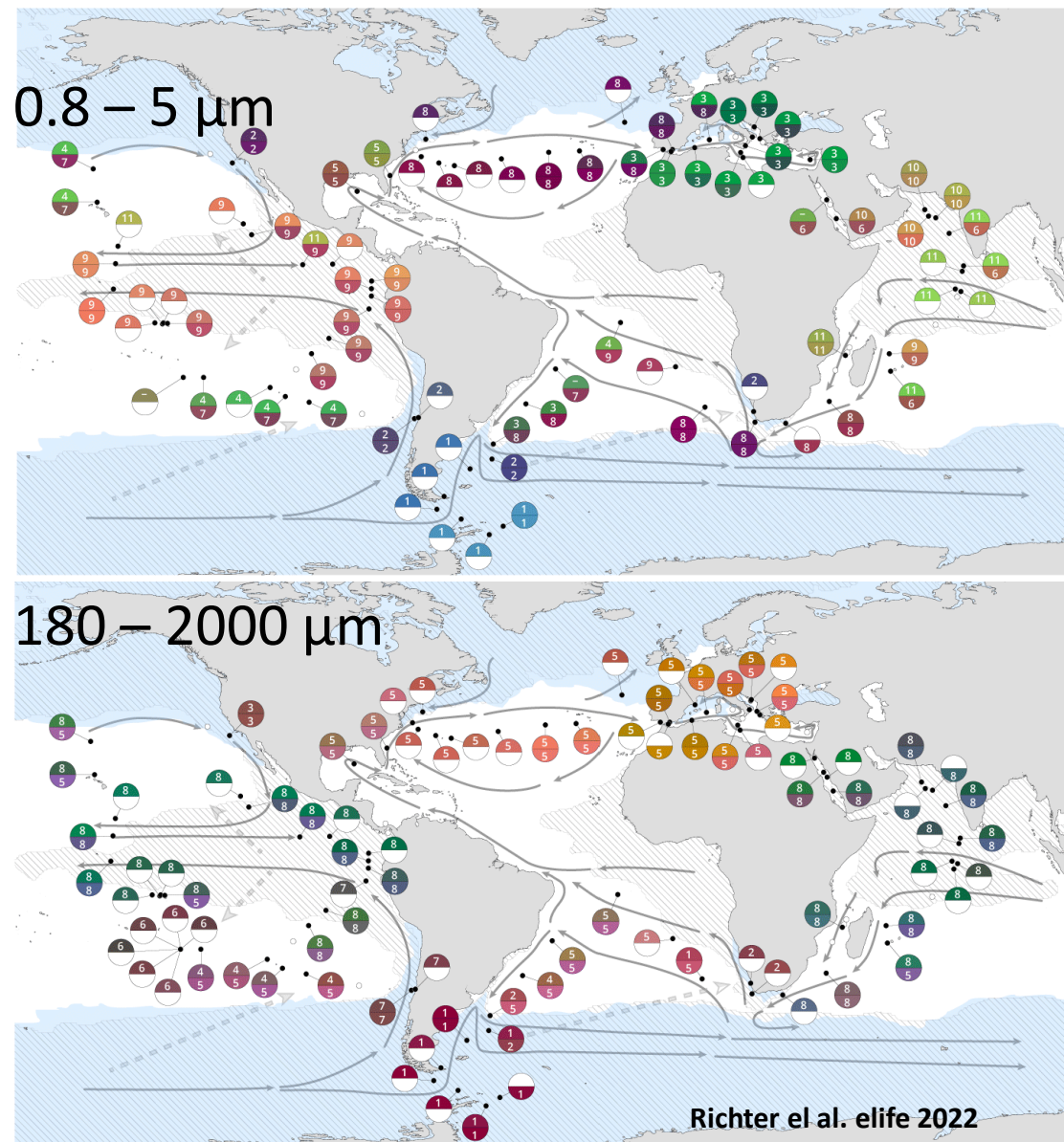
=> Display in a geographical map

A global genomic biogeography of plankton community

- Samples colored in RGB values using PCA 3 main axis component values => differences in colors reflect beta-diversity among samples

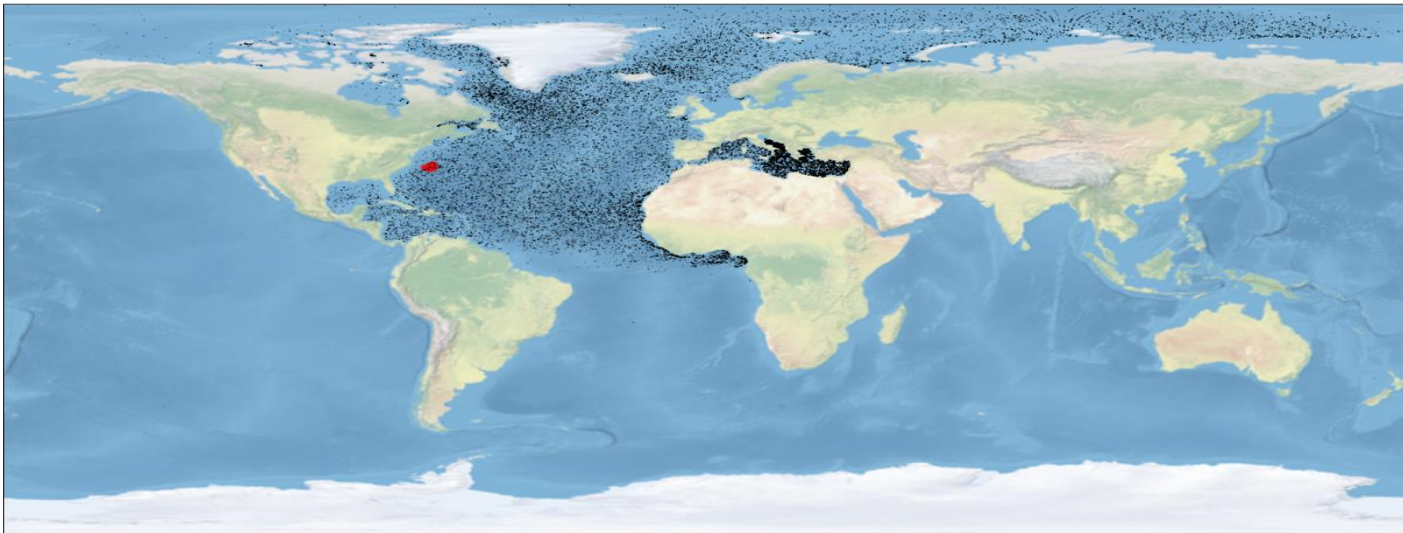


- Spatial organization in *genomic provinces*.
- Different scale of Spatial Distribution among organism size fractions.
 - Smaller size fraction, smaller provinces
 - Reflect different ecology, behavior (reproduction time).
 - Plasticity of biotic interactions ??



Lagrangian distances

- Could we correlate dissimilarity of community composition between 2 stations with transport time ?
- Rationale : Compute Minimum Transport Time (T_{min}) between pairs of samples.



Lagrangian distance (T_{min})

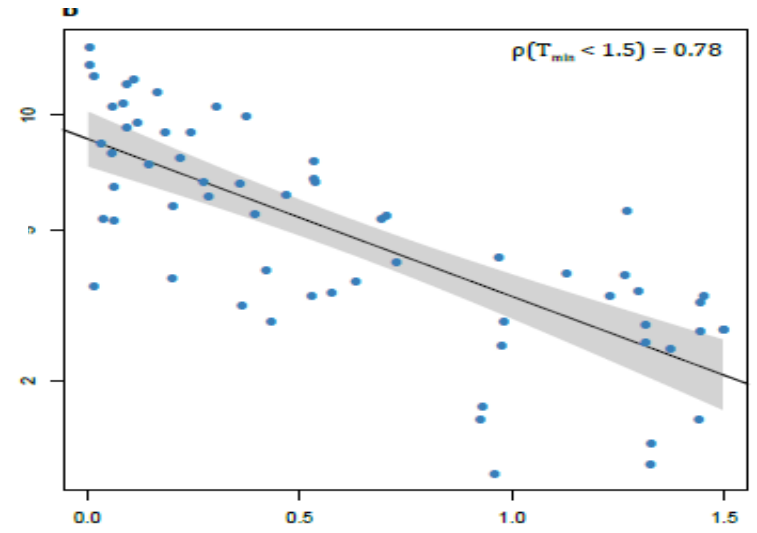


Metagenomic distance (Simka)

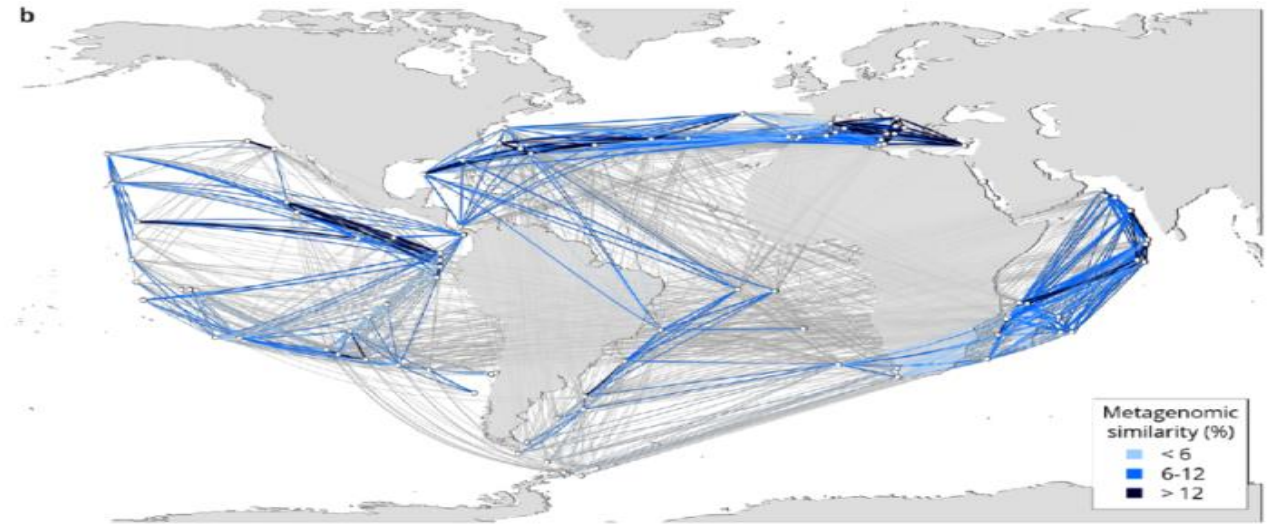
Influence of Transport Time : zoom in North Atlantic

0.8 – 5 μm

Metagenomic Similarity %

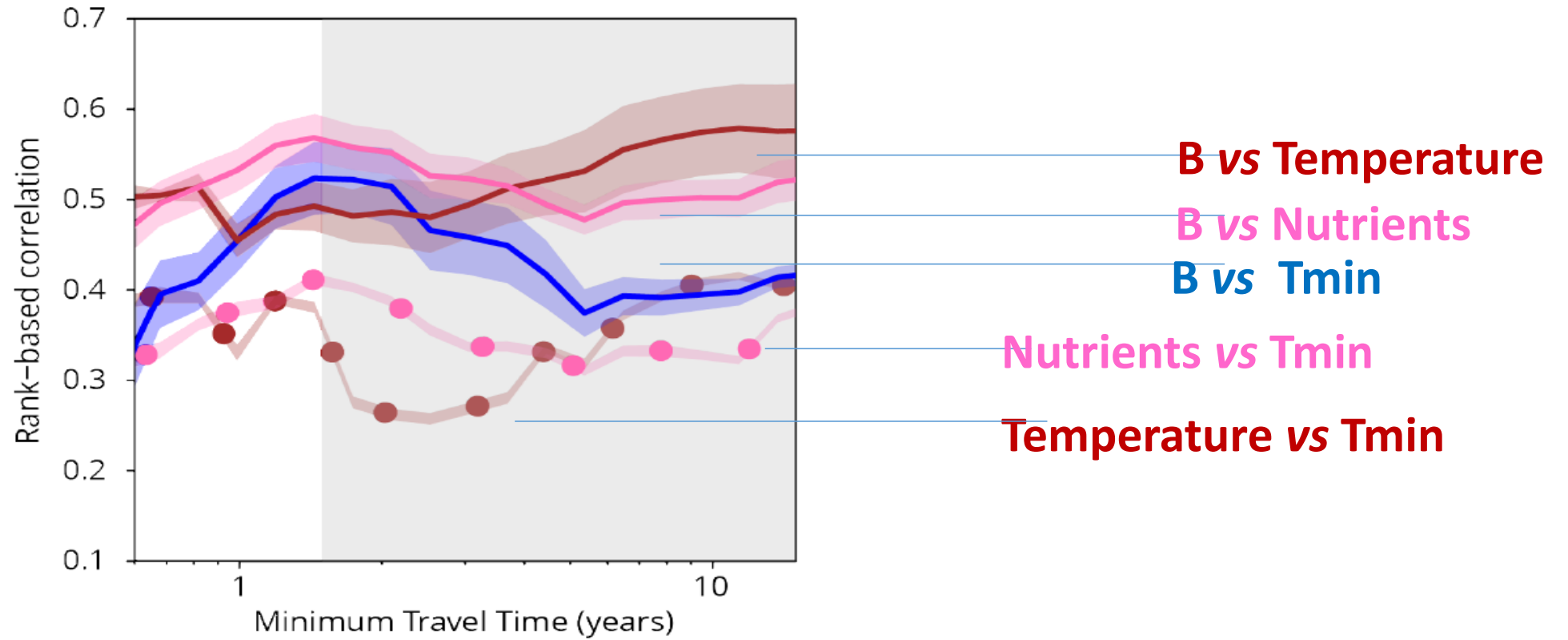


Travel Time (years)

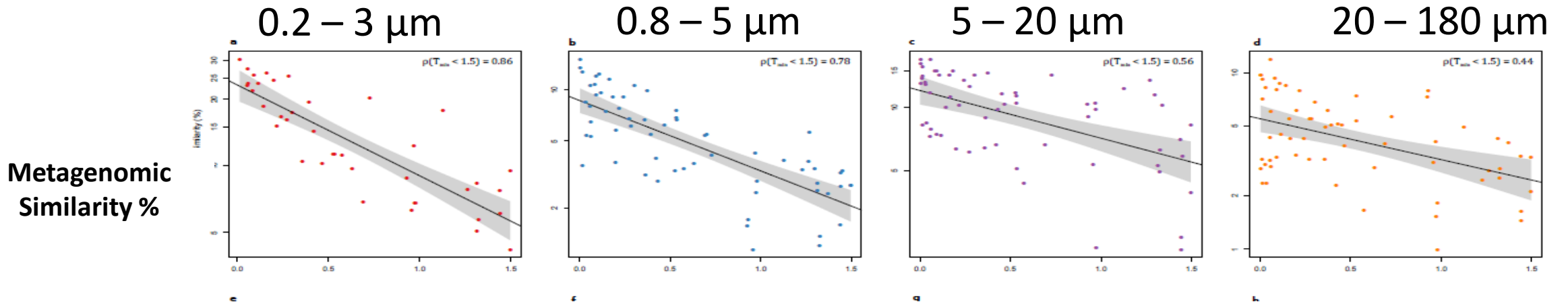


Metagenomic connectivity for pairs of stations < 1,5 year s

Measurable Influences of environmental drivers along plankton travel time

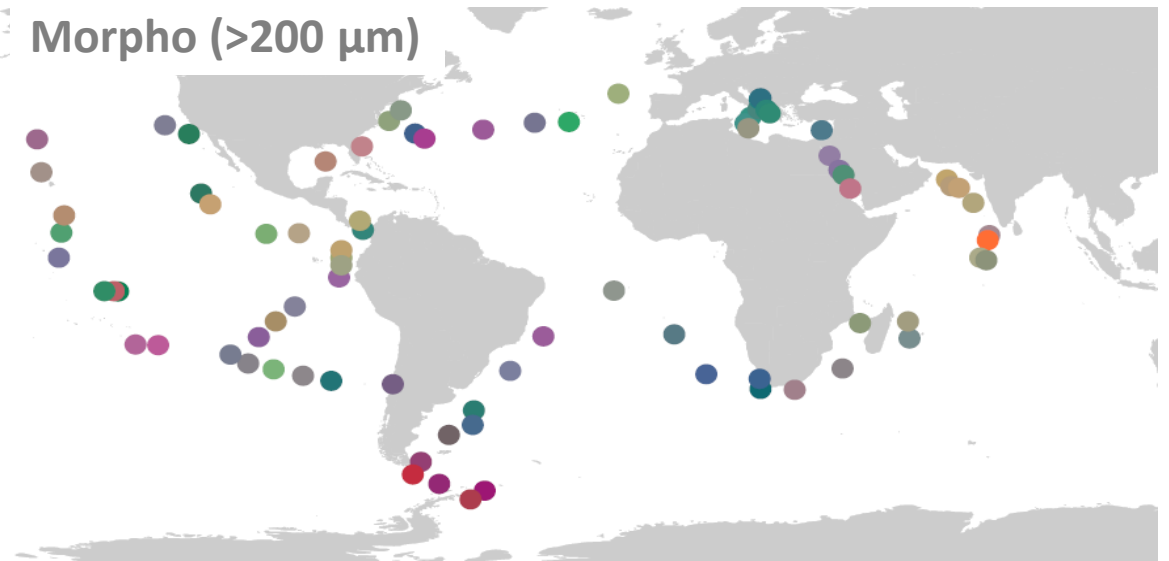
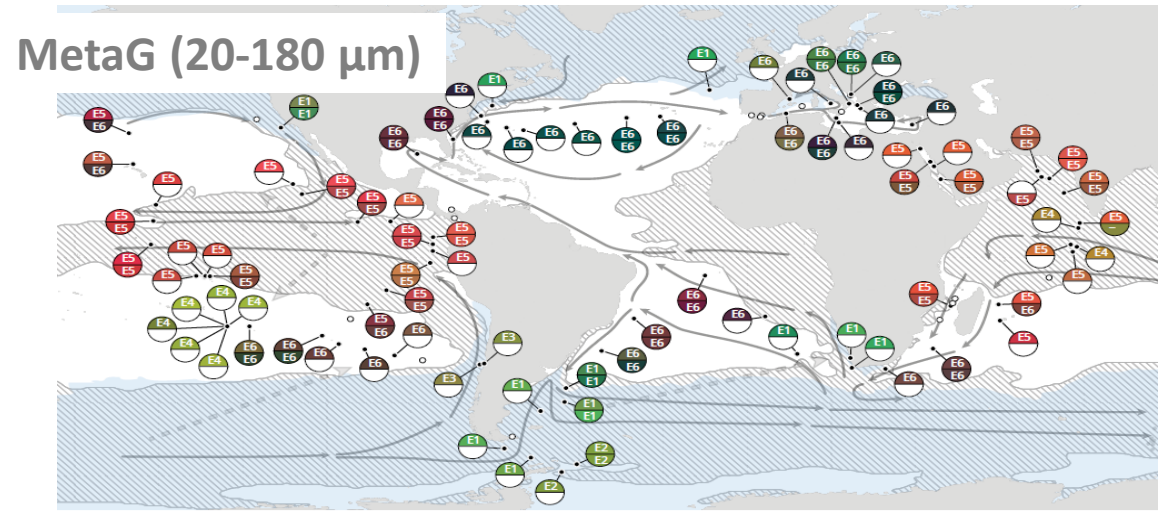
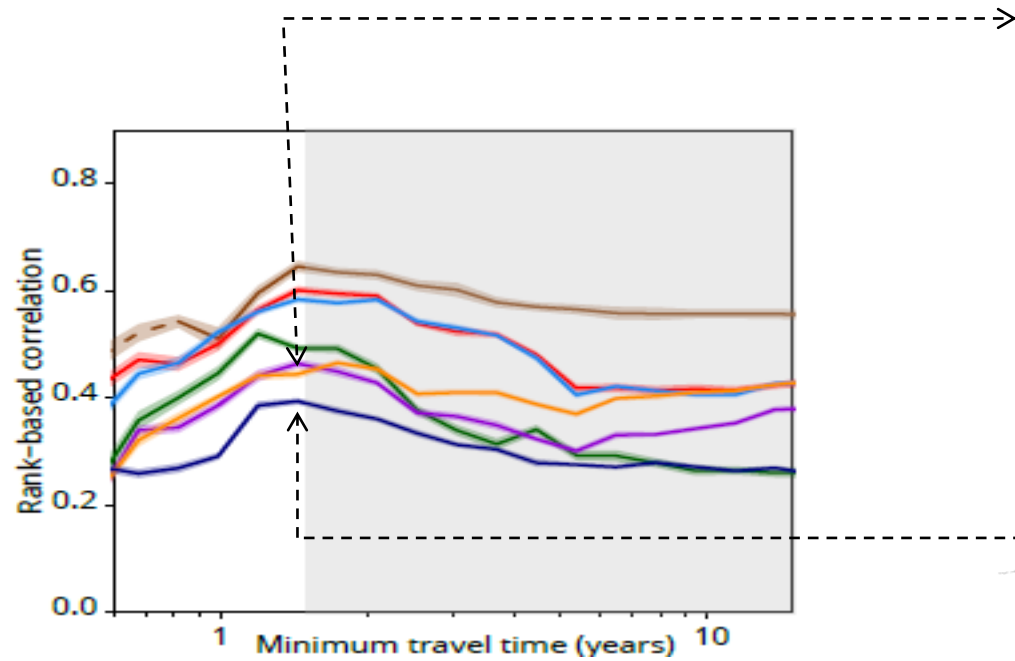


Influence of Transport Time : organism size dependent

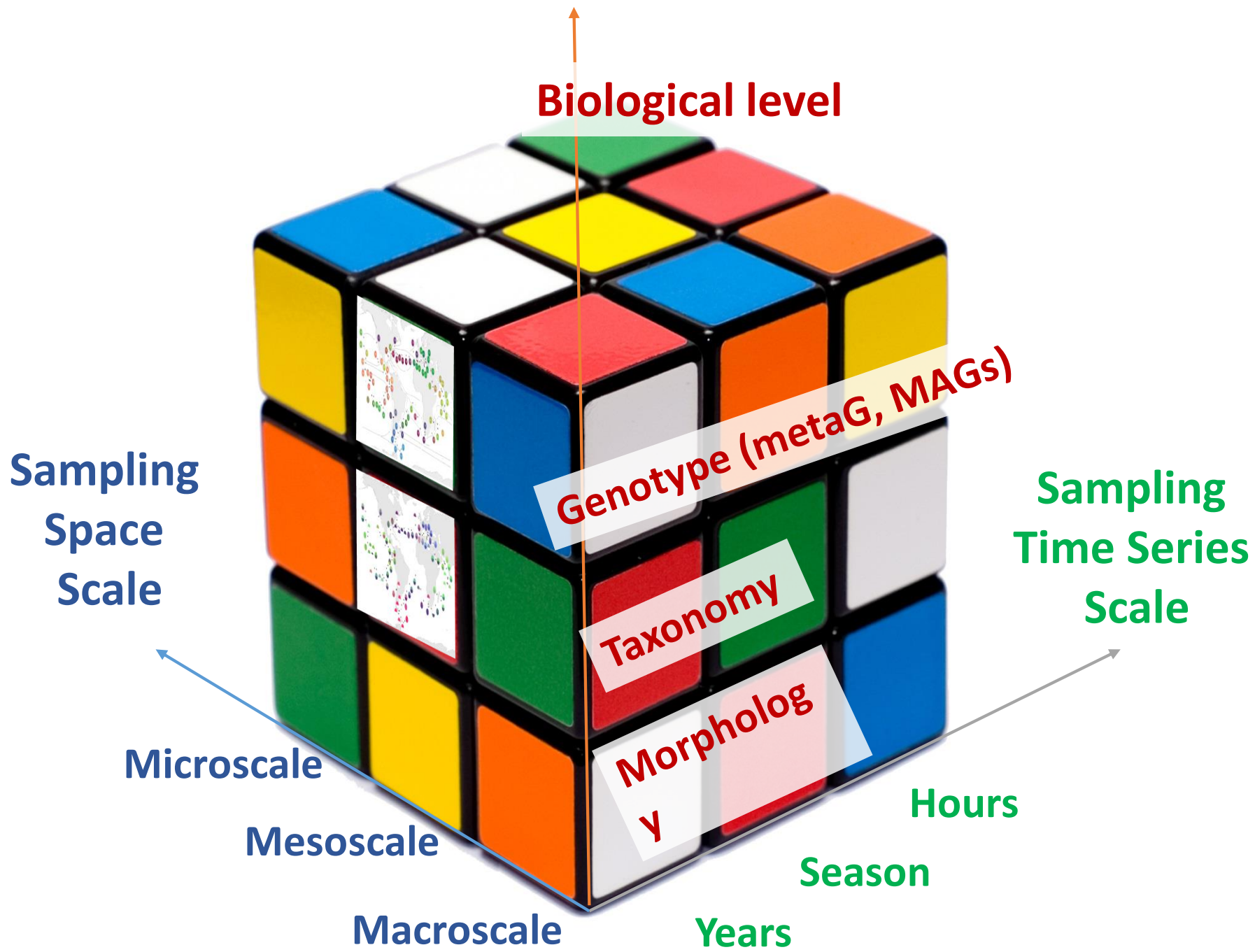


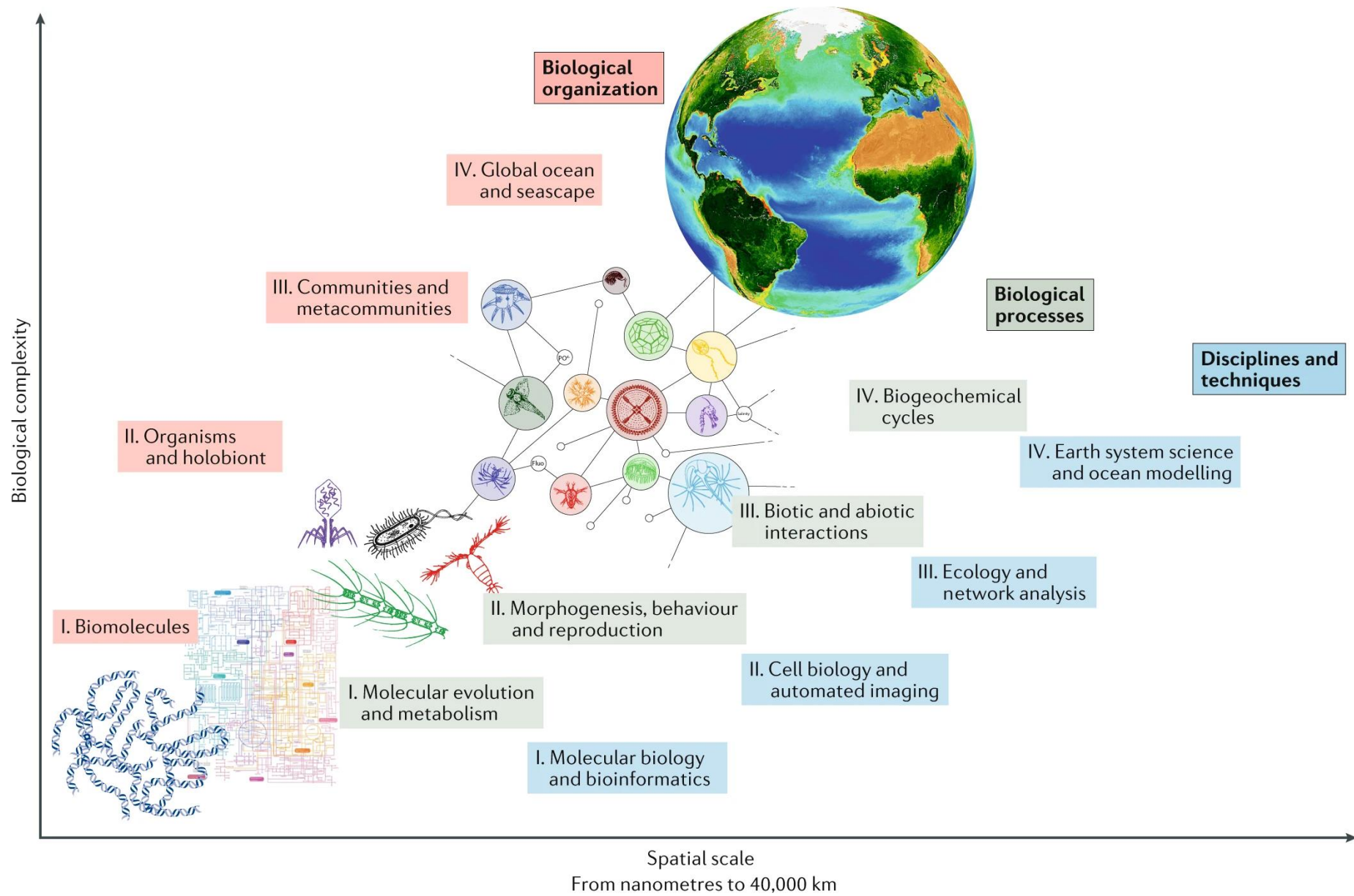
- The slope decreases with organism size
- => At genomic resolution, the turn over rate of populations along currents decreases with organism size

“Omics” resolution vs Morphology resolution



Morpho & MetaG are quite correlated together ($p=0.51$) but morpho is less correlated with transport time and provides a more patchy biogeography.

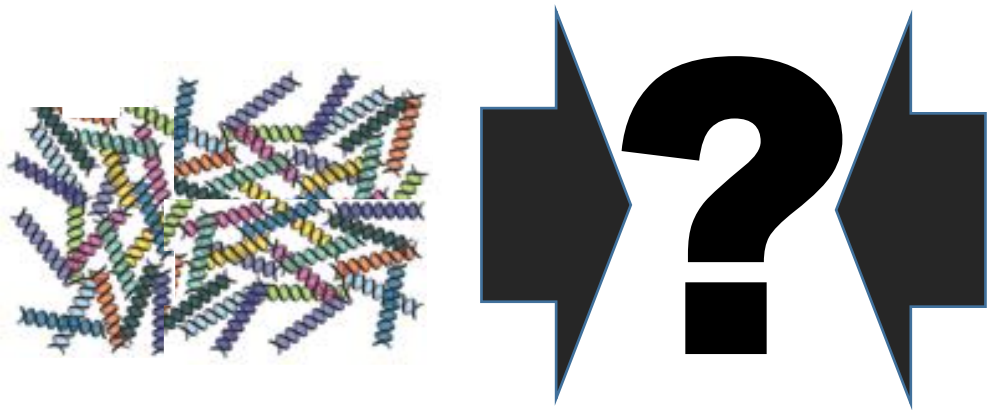




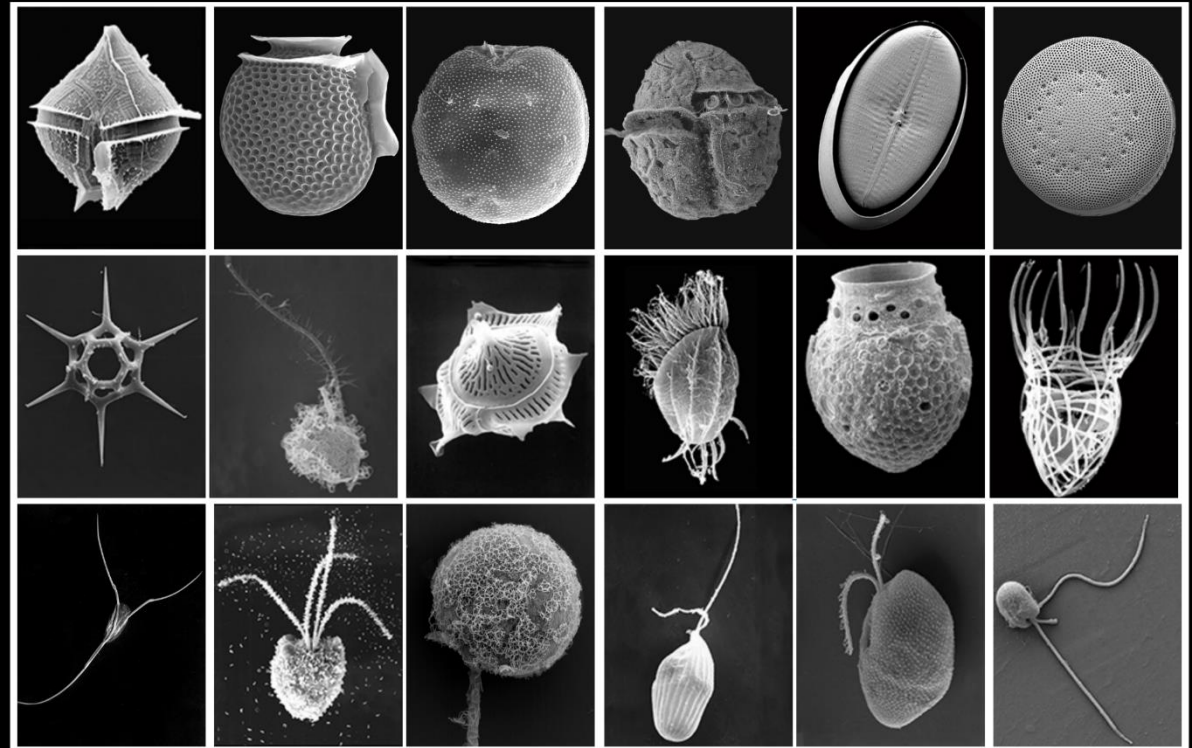
Sunagawa, S., Acinas, S.G., Bork, P. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* 18, 428–445 (2020).

From kmers to genomes :
Community level to genome level, and *vice versa*

A lot of fragments of DNA

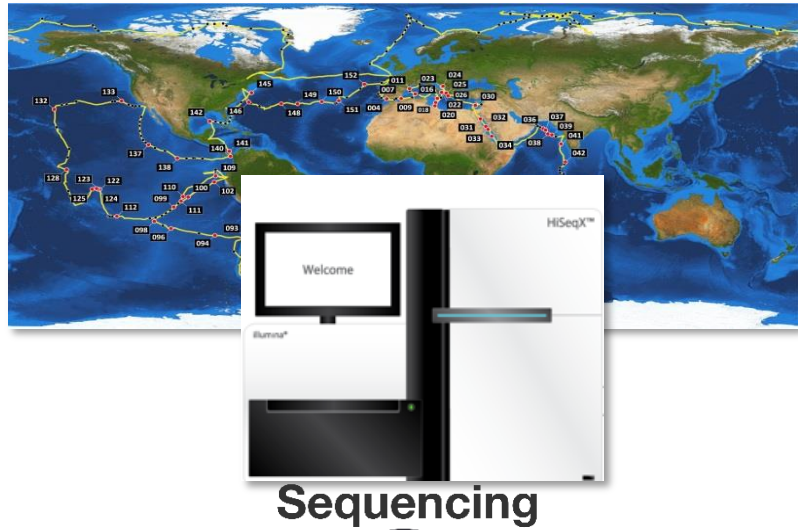


**A lot of beautiful images
but not that many genomes...**

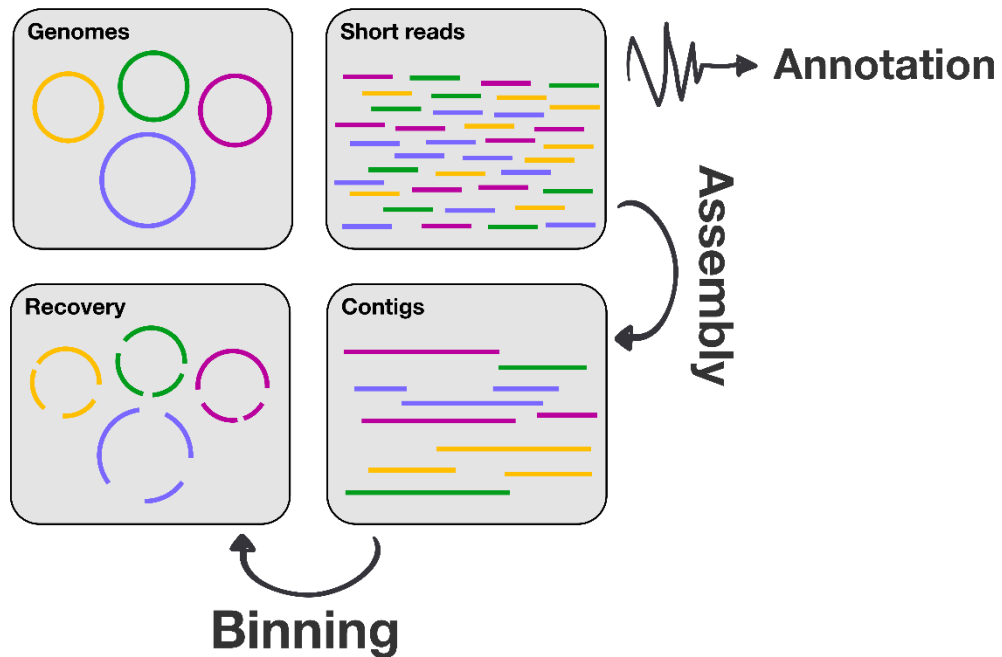


Source: <https://taxonomic.aad.gov.au/>

Reconstructing Genomes : Metagenome-based Assembled Genomes



- **Nearly 1,000 metagenomes**
- **280 billion metagenomic reads**
- **11 large co-assemblies by region**
- **2 million contigs > 2.5 kbp**



For each co-assembly:

- **Automatic binning** (2,550 bins in total)
- **Extensive manual curation**

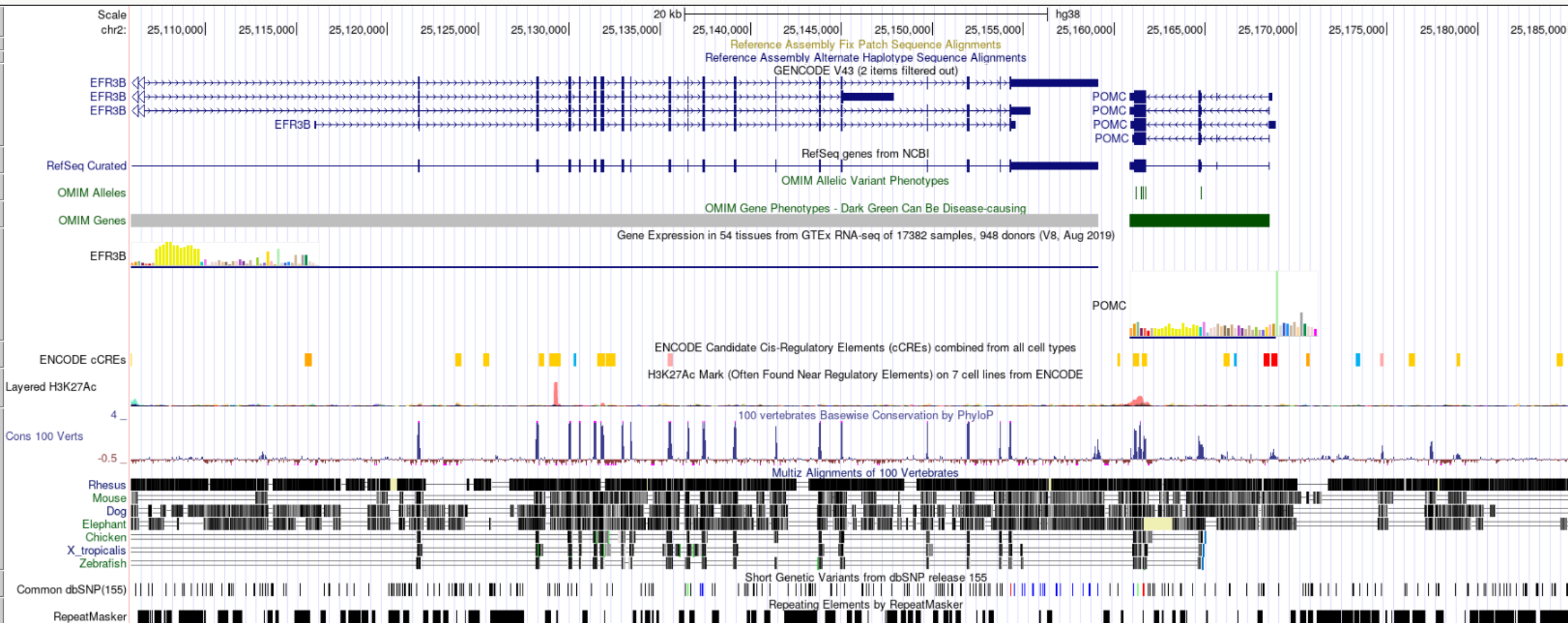
Reconstructing Exon – Intron Gene Structure

UCSC Genome Browser on Human (GRCh38/hg38)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

multi-region chr2:25,105,891-25,186,796 80,906 bp. gene, chromosome range, search terms, help pages, see exan go examples

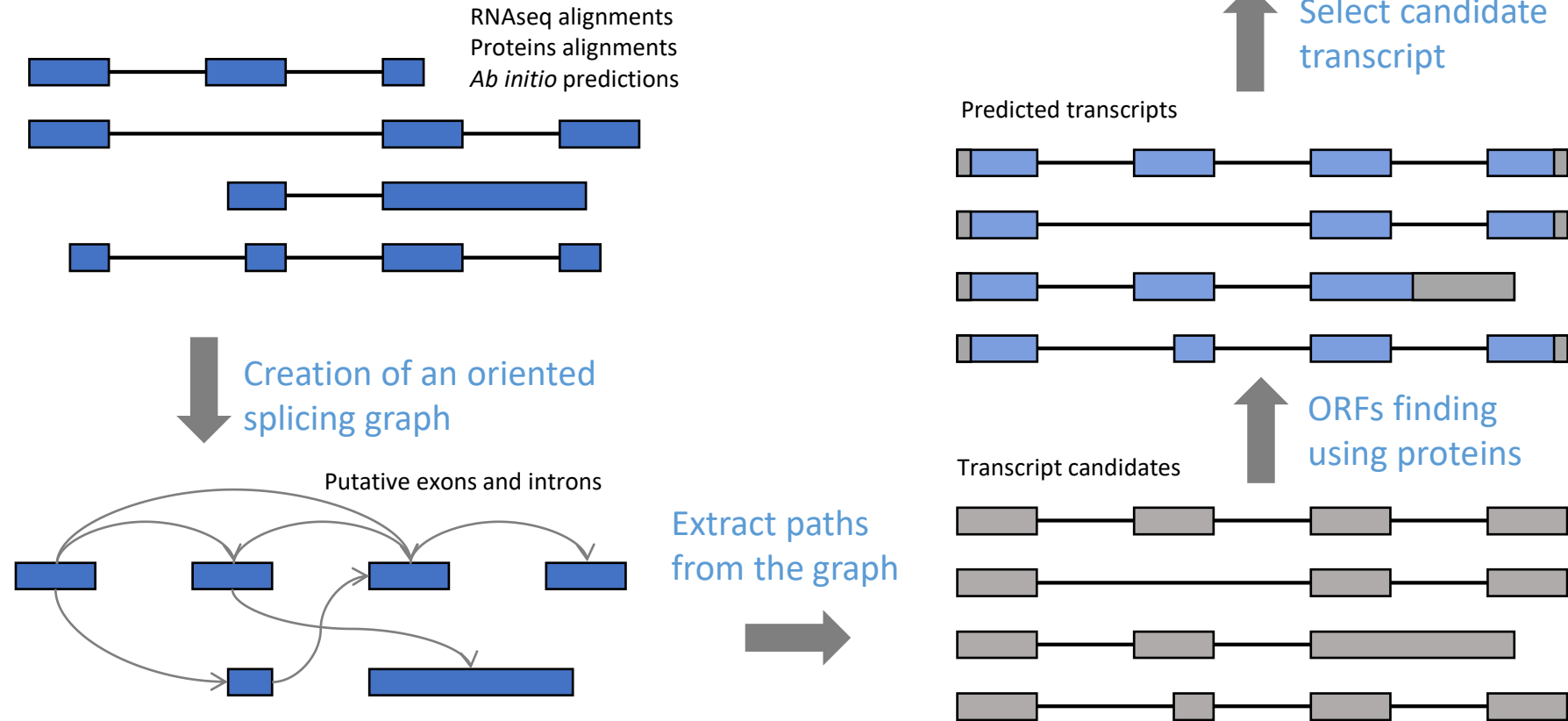
chr2 (p23.3) 25.3 25.1 24.3 24.1 22.3 2p21 16.3 p16.1 15 p14 2p12 2p11.2 q11.2 13 q14.1 2q14.3 22.1 22.3 23.3 24.1 q24.3 2q31.1 q32.1 q32.3 q33.1 2q34 2q35 36.3 37.1 q37.3



Combining data to predict genes structures

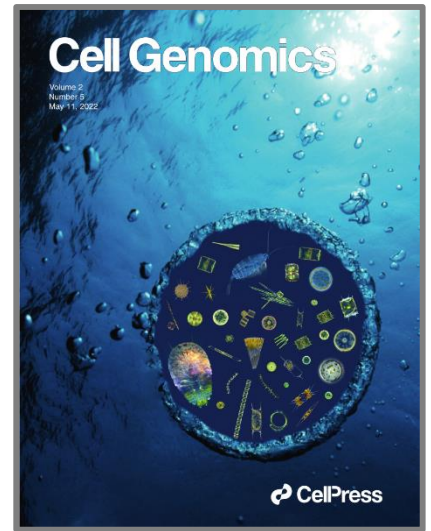
Gmove – Gene Modelling using Various Evidence

<https://github.com/institut-de-genomique/Gmove>

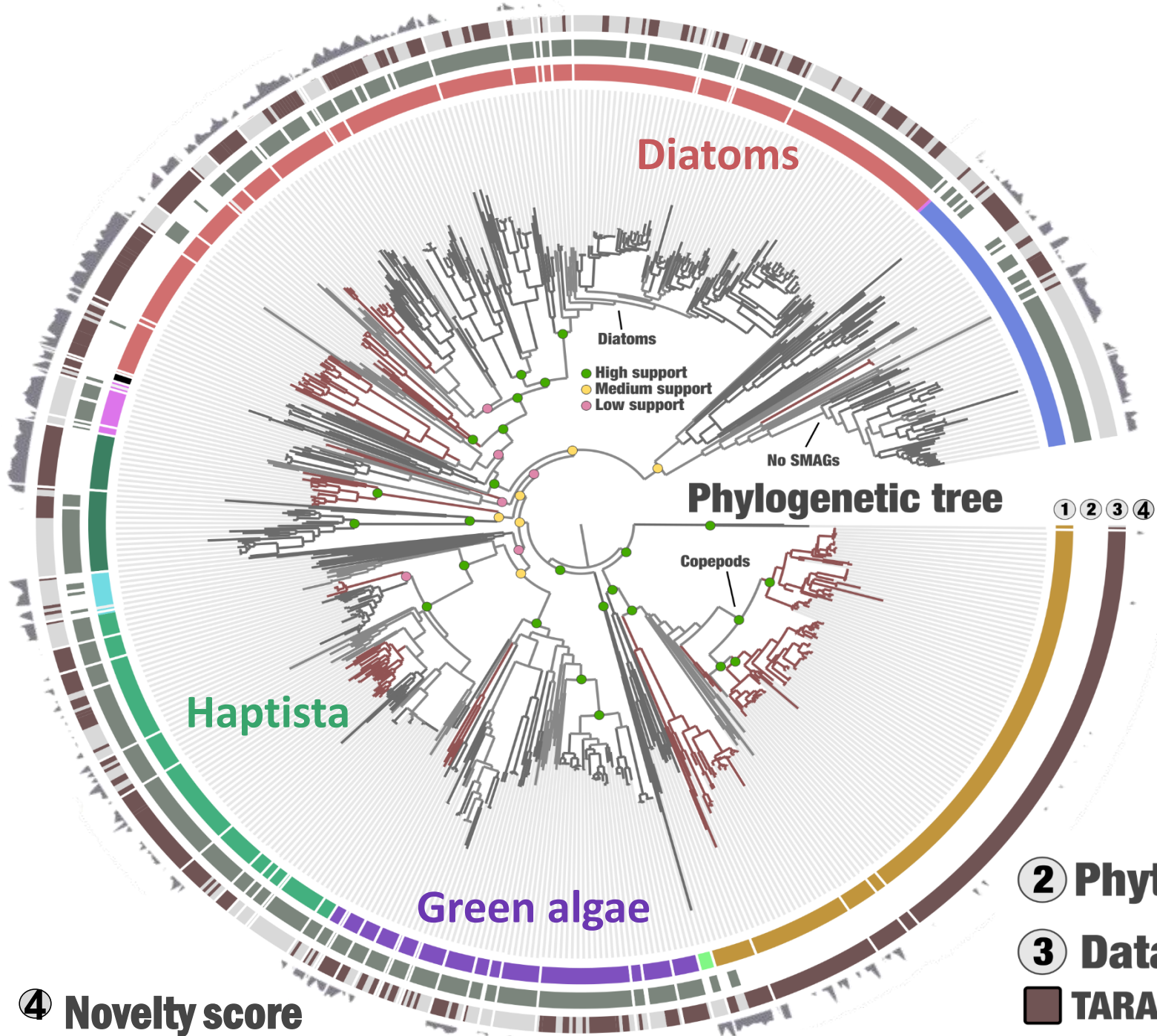


~ 700 eukaryotic genomes

- **Average completion ~40% (up to 93.7%)**
- **Average genomic length 35.4 Mbp (up to 1.32 Gbp)**



Delmont et al. Cell Genomics 2022



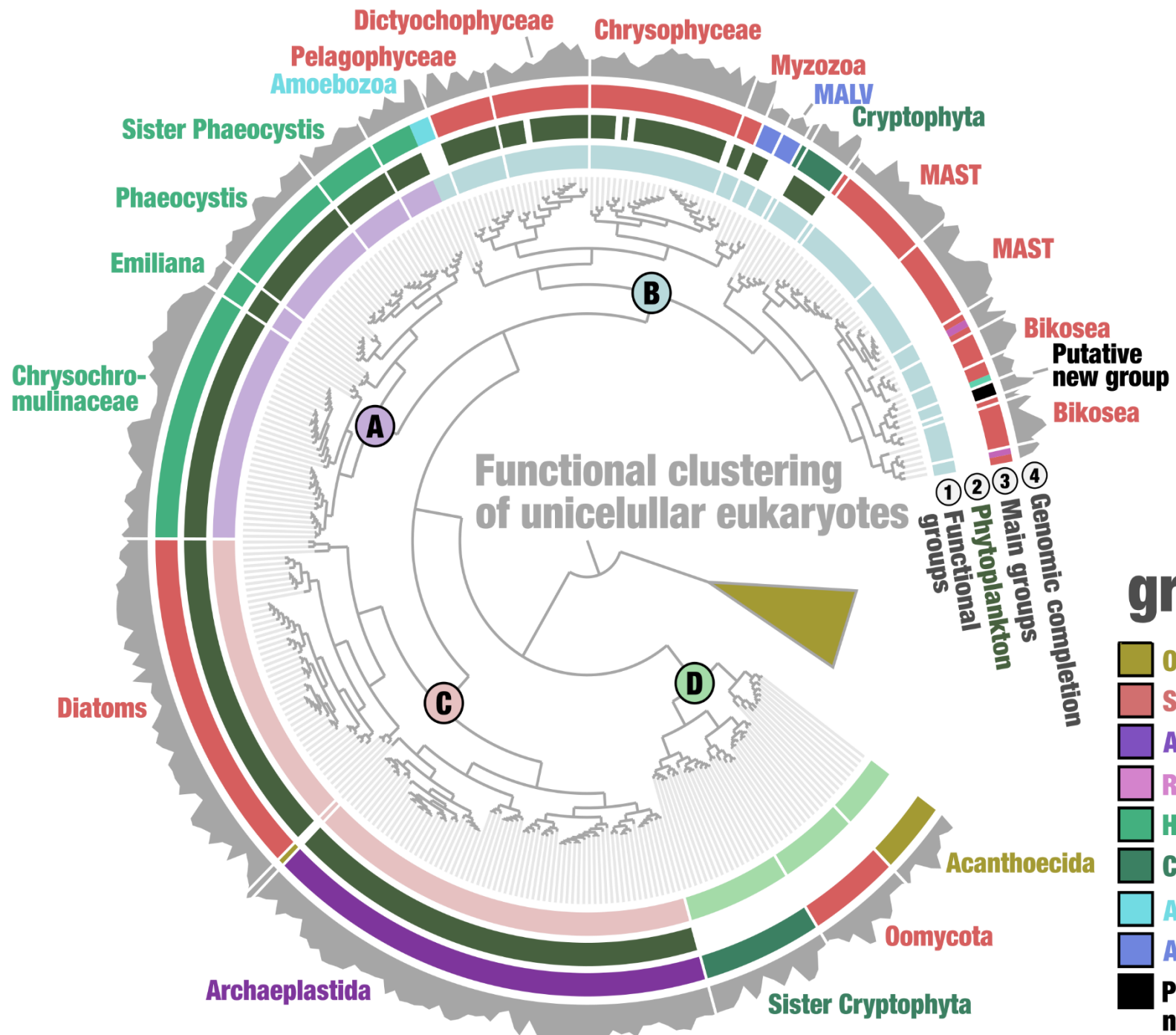
- ① Main groups
- Opisthokonta
- Stramenopiles
- Archaeplastida
- Rhizaria
- Haptista
- Cryptista
- Amoebozoa
- Alveolata
- Excavata
- Putative new group

② Phytoplankton

③ Database:

- TARA SMAGs
- METdb

④ Novelty score

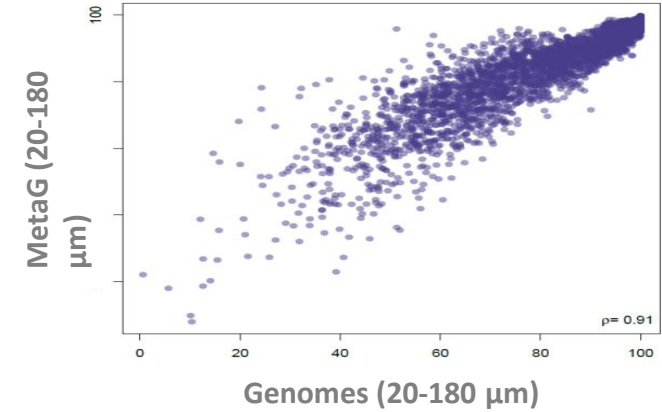
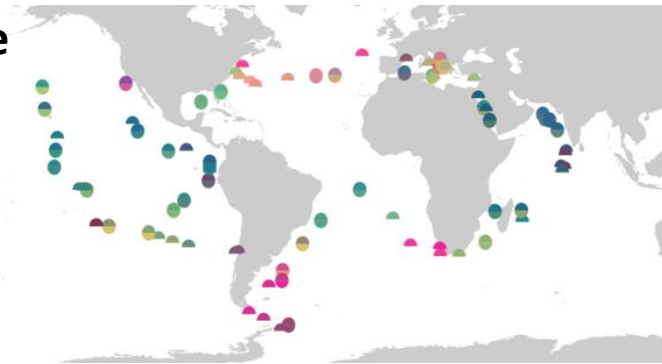
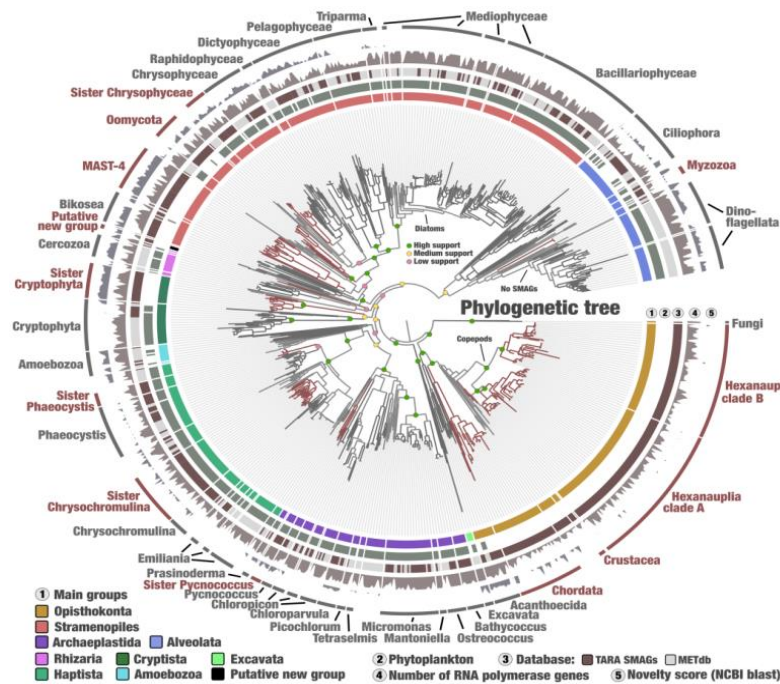


Main groups

- Opisthokonta
- Stramenopiles
- Archaeplastida
- Rhizaria
- Haptista
- Cryptista
- Amoebozoa
- Alveolata
- Putative new group

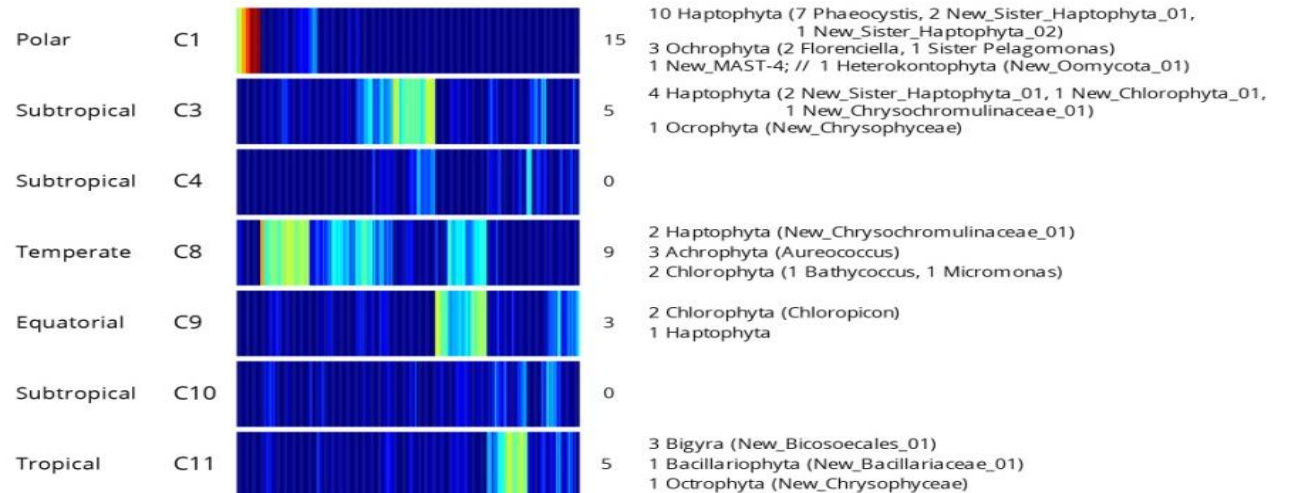
From kmers to genomes : Community level to genome level, and *vice versa*

A collection of ~700 genomes (MAGs) of eukaryote plankton



Genomes "signatures" of provinces

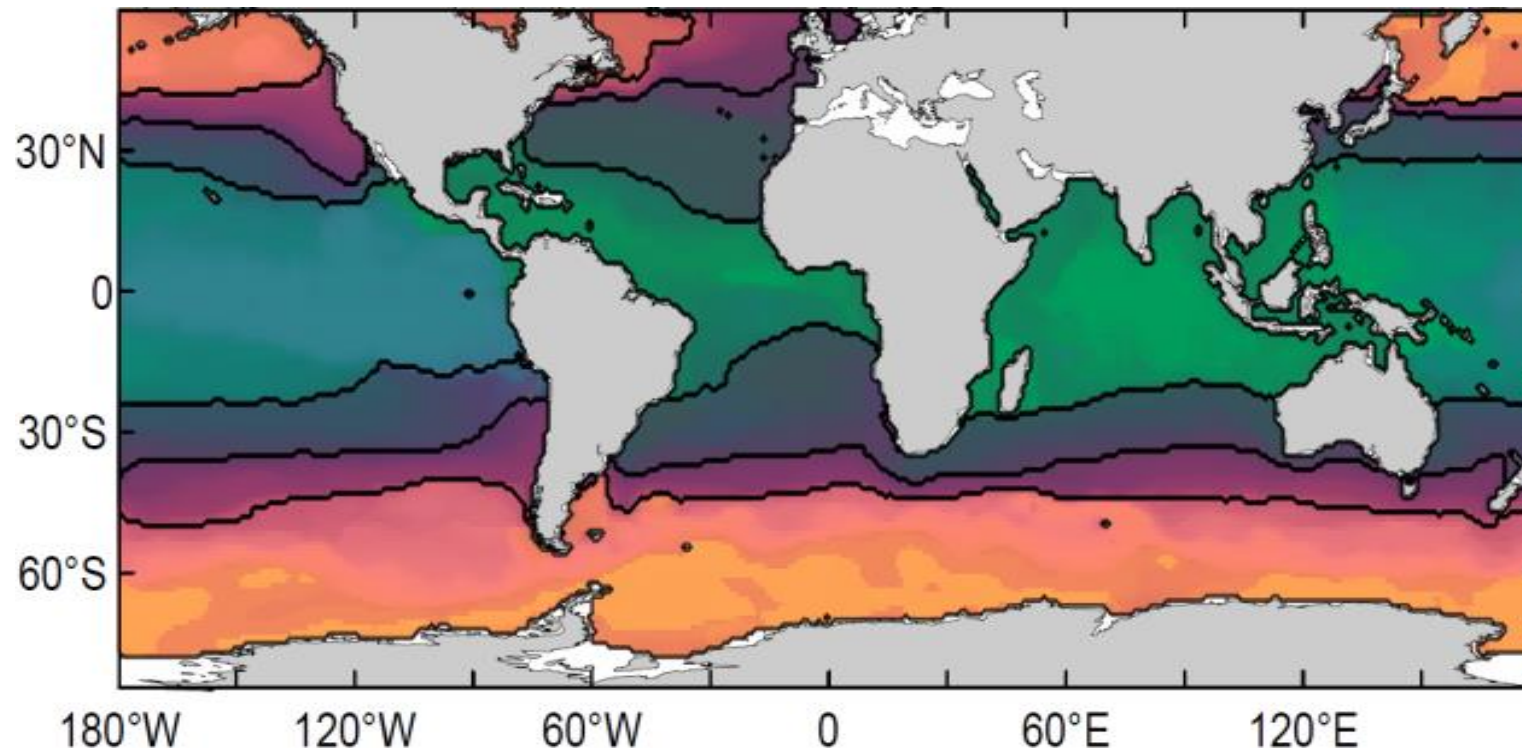
Plankton size 0.8-5 μm



~700 MAGs occurrences

Going Further with Modeling approach (Paul's Talk Teasing)

- Can we extrapolate to :
 - Whole Oceans Biogeography & Merge Plankton size fractions ?
 - End of Century Modification in Biogeography & Biogeochemical Functions
- => ML / Niche Modeling + Climate Models + Carbon Export Models



Patrick Wincker

**Tom Delmont
Morgan Gaia
Eric Pelletier
Quentin Carradec**



Julie Poulain

Jean-Marc Aury

...

**Ian Probert (Roscoff)
Mahendra Mariadassou (Inra)
Pierre Peterlongo (Inria)**

Fabien Lombard (CNRS Villefranche)

Mathieu Vrac (CEA LSCE)

**Daniel
Richter
(Roscoff)**

**Thomas
Vannier
(Genoscope)**

**Daniele
Iudicone
(SZN)**

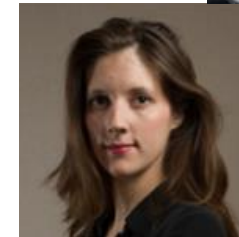
**Romain
Watteaux
(SZN)**



**Paul Frémont
(Genoscope)**

**Marion
Gehlen
(CEA LSCE)**

**Colomban;
De Vargas
(Roscoff))**



**Jade
Leconte
(Genoscope)**



All Tara Oceans Coordinators and actors