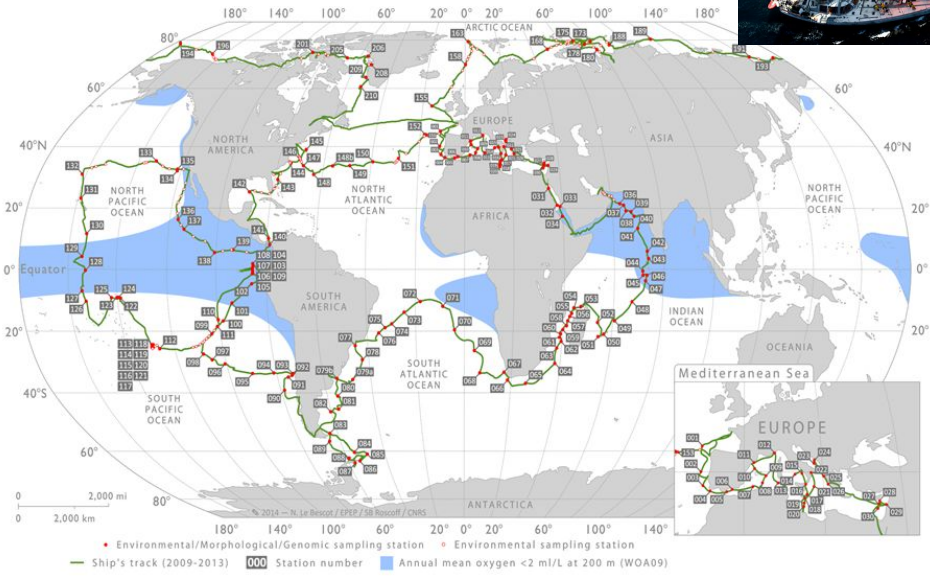
Data intensive research from genes to ecosystems in changing oceans

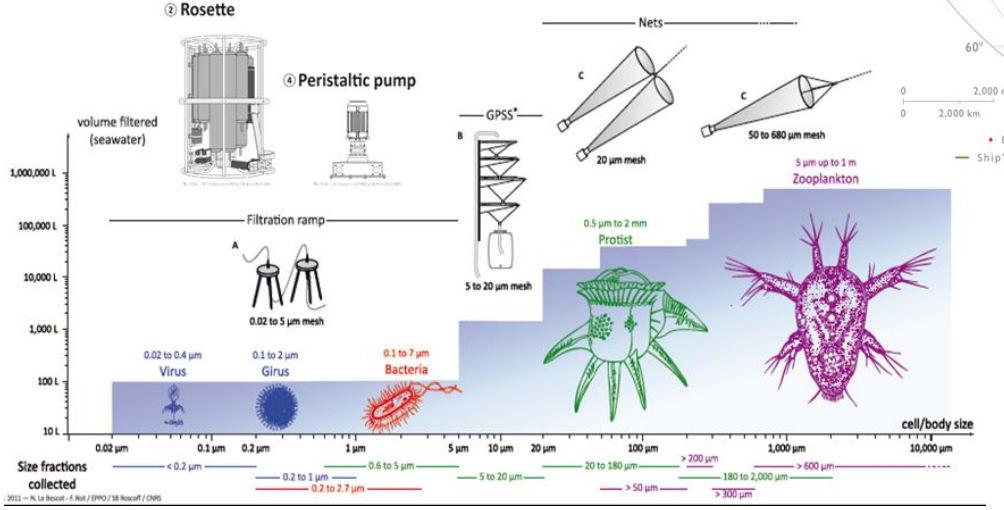
Damien Eveillard
Stéphane Pesant & Guy Cochrane
Tara Oceans Consortium

EU Pilot Blue Cloud consultation, Brussels, 2017.03.28

Use Case: Tara Oceans Expedition 2009-2013

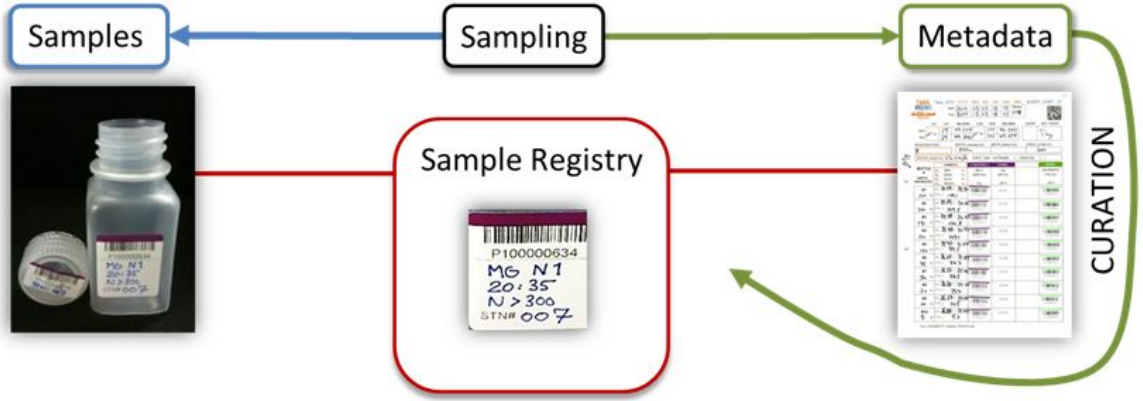
- Global study of Plankton
- 210 stations in 20 biogeo. provinces
- State of the art oceanographic equipment
- Consistent sampling methods



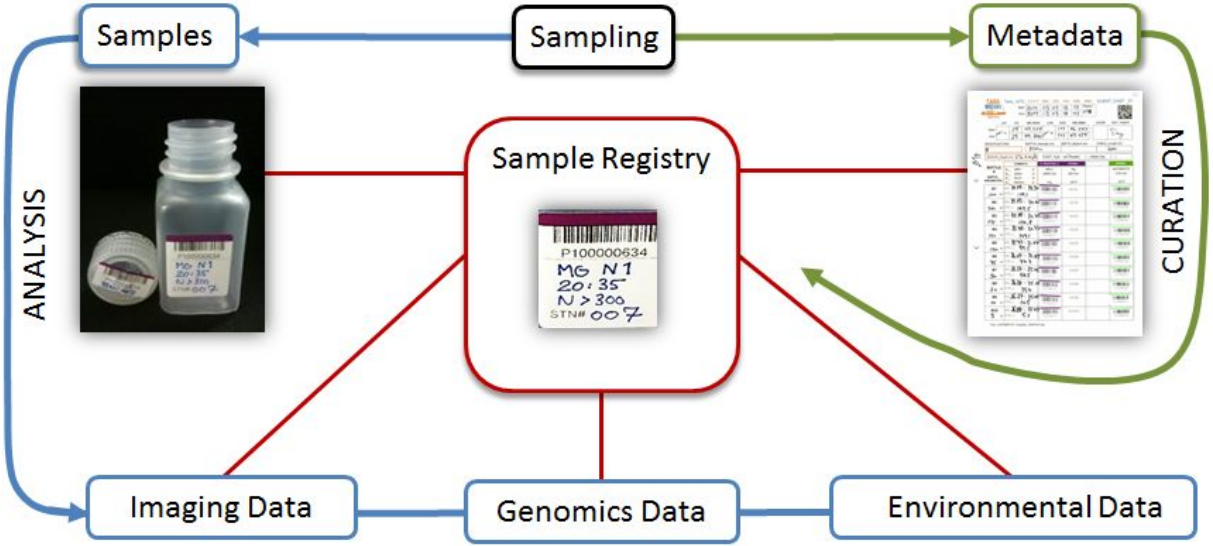


- State of the art plankton protocols
- across size spectra (0.02 µm - 2 m)
- across taxonomic spectra (virus to jellyfish)

Context / Provenance is key to data integration



Context / Provenance is key to data integration



+ oceano. sensors

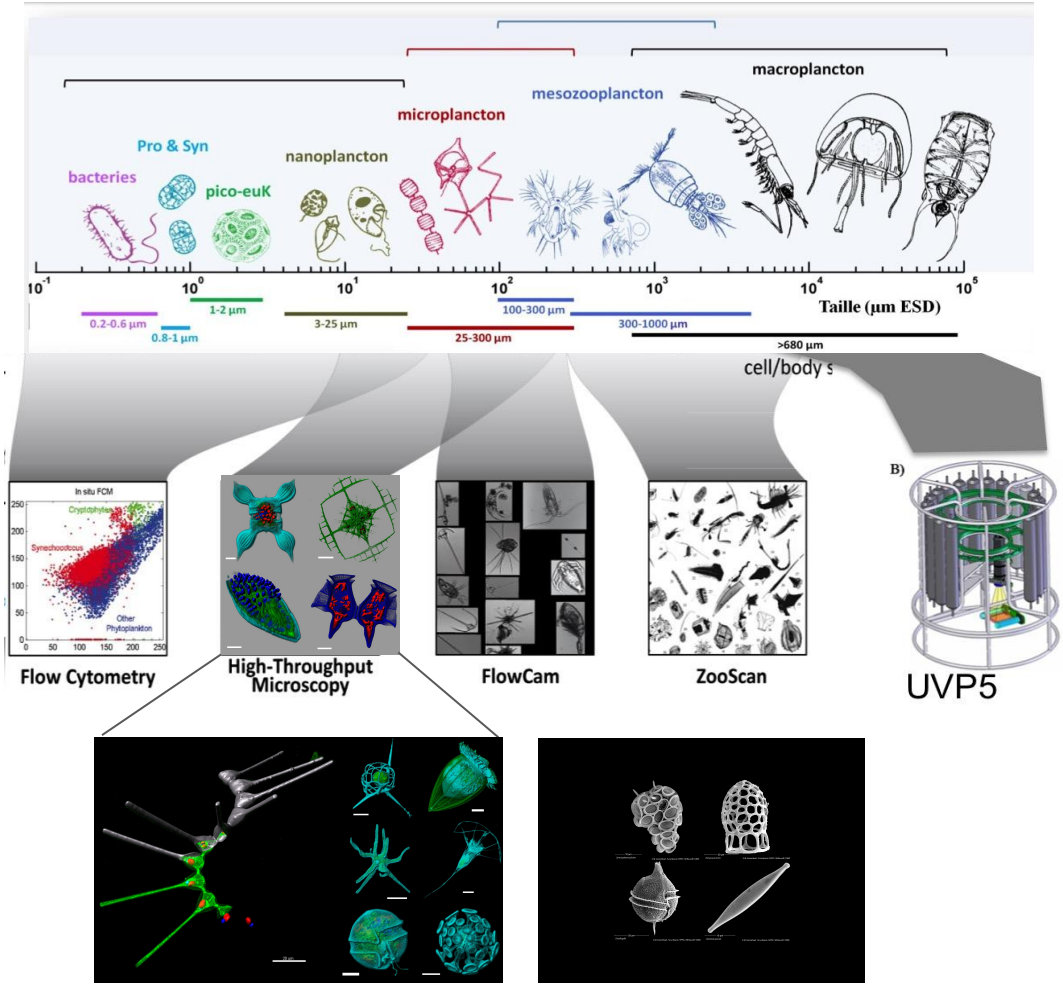
+ climatologies

+ satellite products

#Samples:	17,000	9,000	5,000 (biogeochemistry)

Imaging data

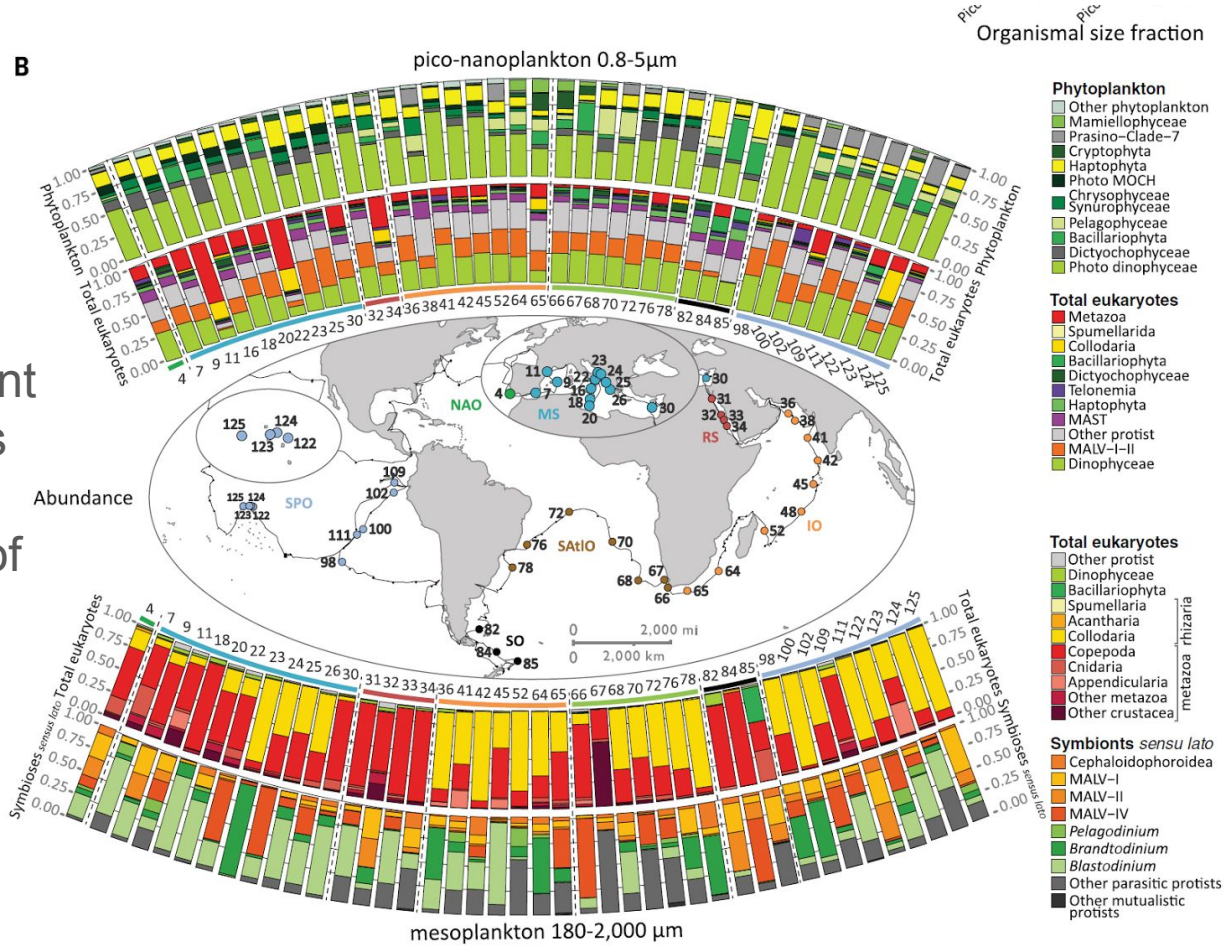
- **>4 million Images**
- **>30 terabytes**
- State of the art automated methods adapted to organism from 0.1µm to 0.1m
- Semi-automatic recognition -- Machine learning (supervised)
- Network approach for data sharing, expert annotation and training



From BioMed Technologies to environmental imaging

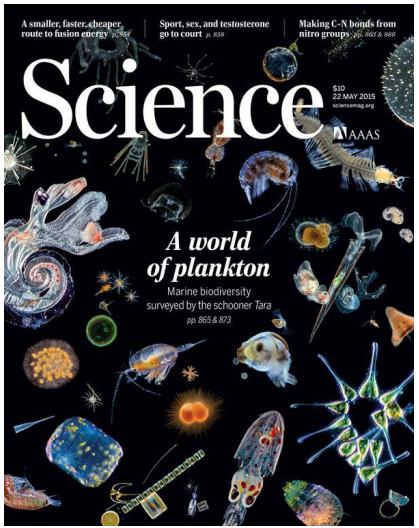
Genomics data

- **40 million genes** -- Largest-ever genomics effort in marine science
- **12 terabytes** -- equivalent to 135x human genomes
- **>7000 data sets** - One of the richest molecular collection in the public domain



[De Vargas *et. al. (2015)*]

Cutting edge & innovative Science

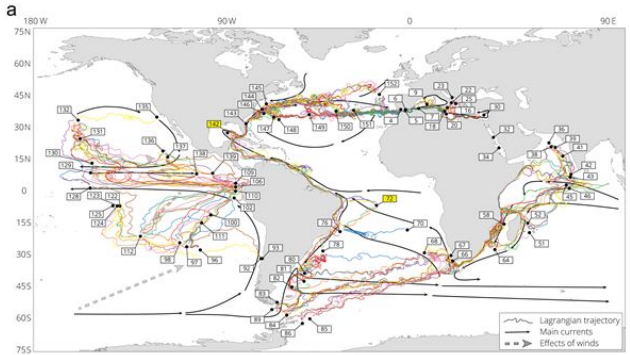


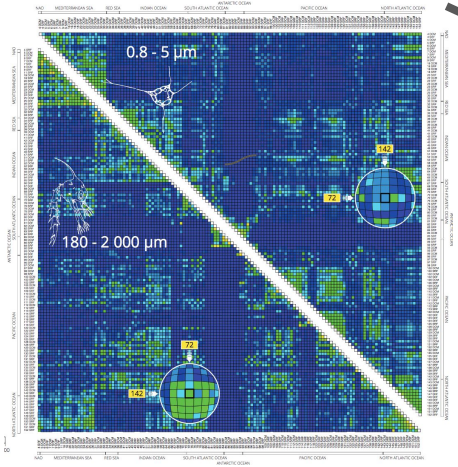
- Plankton at **planetary scale**
- Eukaryotic plankton **diversity**
- Structure and function of the ocean **microbiome**
- Patterns and **ecological drivers** of ocean viruses
- Plankton community structure & “**Interactome**”
- Plankton **transport across the Atlantic**

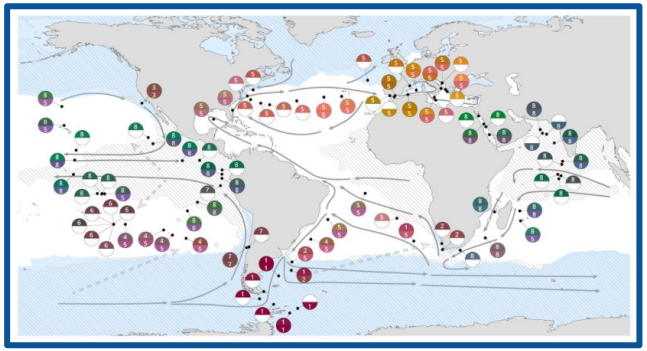
Global Modelling - New approach to Biogeography

[Richter *et. al.*, in revision]

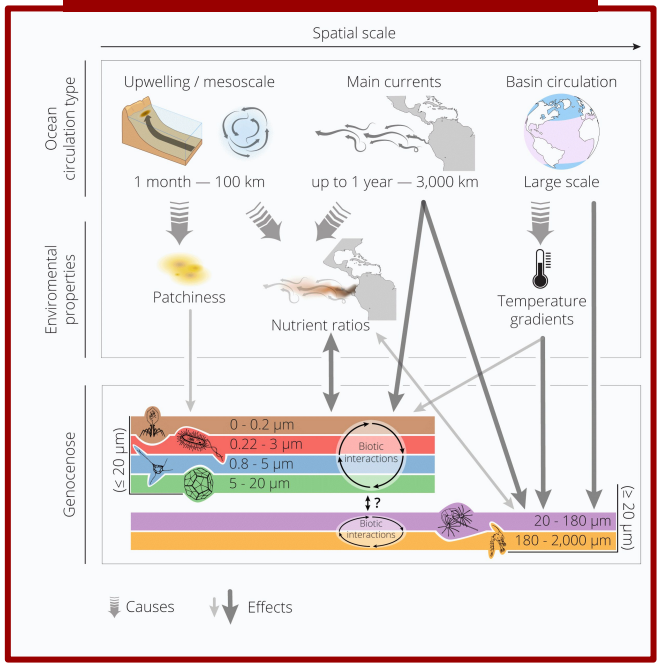
Combining Lagrangian and genomic distances







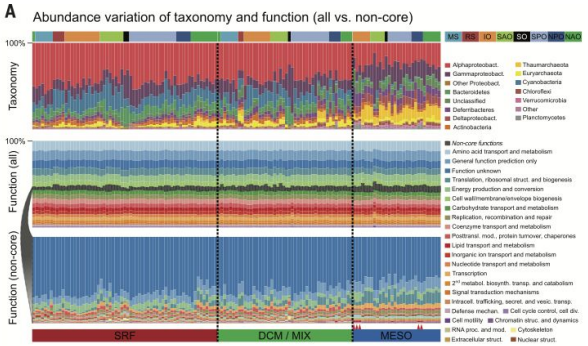
New distribution theory for each size fraction

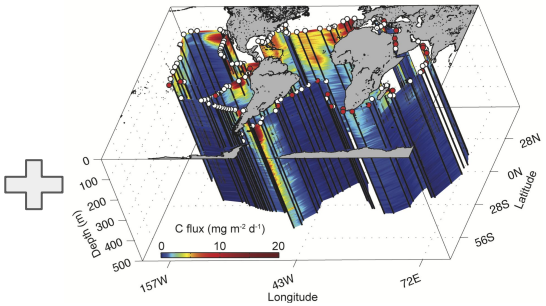


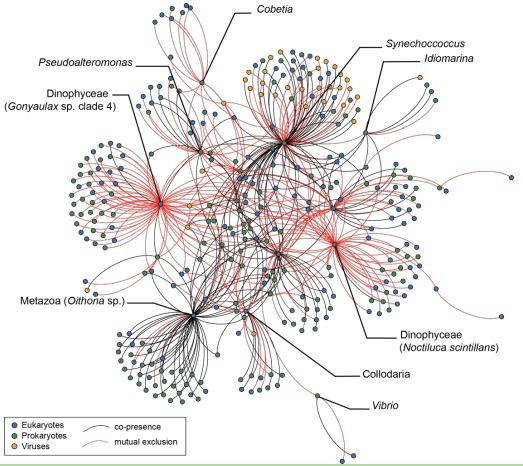
Global Modelling - New approach to Biogeochemistry

[Guidi et. al. Nature 2016]

Predicting Carbon export based on Environmental, Imaging and Genomic Data







Metagenomic knowledge

Physical measurements

Key stone species as a community and marker genes

Few Species, or 51 genes predict quantitatively carbon export with 90% of accuracy

Toward a semi Automatic Computational Pipeline and design of Geochips

FUTURE PERSPECTIVE

Enabling grass-root innovation like Tara Oceans via existing European data infrastructures & services

Needs for a Blue Cloud