ADAPTATION ET DIVERSITE EN MILIEU MARIN - AD2M

CNRS • SORBONNE UNIVERSITÉ

Station Biologique de Roscoff

Fabrice Not – not@sb-roscoff.fr

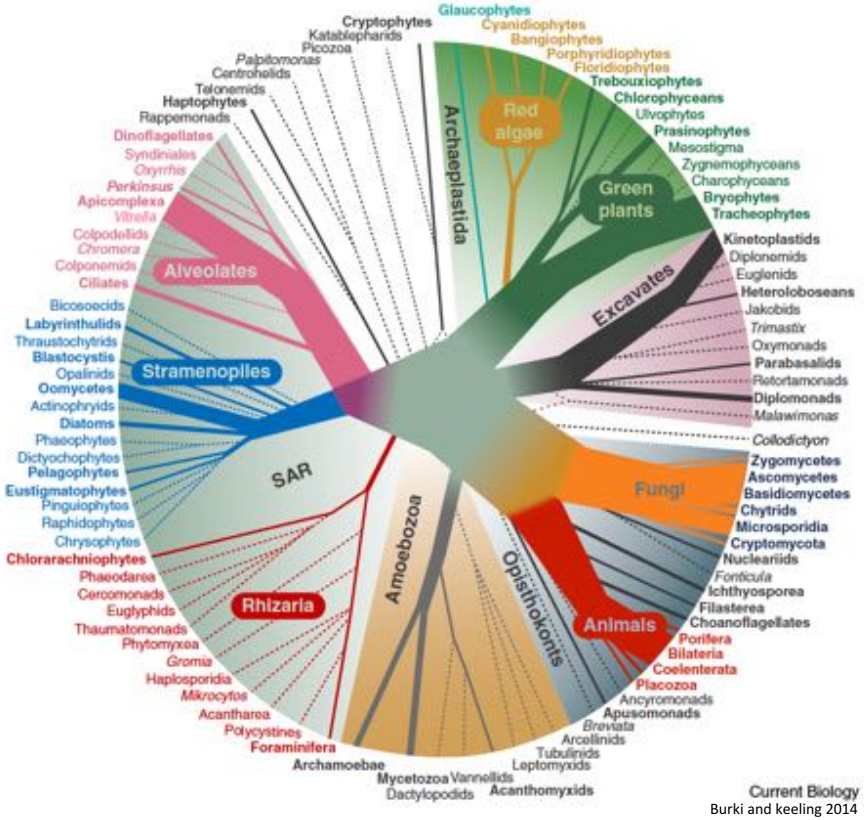
ADAPTATION ET DIVERSITE EN MILIEU MARIN

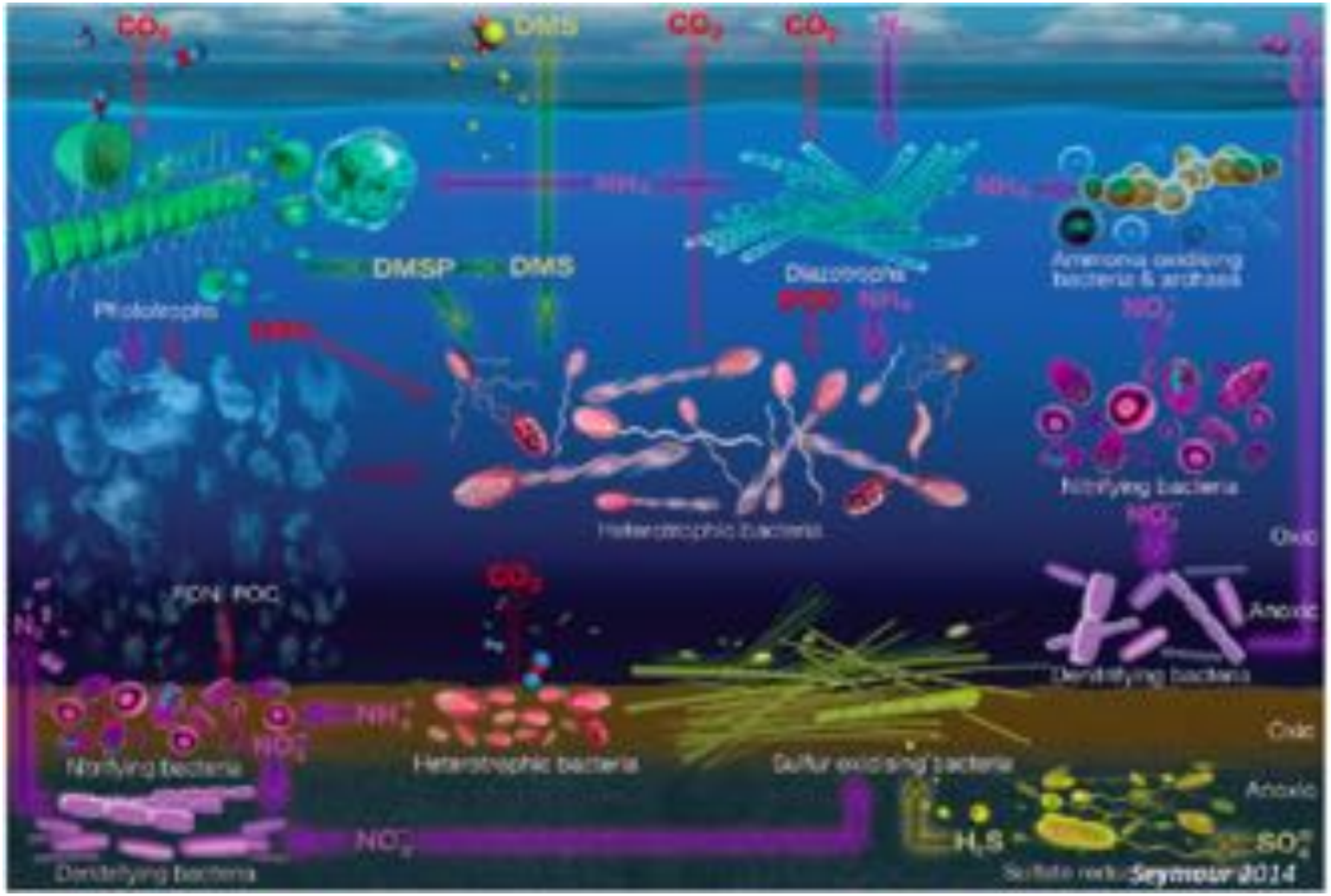
Across broad scale range

- Spatio-temporal coverage
- Molecules, cells, …. up to Ecosystems

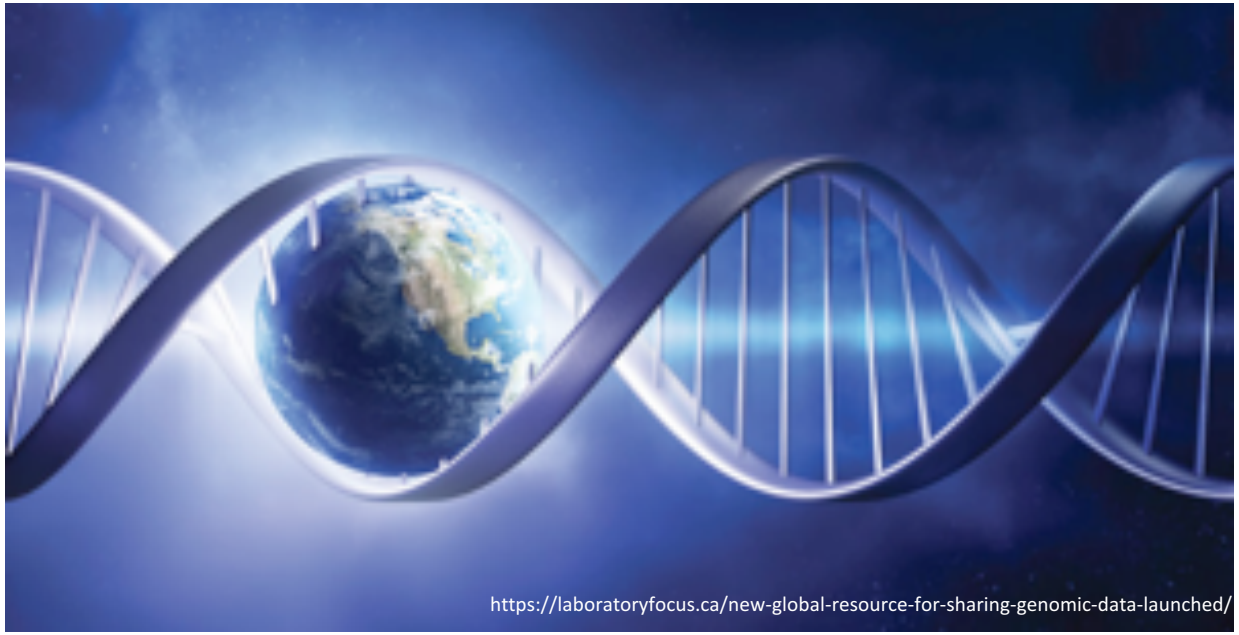
Marine ecosystems are extremely complex

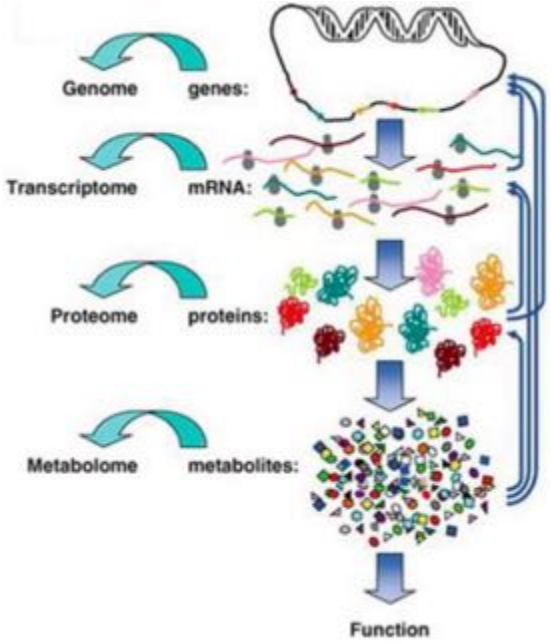
- Sizes (macro-micro)
- Diversity
- Interactions





DNA FOR MOLECULAR ECOLOGY AND EVOLUTION





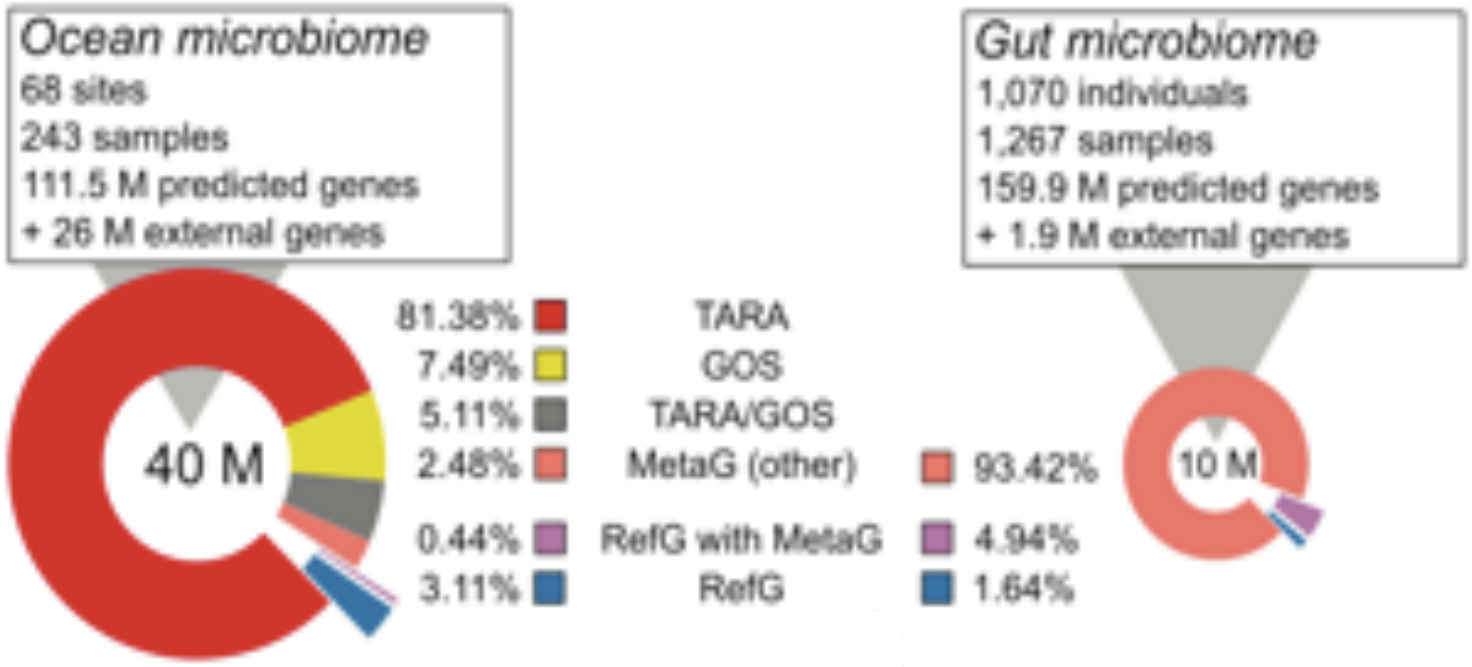
The TARA OCEANS example: Plankton genetic diversity

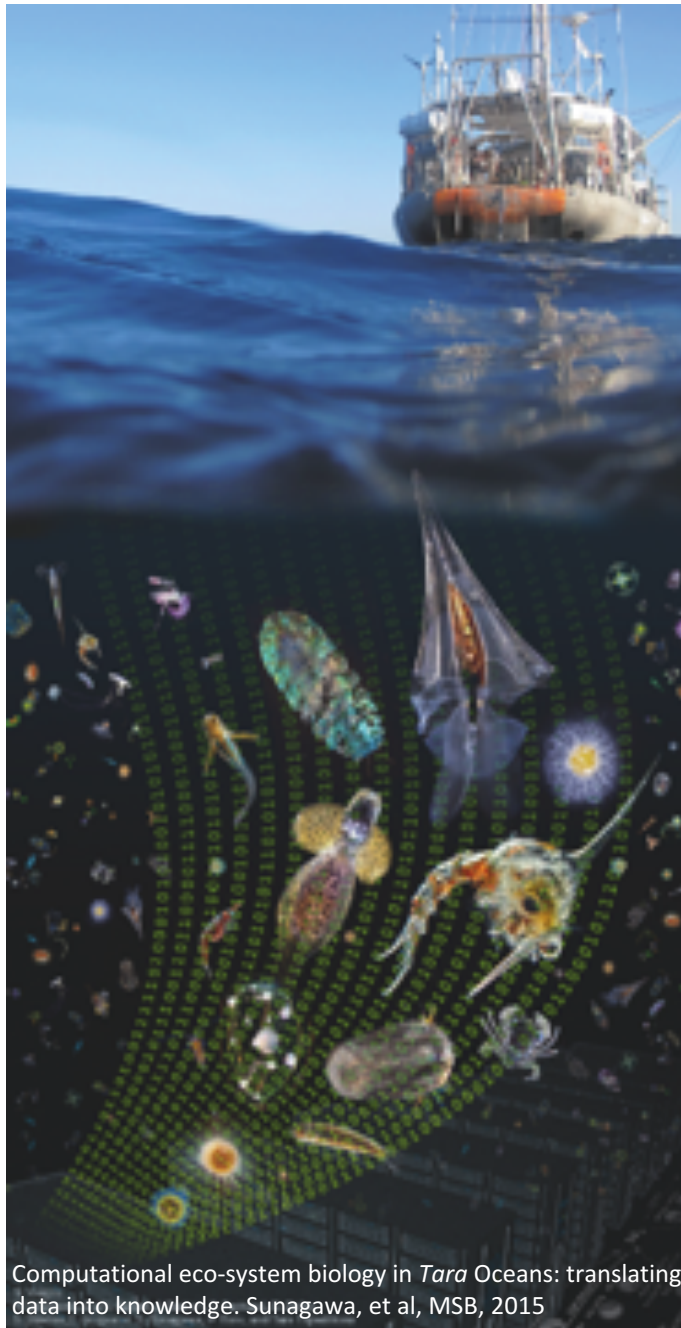
ENVIRONMENTAL GENOMICS

- 68 stations
- 3 depths
- 243 samples

7.2 Tbp DNA data
in the context of
the environment

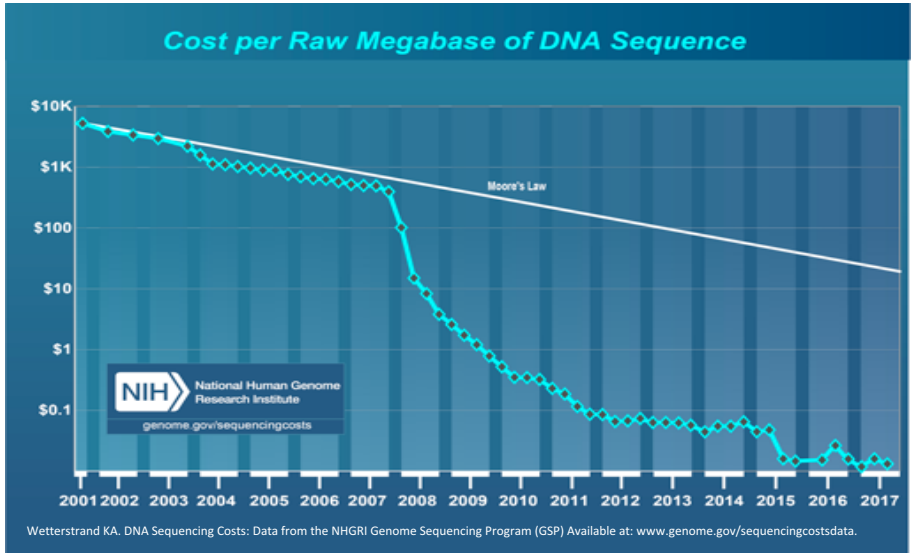
Integration of *Tara* Oceans and public data





Computational eco-system biology in *Tara* Oceans: translating data into knowledge. Sunagawa, et al, MSB, 2015

A BIG DATA ISSUE !!



Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

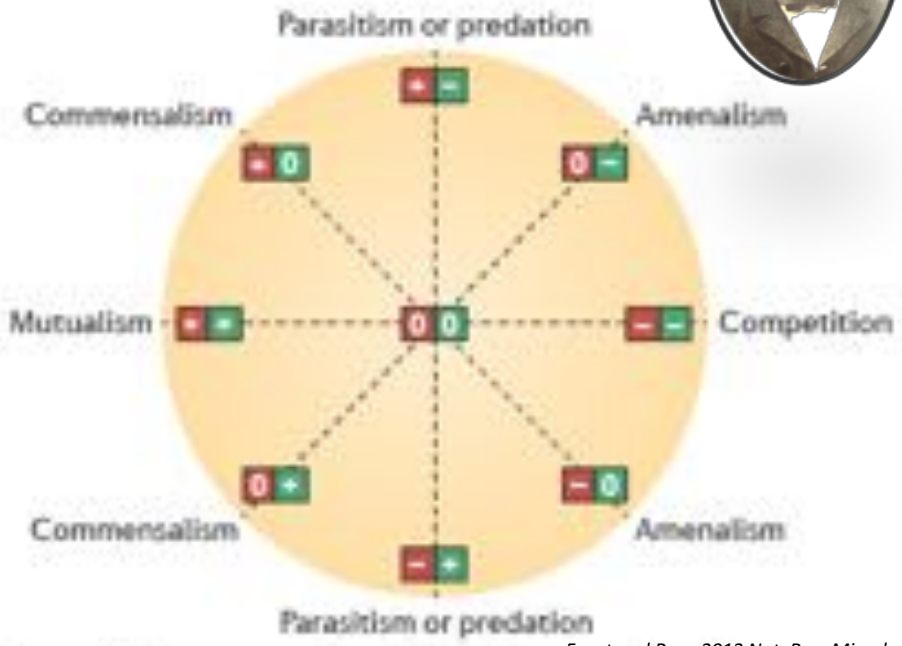
https://cloud.google.com/genomics/

(4) Analysis

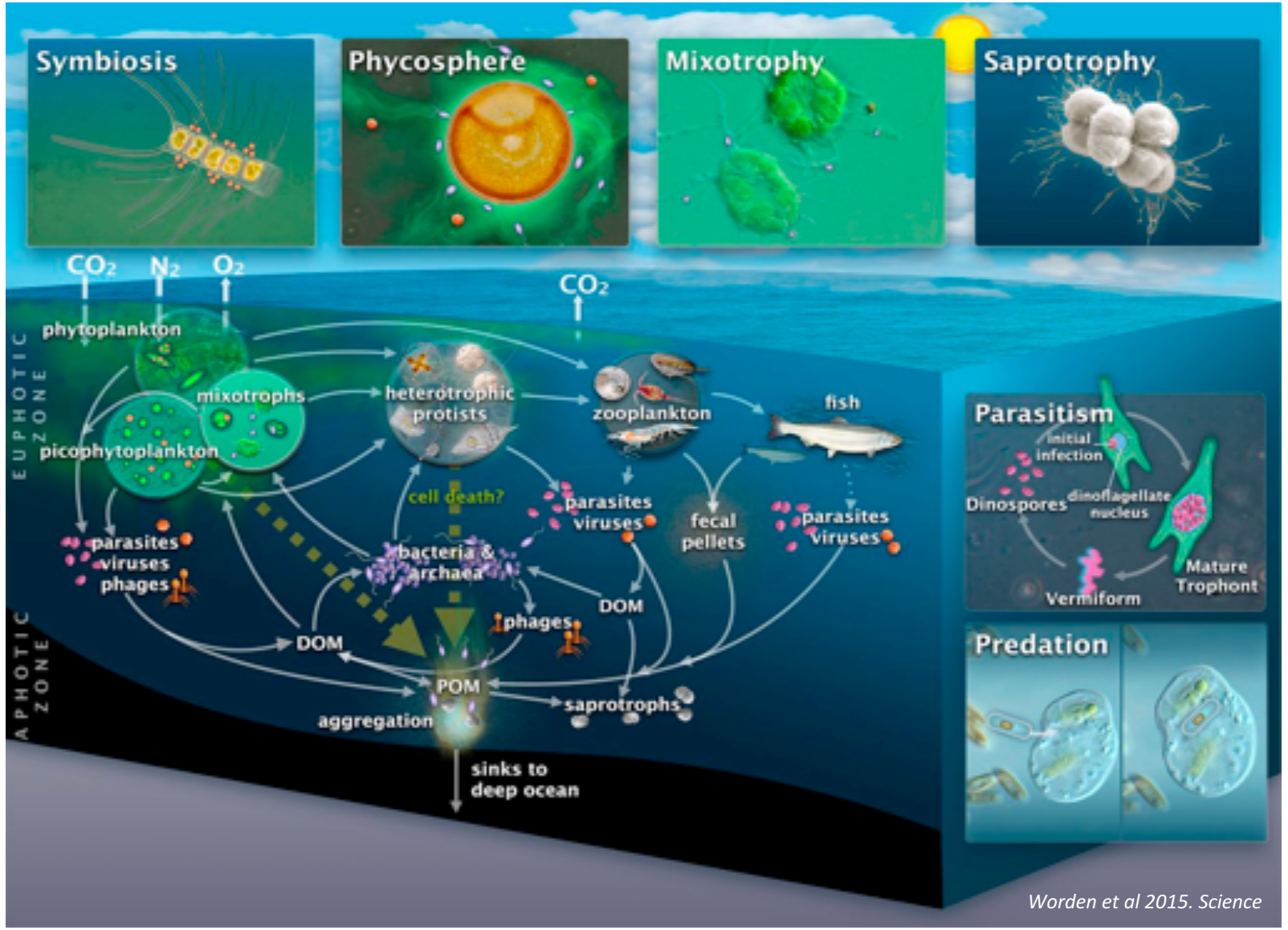
Our ultimate goal is to be able to interpret genomic sequences and answer how DNA mutations, expression changes, or other molecular measurements relate to disease, development, behavior, or evolution. Accomplishing this goal will clearly require integration of biological domain expertise, large-scale machine learning systems, and a computing infrastructure that can support flexible and dynamic queries to search for patterns over very large collections in very high dimensions. A number of "data science technologies," including R, Mahout, and other machine learning systems powered by Hadoop and other highly scalable systems, are a start, but the current offerings are still difficult and expensive to use. The community would also benefit from libraries of highly optimized algorithms within a simple interface that can be combined and reused in many contexts as the problems emerge. Data science companies as well as open-source initiatives are already starting to develop such components, such as Amazon's recent "Amazon Machine Learning" prediction system. But because genomics poses unique challenges in terms of data acquisition, distribution, storage, and especially analysis, waiting for innovations from outside our field is unlikely to be sufficient. We must face these challenges ourselves, starting with integrating data science into graduate, undergraduate, and high-school curricula to train the next generations of quantitative biologists, bioinformaticians, and computer scientists and engineers [49].

BIOTIC INTERCATIONS

Symbiosis: In 1879, H. Anton de Bary, defined symbiosis as "***the living together of unlike organisms*".** It is the close and long-term interactions between different biological taxa.



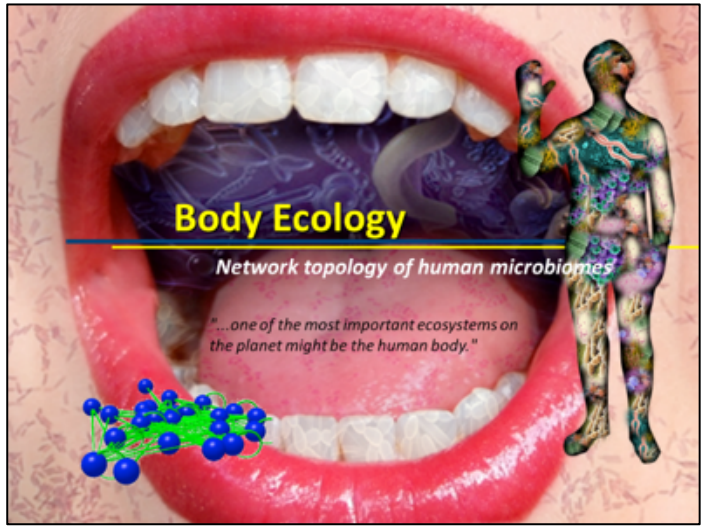
Faust and Raes 2012 Nat. Rev. Microb.

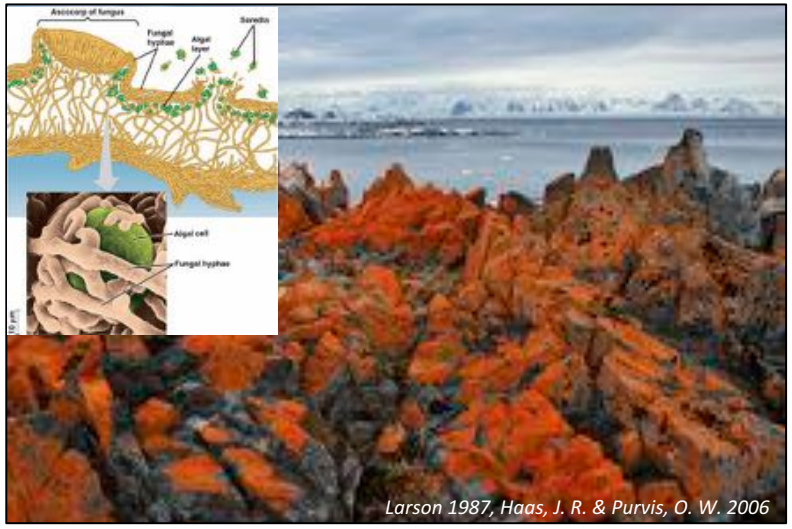


Worden et al 2015. Science

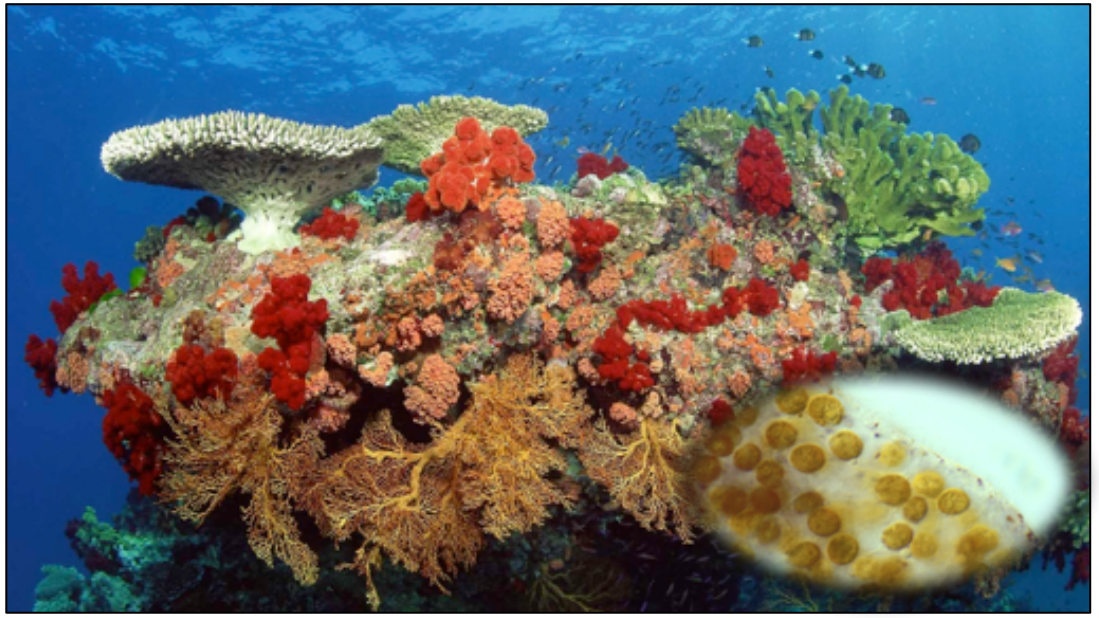
SYMBIOSIS IN ECOLOGY

Symbiosis on land



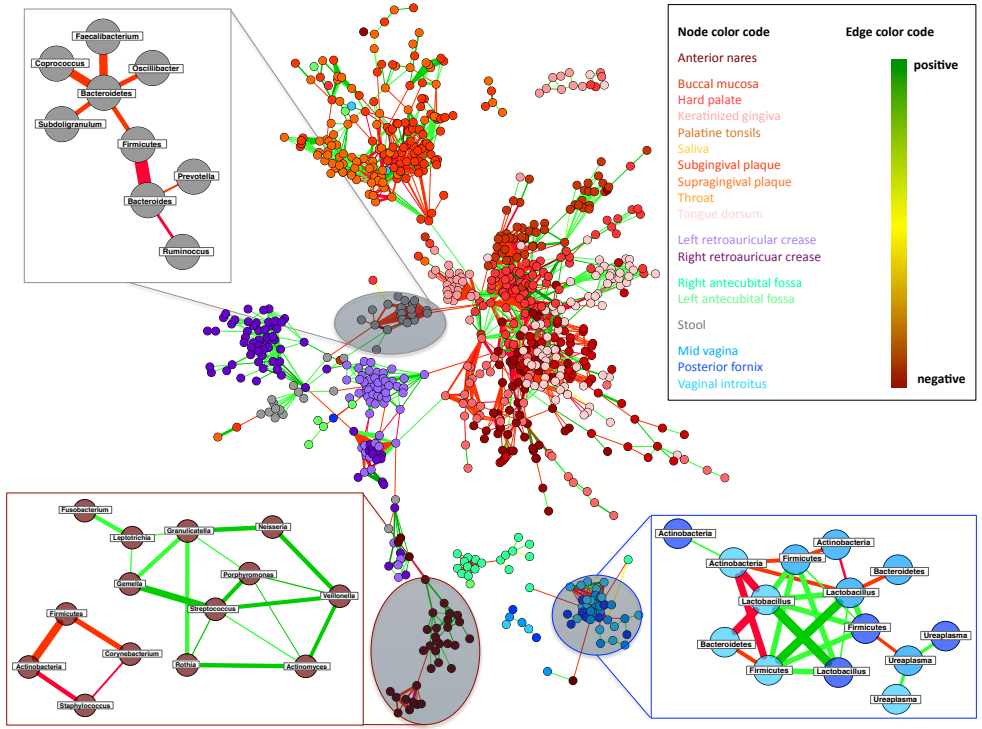


Marine benthic symbiosis



INTERCATIONS NETWORKS

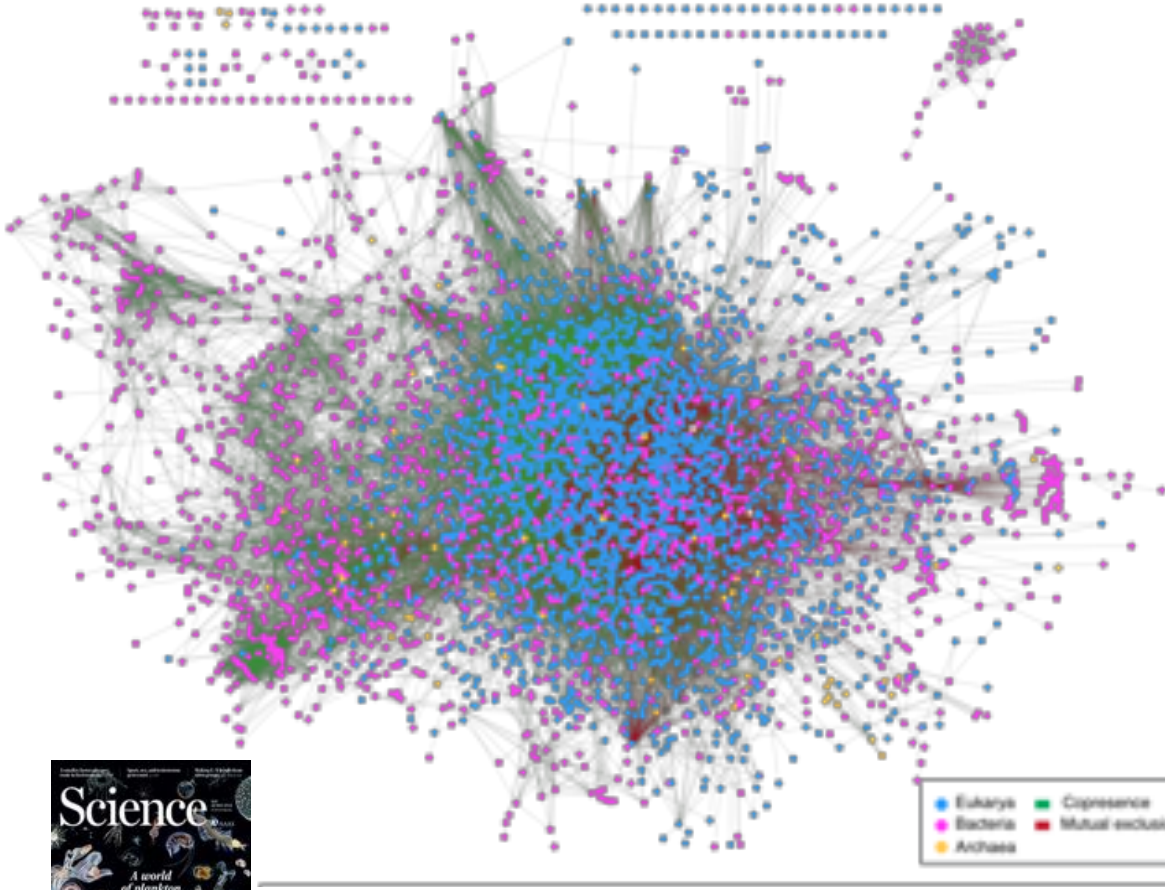
Reconstruction of the species interaction network of the whole human microbiome in 265 individuals and 18 bodysites



Huttenhower et al Nature 2012; Faust et al PLoS Comp Biol 2012

The global ocean interactome an integrated "network of networks"

A



127995 associations
92633 taxon-taxon
35362 taxon-env

Lima-Mendez et al Science 2015

COMPARATIVE GENOMICS = FUNCTION IDENTIFICATION

WILEY MOLECULAR ECOLOGY

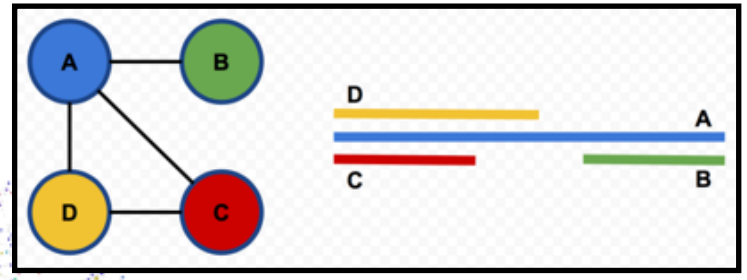
Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network

Arnaud Meng[1] | Erwan Corre[2] | Ian Probert[3] | Andres Gutierrez-Rodriguez[4] | Raffaele Siano[5] | Anita Annamale[6,7,8] | Adriana Alberti[6,7,8] | Corinne Da Silva[6,7,8] | Patrick Wincker[6,7,8] | Stéphane Le Crom[1] | Fabrice Not[9†] | Lucie Bittner[1†]

https://www.biorxiv.org/content/early/2017/10/30/211243

Sequence Similarity Network

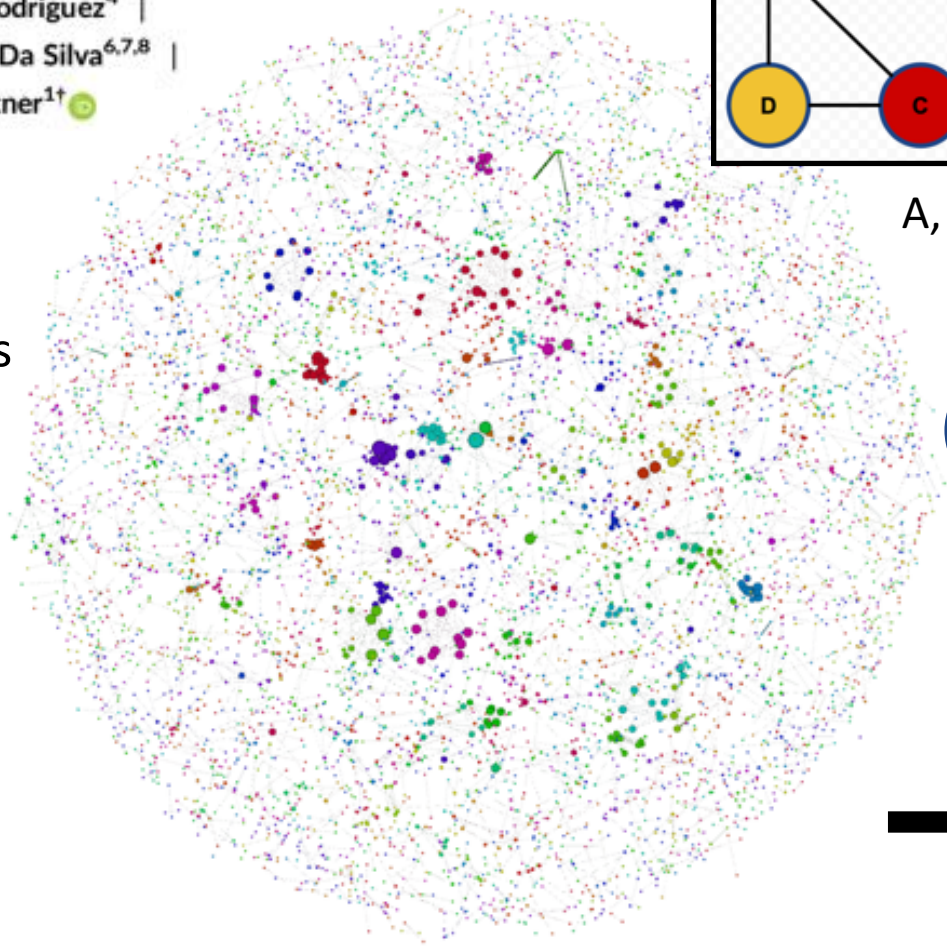
Connected component = CC

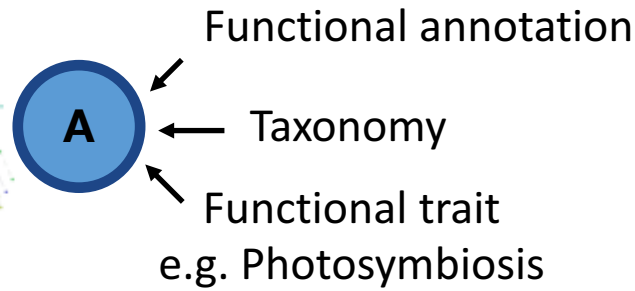


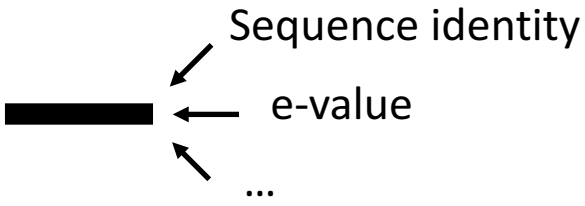
A, B, C and D are sequences

Goal:

Identify gene set involved in functional traits







≈ 80 individuals

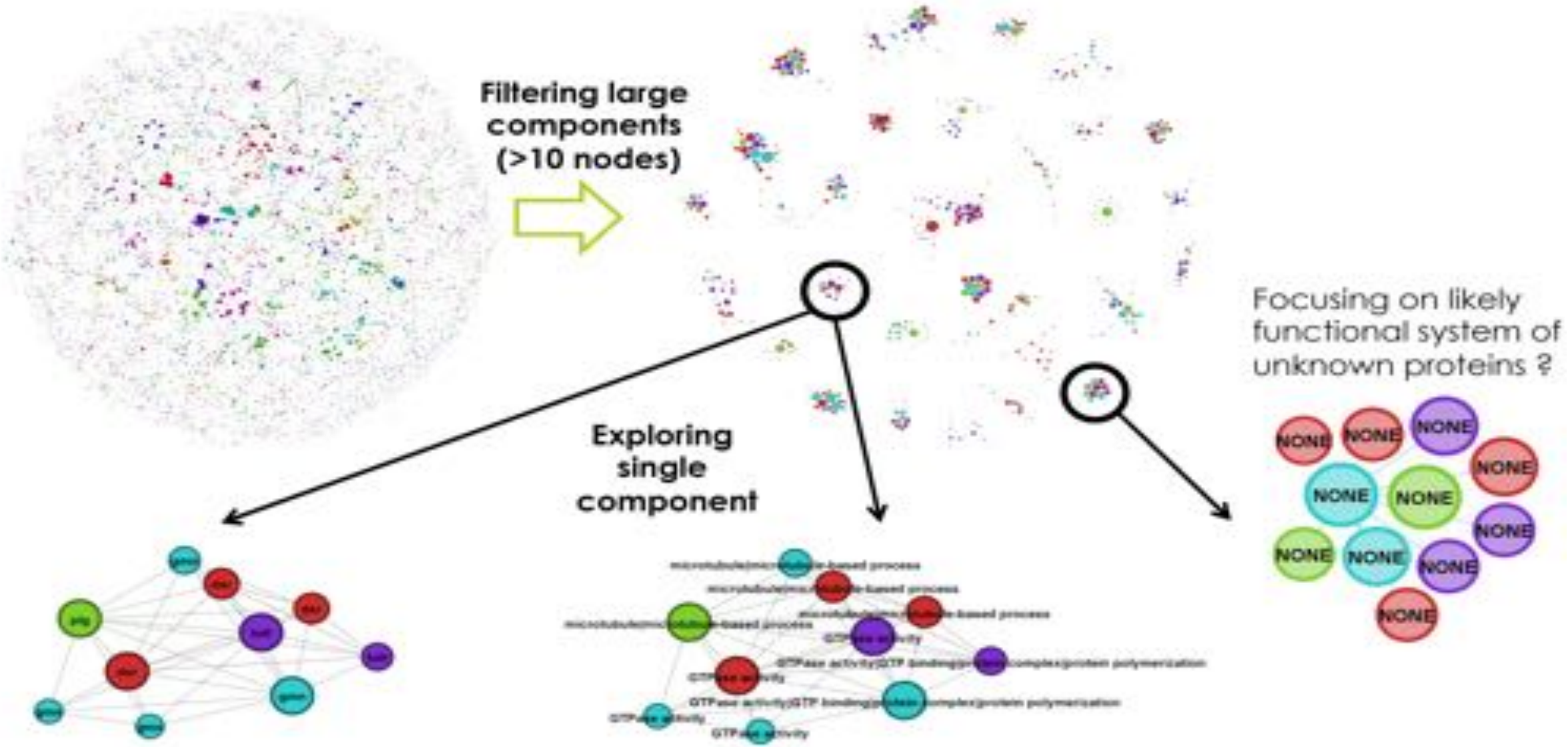
✓ 2,933,509 Sequences

✓ 31,023,513 Alignments

✓ 295,142 Connected components
(=protein coding domains)

✓ 186,697 no functional annotation (ca 60%)

PREDICTED PROTEIN SEQUENCE SIMILARITY NETWORK

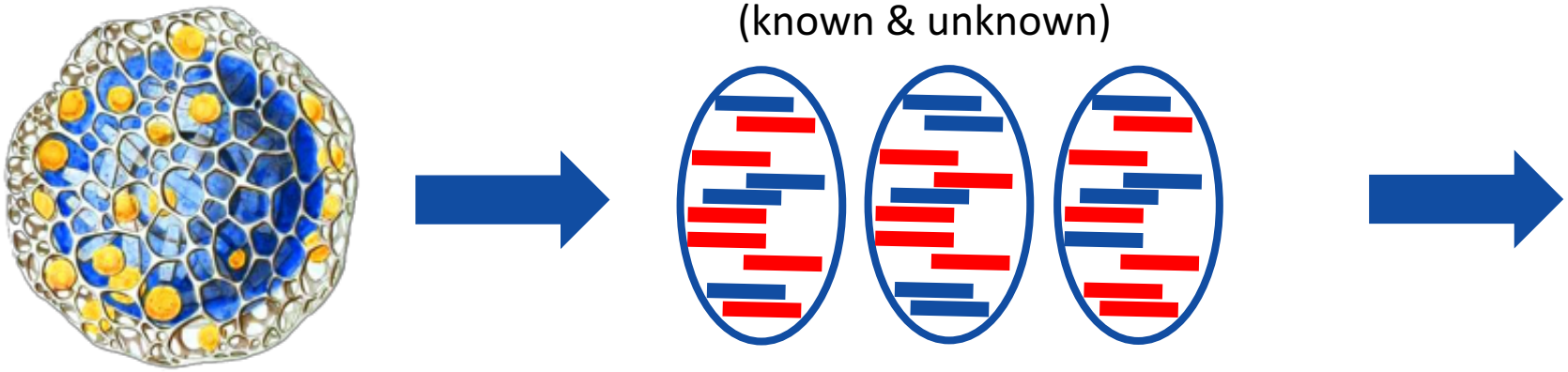


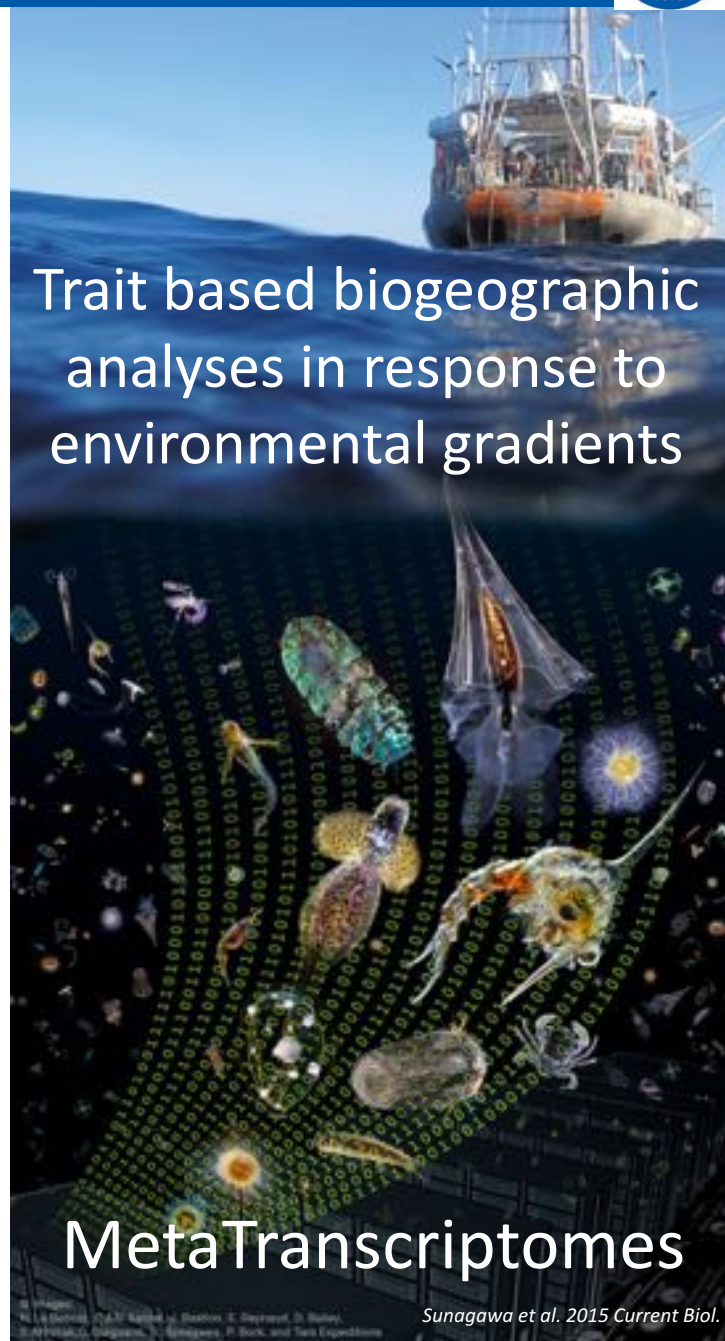
TRAIT BASED ECOLOGY WITH “OMICS” DATA

❖ Characterization of Genomic Functional Trait (*i.e.* GFT)

✓ Dedicated experiments (e.g. stress)

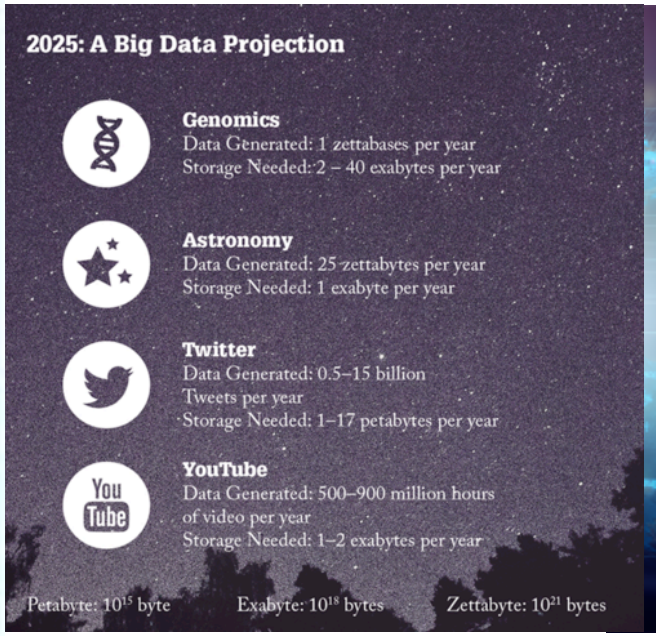
✓ Large scale comparative genomics
(using uncultured protists and unassigned genes….)





NEXT CHALLENGE: GENOMIC OBSERVATORY

http://www.assembleplus.eu/



Genes **2019**, *10*(1), 18; https://doi.org/10.3390/genes10010018

Open Access | Feature Paper | Perspective

Artificial Intelligence and Integrated Genotype–Phenotype Identification

Lewis J. Frey [1,2]

[1] Department of Public Health Sciences, Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC 29425, USA

[2] Health Equity and Rural Outreach Innovation Center (HEROIC), Ralph H. Johnson Veteran Affairs Medical Center, Charleston, SC 29401, USA

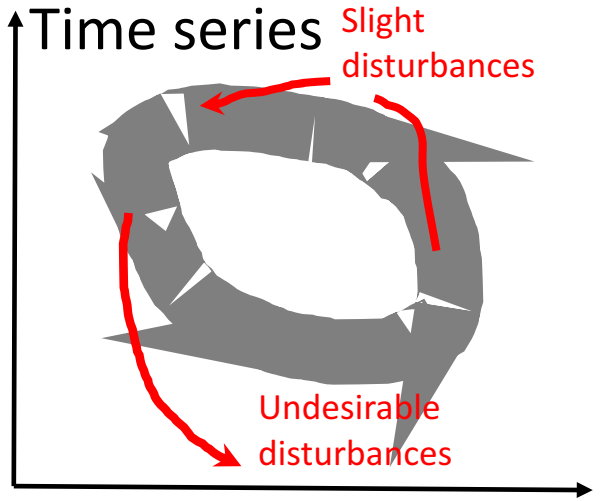
Received: 9 December 2018 / Revised: 20 December 2018 / Accepted: 21 December 2018 / Published: 28 December 2018

(This article belongs to the Special Issue Systems Analytics and Integration of Big Omics Data)

OBJECTIVES AND CHALLENGES

Predict ecosystems' functions evolution under changing conditions

(fundamental knowledge, law of Nature, Conservation, Ecosystem health, society etc…)

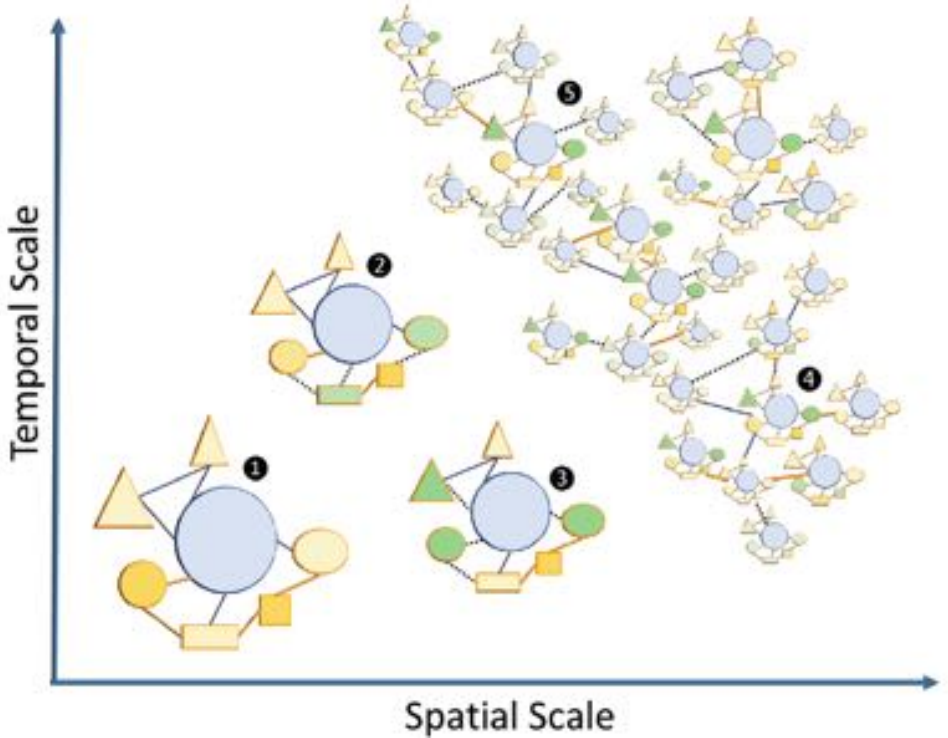


- Describing & understanding patterns, drivers
- Natural enveloppe of variations (vs disturbed ecosystem)

A community perspective on the concept of marine holobionts: state-of-the-art, challenges, and future directions

The Holomarine working group*: Simon M. Dittami†, Enrique Arboleda, Jean-Christophe Auguet, Arite Bigalke, Enora Briand, Paco Cárdenas, Ulisse Cardini, Johan Decelle, Aschwin Engelen, Damien Eveillard, Claire M.M. Gachon, Sarah Griffiths, Tilmann Harder, Ehsan Kayal, Elena Kazamia, Francois H. Lallier, Mónica Medina, Ezequiel M. Marzinelli, Teresa Morganti, Laura Núñez Pons, Soizic Pardo, José Pintado Valverde, Mahasweta Saha, Marc-André Selosse, Derek Skillings, Willem Stock, Shinichi Sunagawa, Eve Toulza, Alexey Vorobev, Catherine Leblanc†, and Fabrice Not†

Russian dolls models



- Functional trait in ecology based on omics (ref database for ID or IA)

- Integrate various kinds of data and scales (DNA, metabo, physio, physics, chemistry, …)

- Deal with and make sense of the unknown !!

- Holobionts. Fundamental paradigm change in bio and eco

- Detailed microbial-macrobial ecological models. Would include "individual based models" where each individual organism is explicitly modelled, including internal functions and external interactions.