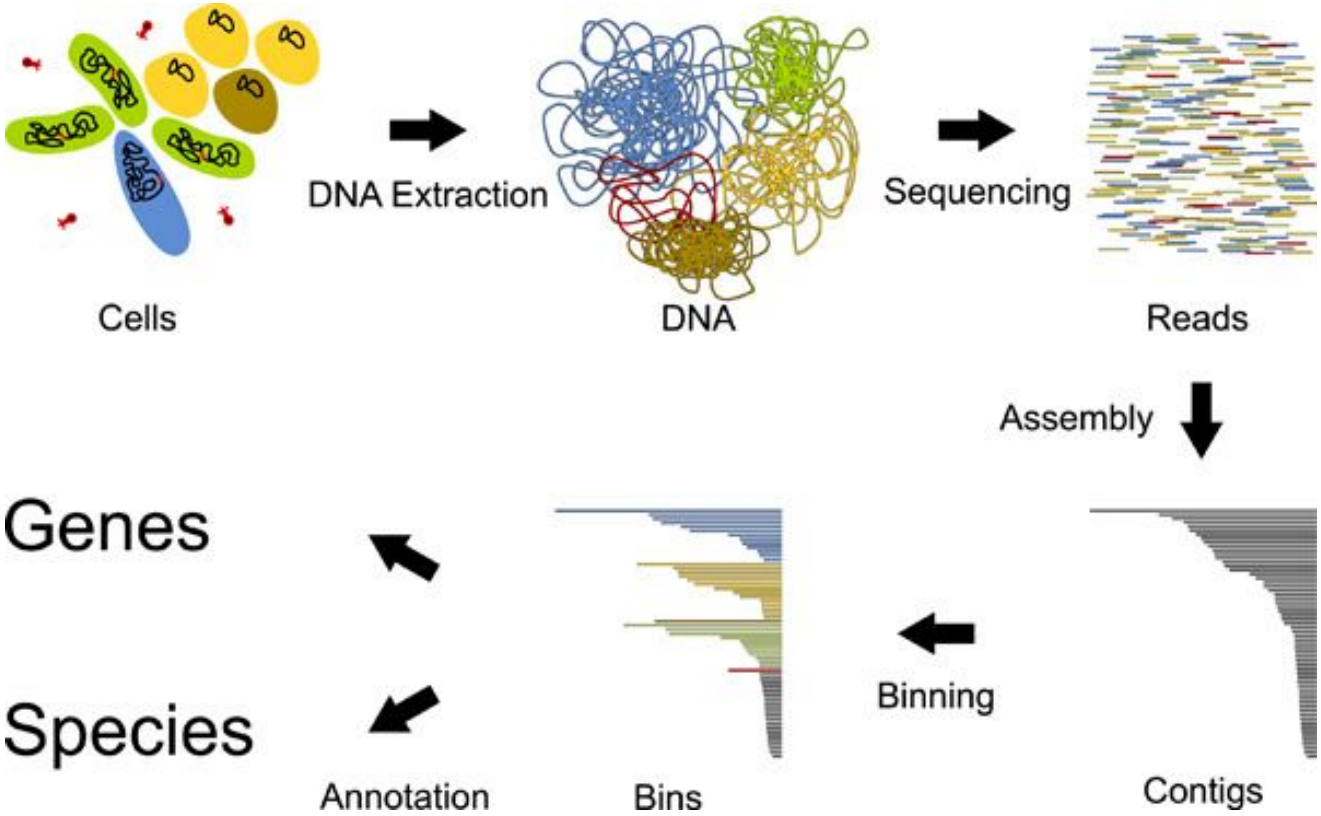
Metagenomics and ecological networks

From a sample to an ecosystem

- Metagenomics
- The current state of the art
  - Assembly
  - Binning
- Extant biodiversity and the unknown
- Functional annotation
  - Structures and AI
- Finding networks and characterizing them.
- Taking an evolutionary angle relationships?

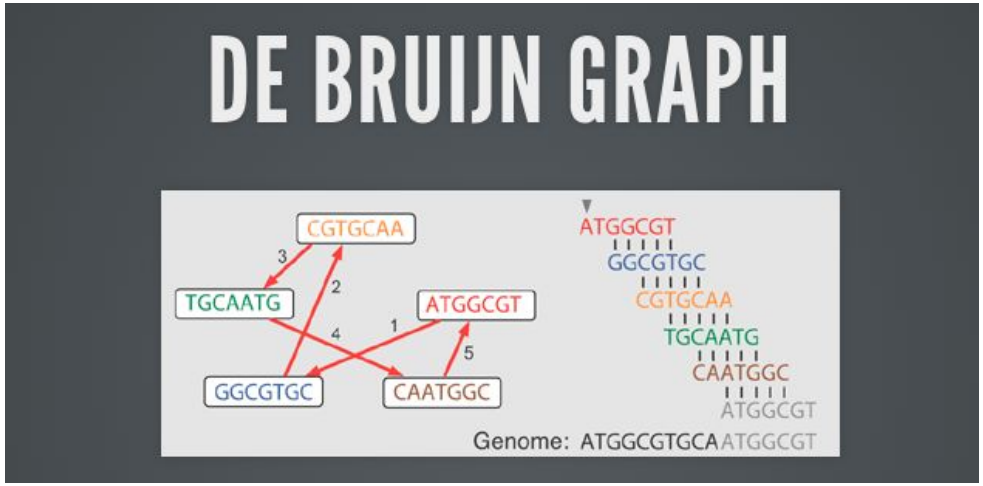
How do we even get this data?

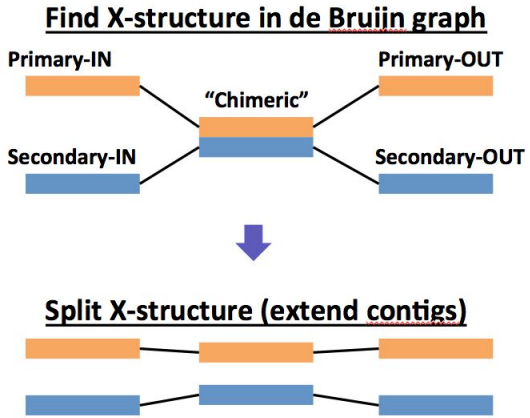
In a nutshell



Metagenomic Assembly

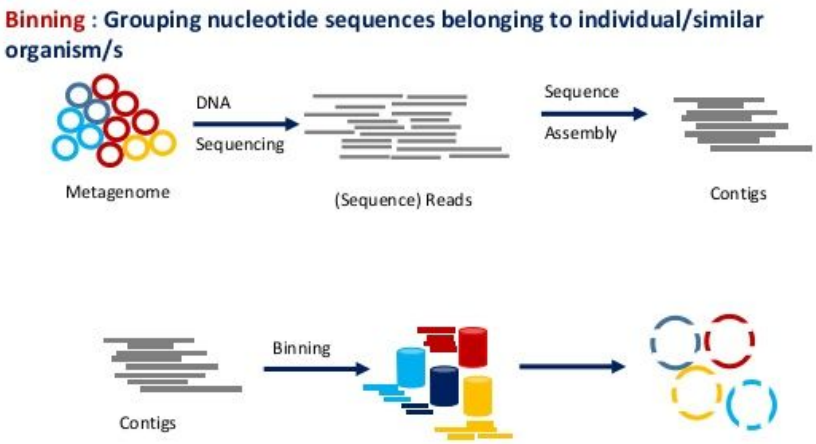
- Most approaches will use a De Bruijn graph.
  - Link all reads by the presence of shared KMERs
  - The reads are nodes on a graph.
  - The final sequence is dictated by the series of nodes the graph passes through.
- Metagenome-assembled genomes (MAGs)





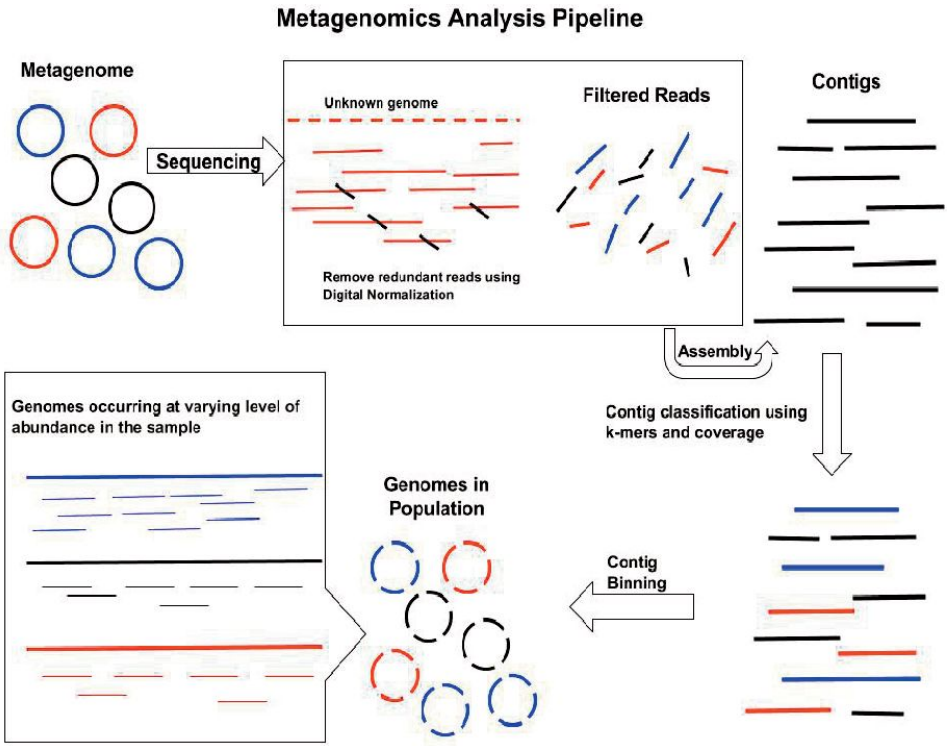
Binning

- meta-assemble a sample of short reads into contigs
  - Then cluster contigs based on some criteria.
  - In theory each bin corresponds to a genome.
- Two criteria: KMERs and Abundance.
- Time complexity remains an issue



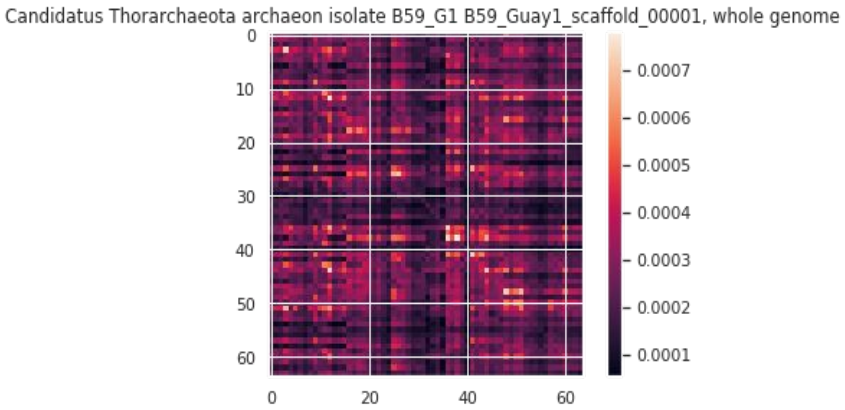
Co-abundance binning

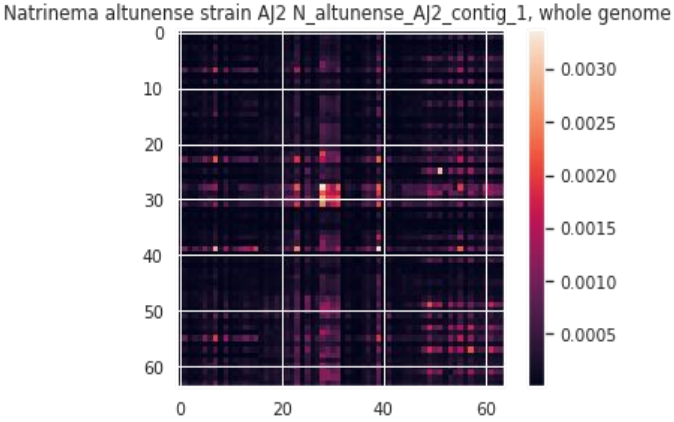
- Across different samples
- The abundance of organisms can help group reads
- Binning the reads and then assembling increases accuracy
- Still requires substantial coverage



Kmer spectra Binning

- Words of length k in the DNA
- These are optimised by each organism for its niche
- Can be used to group reads
- Here k=6, 4096 possible words of length 6





Metaproteomics: PLASS

- What if we don’t even care about genomes?
- We lose synteny, intragenic regions etc
- We gain sensitivity!
- Bag of proteins

Brief Communication | Published: 24 June 2019

Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold

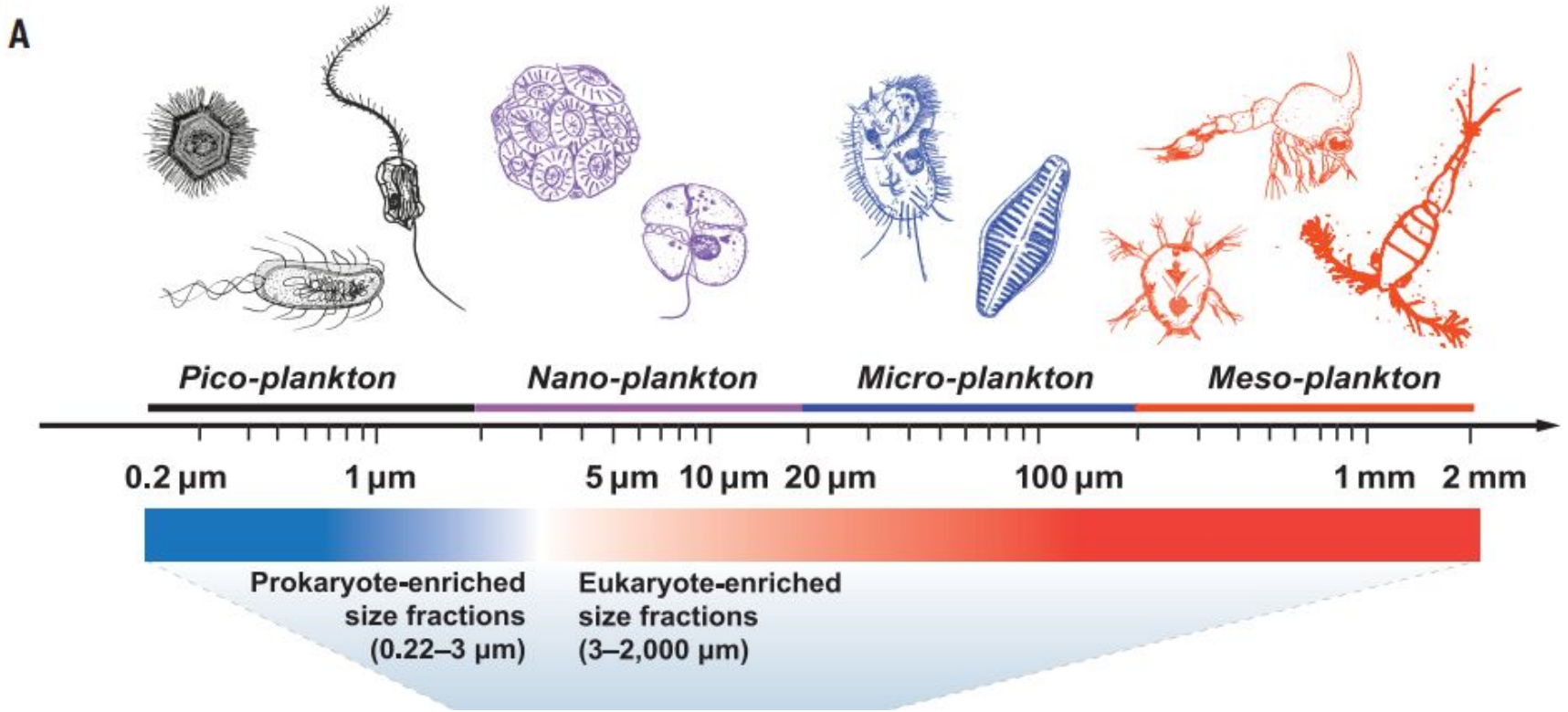
Martin Steinegger, Milot Mirdita & Johannes Söding

Nature Methods **16**, 603–606 (2019) | Cite this article

8172 Accesses | **131** Citations | **61** Altmetric | Metrics

What’s out there?

Filtering is already informative



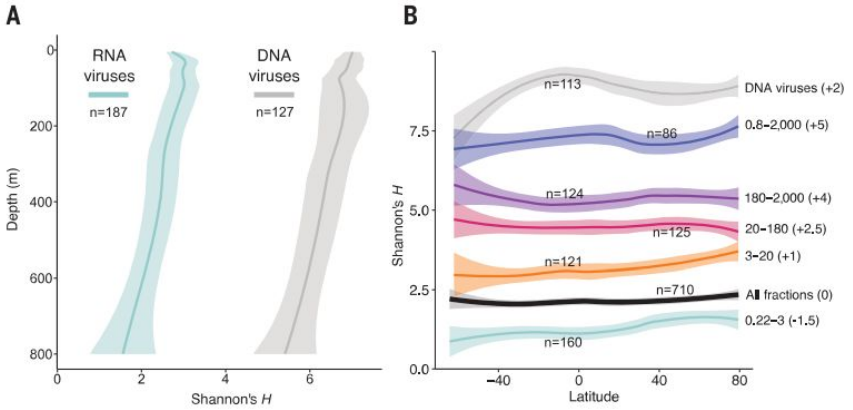
Extant diversity: Viruses

- In this Tara oceans they quantify the diversity using entropic measures.
  - $$H = -\sum p(x)\log p(x)$$
- It varies along the sampling depth and latitude.
- We don't know what most of these viruses infect or what their viral cycle is like.
- Another study: (<6% of reads matched known viruses)

MARINE VIROME

Diversity and ecological footprint of Global Ocean RNA viruses

Guillermo Dominguez-Huerta[1,2,3]†, Ahmed A. Zayed[1,2,3]†, James M. Wainaina[1,3], Jiarong Guo[1,2,3], Funing Tian[1,3], Akbar Adjie Pratama[1,2], Benjamin Bolduc[1,2,3], Mohamed Mohssen[1,3,4], Olivier Zablocki[1,2,3], Eric Pelletier[5,6], Erwan Delage[6,7], Adriana Alberti[5,6]‡, Jean-Marc Aury[5], Quentin Carradec[5,6], Corinne da Silva[5], Karine Labadie[5,6], Julie Poulain[5,6], Tara Oceans Coordinators§, Chris Bowler[6,8], Damien Eveillard[6,7], Lionel Guidi[6,9], Eric Karsenti[6,8,10], Jens H. Kuhn[11], Hiroyuki Ogata[12], Patrick Wincker[5,6], Alexander Culley[13], Samuel Chaffron[6,7], Matthew B. Sullivan[1,2,3,4,14]*



Open Problem: Eukaryotes!

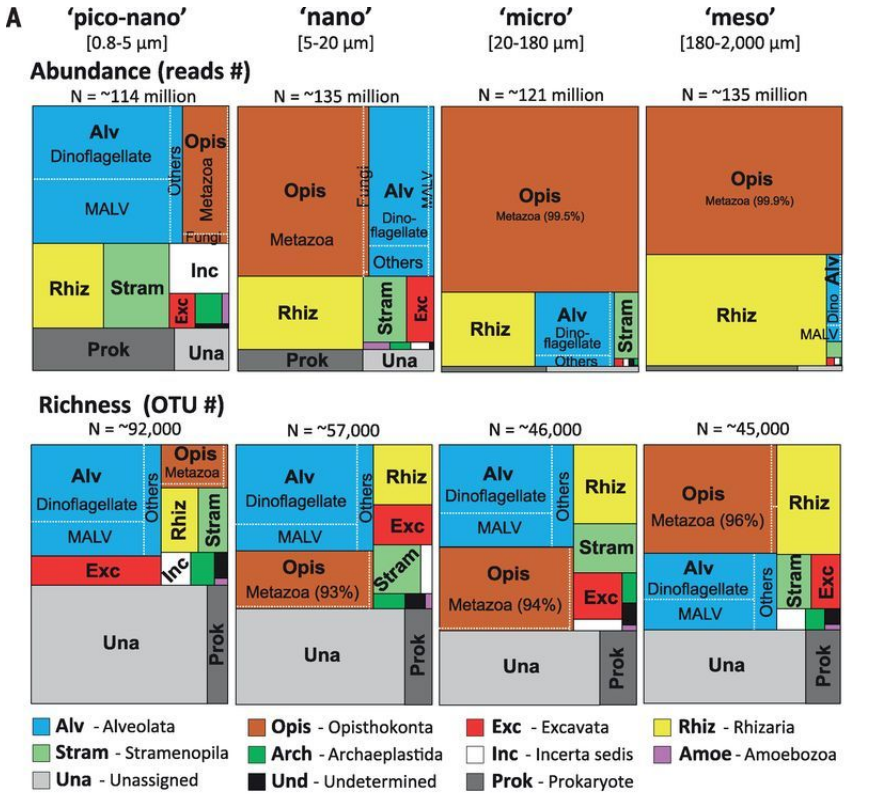
- Metagenomics assemblies are still out of reach for eukaryotes.
- Low coverage
  - We can't even attribute their rDNA!
- Many early branching species.
  - Could be crucial in understanding the emergence of eukarya from archaea.

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas[1,2,*,†], Stéphane Audic[1,2,†], Nicolas Henry[1,2,†], Johan Decelle[1,2,†], Frédéric Mahé[3,1,2,†], Ramiro Logares[4]...

+ See all authors and affiliations

Science 22 May 2015:
Vol. 348, Issue 6237, 1261605
DOI: 10.1126/science.1261605



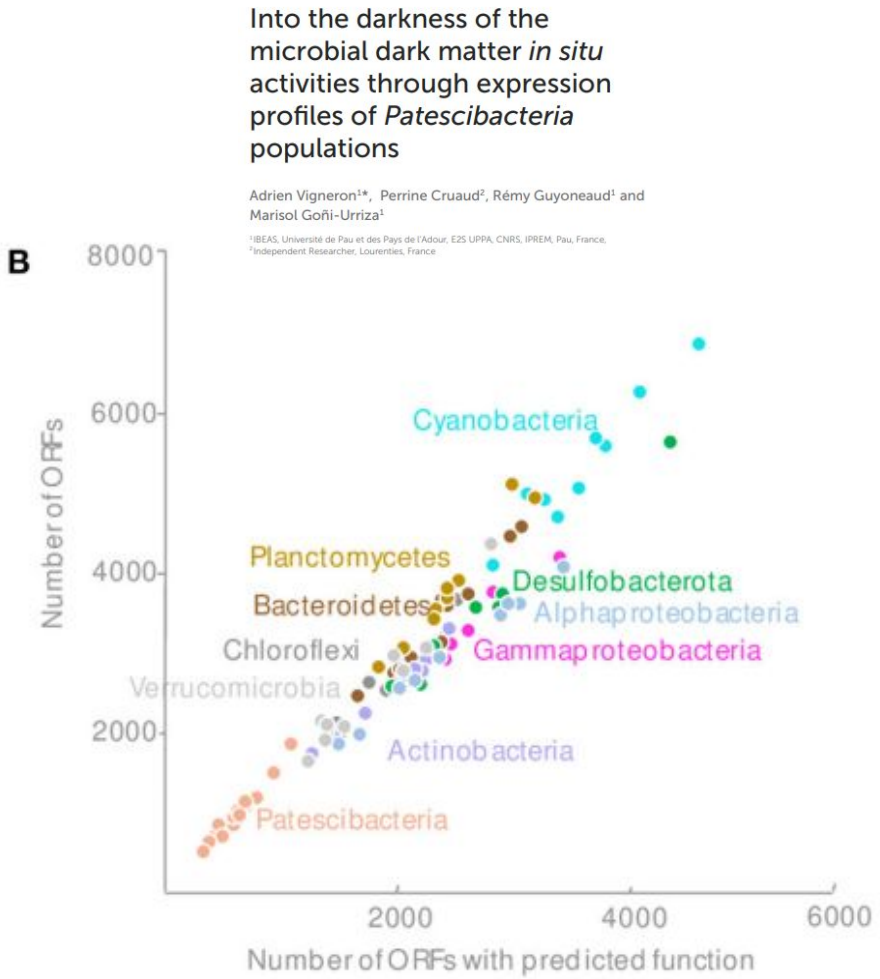
What’s in their genomes?

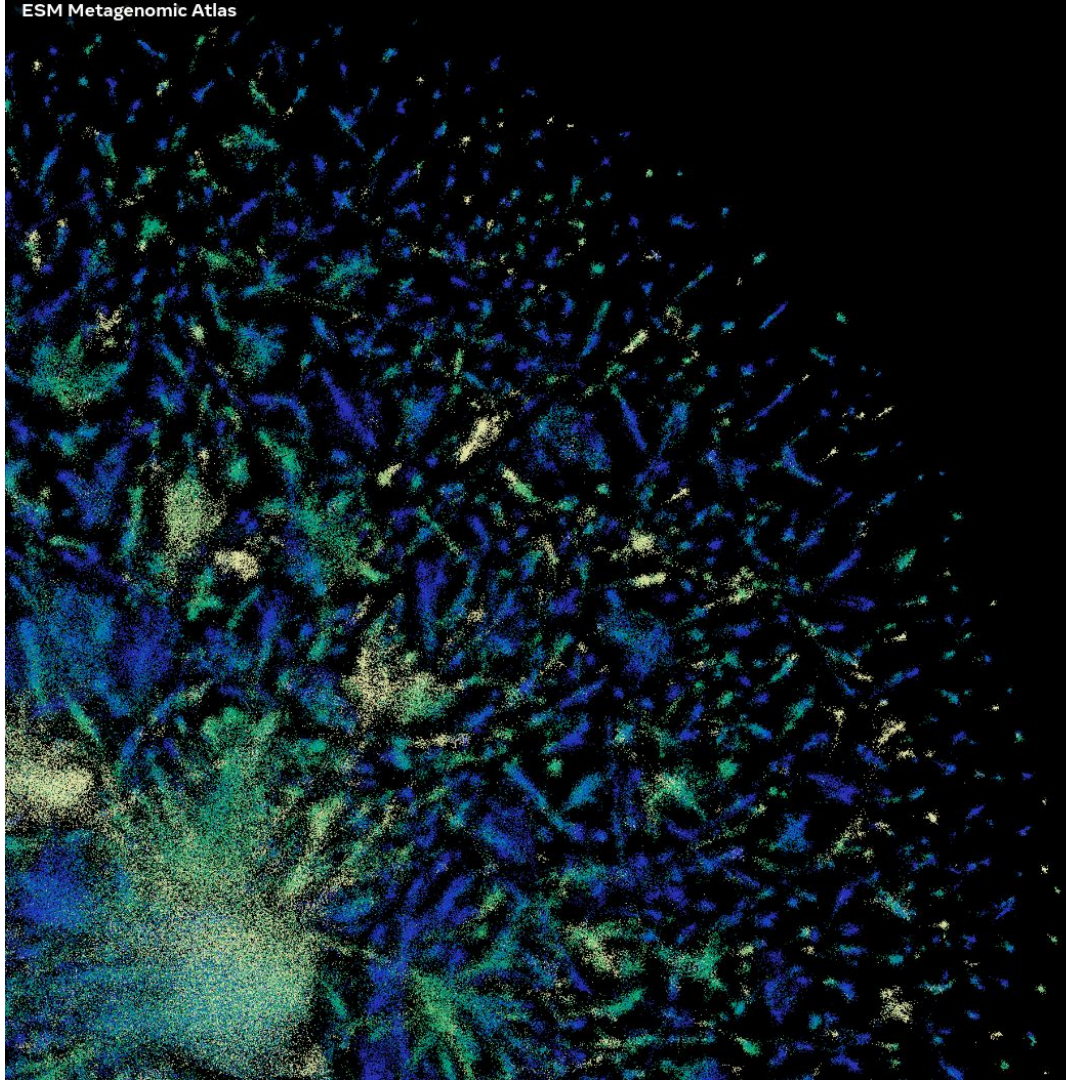
Functional annotation

- Map to something you know…
- KEGG
- UniProt
- eggNOG (and OMA we hope soon)
- Pfam
- Transfer knowledge based on homology… to model organisms

Metagenomic dark matter

- 25% or more in some cases have No Detectable Homology to anything we know.
- In the mobile fraction / pangenome this often reaches above 50%.
- In the other cases where we can map our knowledge is limited.
- How do we proceed?





ESM Metagenomic Atlas

An open atlas of 772 million predicted metagenomic protein structures

Explore →

Fold sequence ↗

Read blog post ↗

Read research paper ↗

ESM: evolutionary scale model

- Form is tied to function
- We can use structural information now
- ESM folded the Mgnify database
- We now have 600 M structures attributed to MAGs

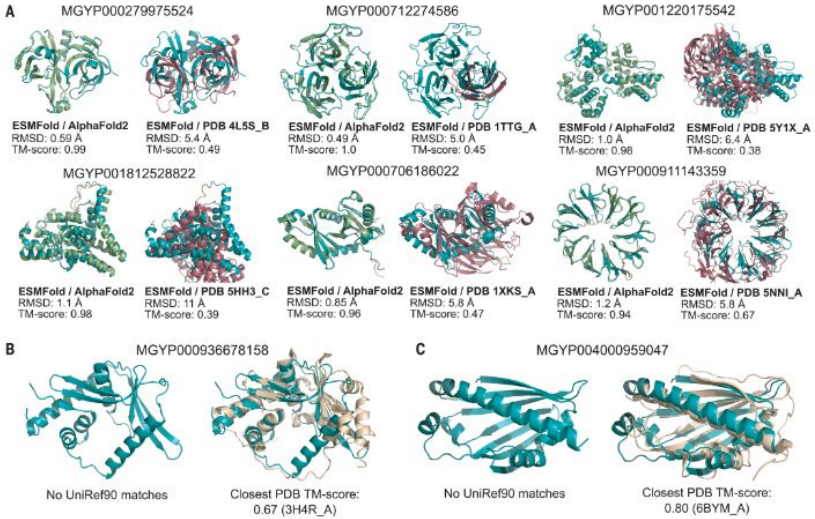


Fig. 4. Example ESMFold structure predictions of metagenomic sequences.

RESEARCH ARTICLE | BIOLOGICAL SCIENCES

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives, Joshua Meier, Tom Sercu, +7, and Rob Fergus Authors Info & Affiliations

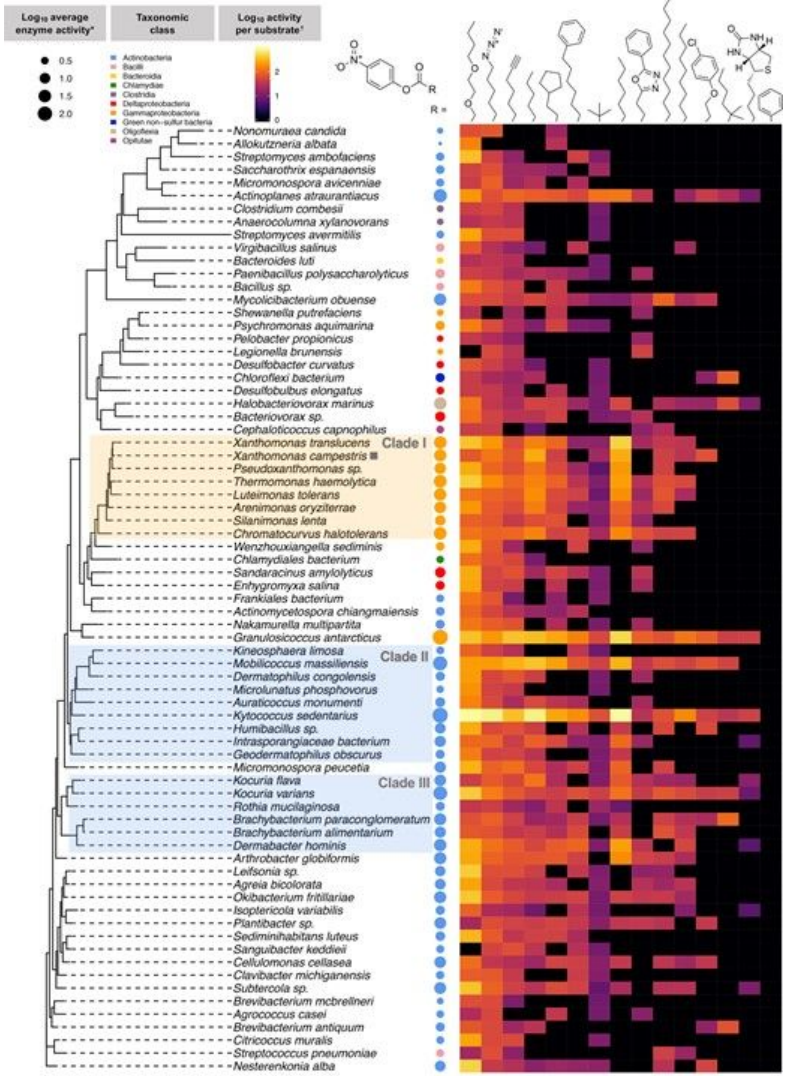
Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)

April 5, 2021 | 118 (15) e2016239118 | https://doi.org/10.1073/pnas.2016239118

Evolutionary-scale prediction of atomic-level protein structure with a language model

ZEMING LIN, HALIL AKIN, ROSHAN RAO, BRIAN HIE, ZHONGKAI ZHU, WENTING LU, NIKITA SMETANIN, ROBERT VERKUIL, ORI KABELI, [...], AND ALEXANDER RIVES +5 authors Authors Info & Affiliations

SCIENCE · 16 Mar 2023 · Vol 379, Issue 6637 · pp. 1123-1130 · DOI: 10.1126/science.ade2574



Inferring function: the AI revolution

Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily

Serina L Robinson, Megan D Smith, Jack E Richman, Kelly G Aukema, Lawrence P Wackett

Synthetic Biology, Volume 5, Issue 1, 2020, ysaa004,
https://doi.org/10.1093/synbio/ysaa004
Published: 27 May 2020 **Article history**

- One class of enzymes
- They are interested in biosynthetic gene clusters
- What do organisms make?
- They wanted to assign the substrate
- Trained a machine learning model on structural features!

Each actor has its ecological niche

Niche specificity: Generalists vs specialists

- Varying
  - Geographic distribution
  - Functional content
  - Coupling to other organisms
- In this work they explore the spread of each niche as a function of an organism's genomic/pangenomic content

nature ecology & evolution

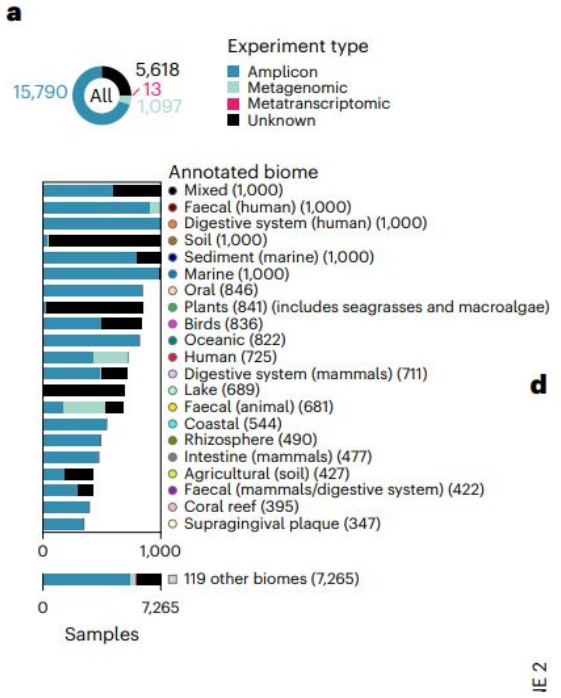
Article

https://doi.org/10.1038/s41559-023-02027-7

A social niche breadth score reveals niche range strategies of generalists and specialists

Received: 22 July 2022

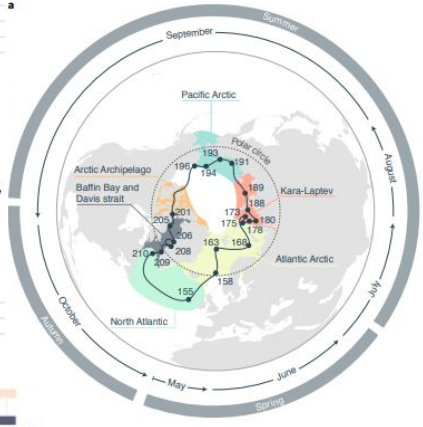
F. A. Bastiaan von Meijenfeldt[1,3], Paulien Hogeweg[1] & Bas E. Dutilh[1,2]

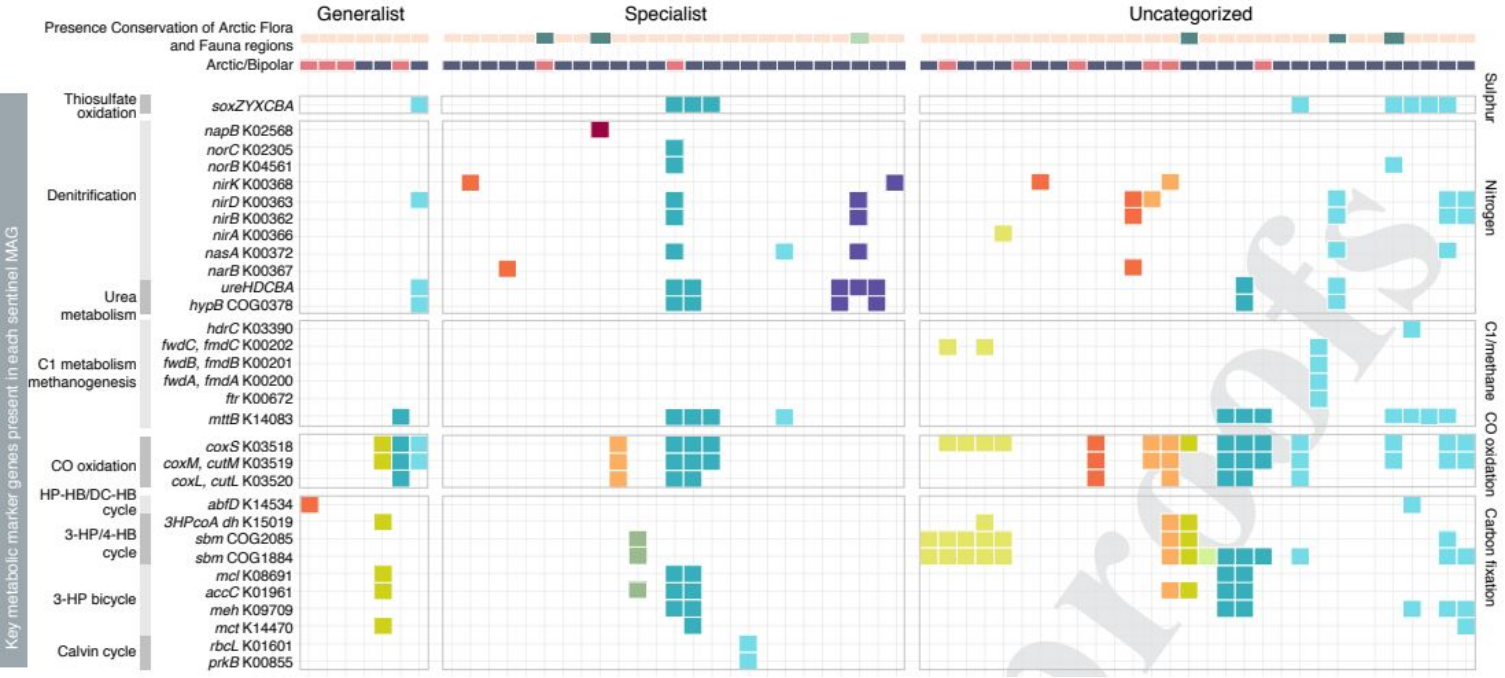


One example: Arctic prokaryotes

Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean

Marta Royo-Llonch[1], Pablo Sánchez[1], Clara Ruiz-González[1], Guillem Salazar[2], Carlos Pedrós-Alió[3], Marta Sebastián[1], Karine Labadie[4], Lucas Paoli[2], Federico Ibarbalz[5], Lucie Zinger[5], Benjamin Churcheward[6], Tara Oceans Coordinators*, Samuel Chaffron[6,7], Damien Eveillard[6,7], Eric Karsenti[5,7,8], Shinichi Sunagawa[2], Patrick Wincker[4,7], Lee Karp-Boss[9], Chris Bowler[5,7] and Silvia G. Acinas[1]✉





Let’s bring it together!

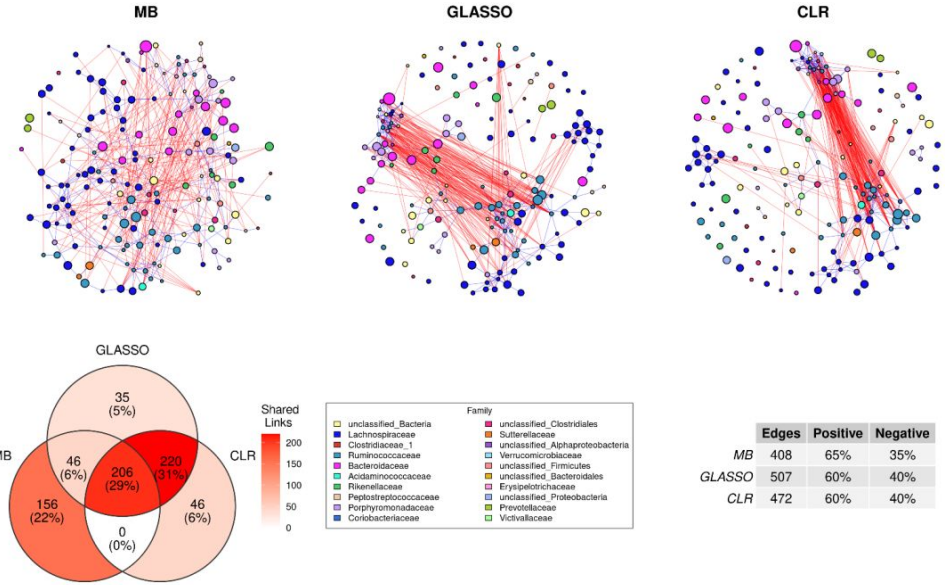
Building networks: Correlation and anticorrelation

- Discussing different network building approaches using co-abundance
- Methods will pick up on different associations
- Some links are shared

Comprehensive analysis of network reconstruction approaches based on correlation in metagenomic data

Alessandro Fuschi[1], Alessandra Merlotti[1], Thi Dong Binh Tran[2], Hoan Nguyen[2], George M. Weinstock[2,3], Daniel Remondini[1*],

1 Department of Physics and Astronomy, University of Bologna, Bologna 40127, IT.
2 The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030 USA
3 Dept. Genetics and Genome Science, University of Connecticut Health Center, Farmington, CT 06032 USA



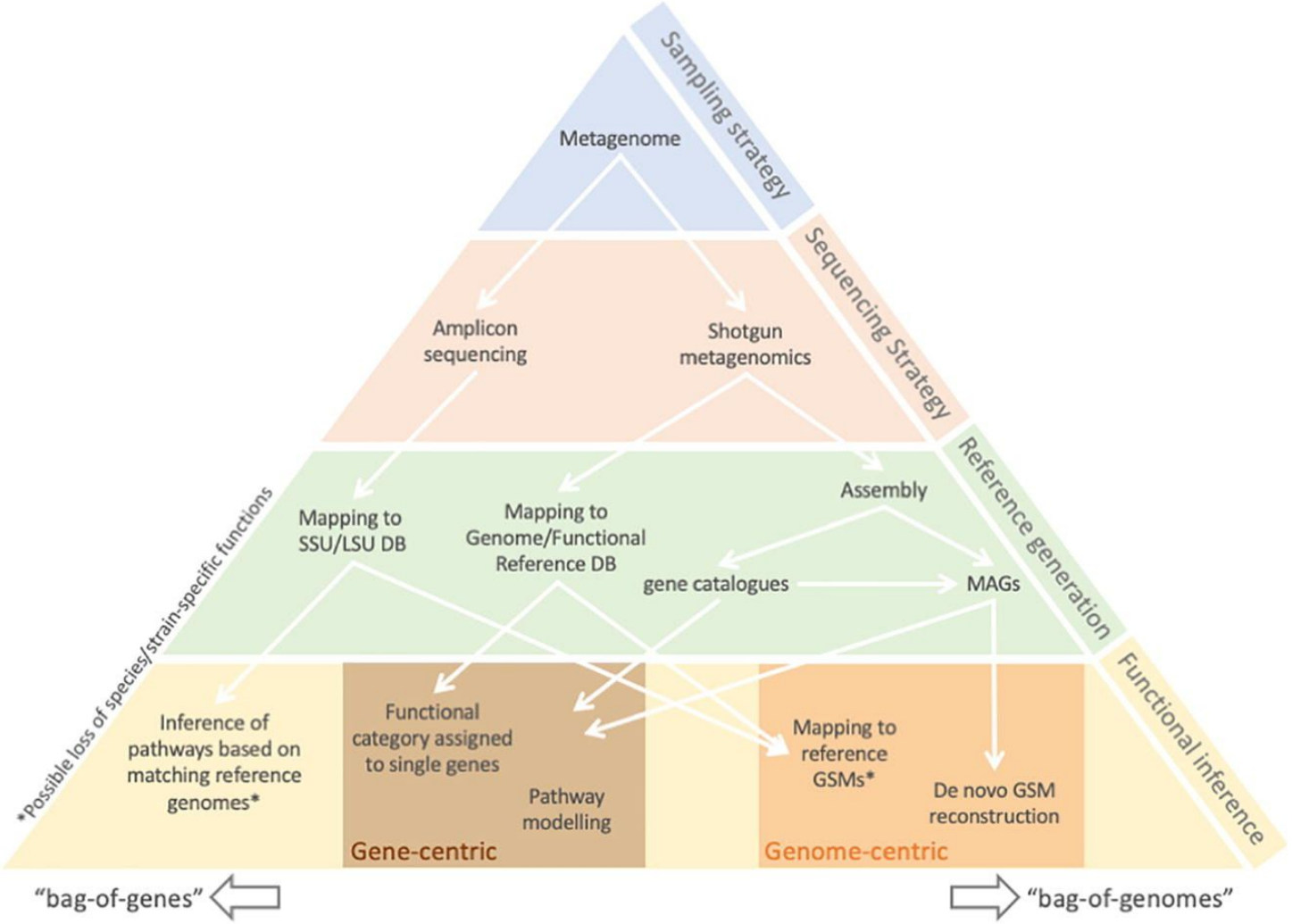
	Edges	Positive	Negative
MB	408	65%	35%
GLASSO	507	60%	40%
CLR	472	60%	40%

Greater than the sum of its parts: The ecological net.

- Organisms are metabolically complementary
- Predator prey / Symbiotic / Parasitic etc
- With each mag we build a genome scale model GSM
- Can we find the metabolic flux through each organism?
- This is more of an opinion piece…

From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes

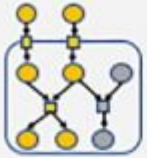
Clémence Frioux [a b], Dipali Singh [c], Tamas Korcsmaros [b d], Falk Hildebrand [b d]


Sampling strategy
Metagenome
Sequencing Strategy
Amplicon sequencing
Shotgun metagenomics
Reference generation
Mapping to SSU/LSU DB
Mapping to Genome/Functional Reference DB
Assembly
gene catalogues
MAGs
Functional inference
*Possible loss of species/strain-specific functions
Inference of pathways based on matching reference genomes*
Functional category assigned to single genes
Pathway modelling
Gene-centric
Mapping to reference GSMs*
De novo GSM reconstruction
Genome-centric
"bag-of-genes"
"bag-of-genomes"

Type of modeling

Graph-based

Network expansion



Steady-state

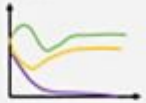
$$S = \begin{pmatrix} +1 & \cdots & 0 \\ \vdots & \ddots & +2 \\ 0 & \cdots & -1 \end{pmatrix}$$

$$S \cdot v = 0$$

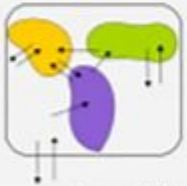
$$v_{min} < v < v_{max}$$

Dynamic models

$$\frac{dX_1}{dt} = \cdots$$



Lumped
'bag-of-genes'



Compartmentalized
'bag-of-genomes'

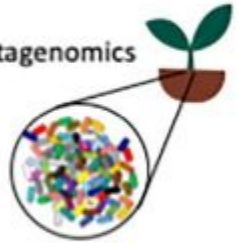
Level of modeling

Applications

- Inference of knowledge – prediction of response
- Selection of communities – keystone species
- Synthetic design – alteration of communities

Top-down

Metagenomics



Bottom-up

Small-scale experiments

Approach

An example: Viral communities

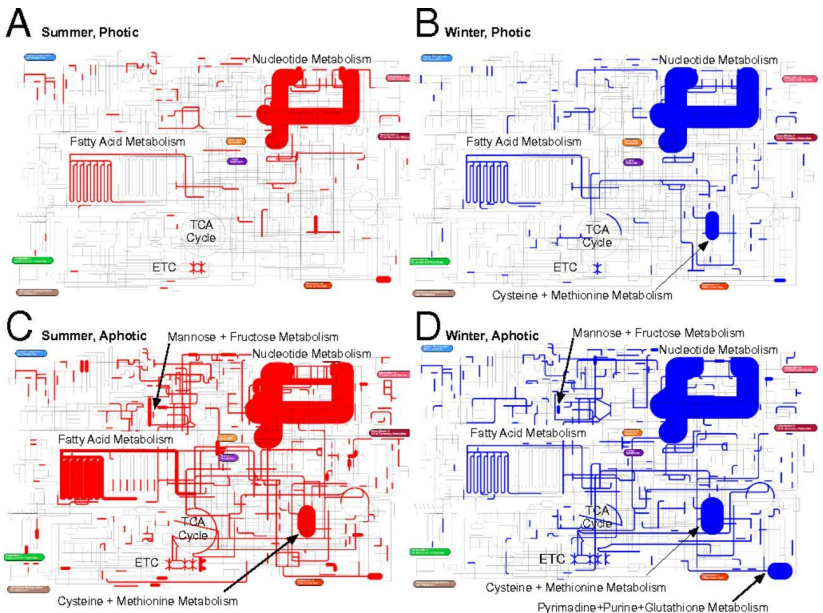
- (<6% of reads matched known viruses)
- **These analyses resulted in (i) a unified comparative network of viromes based on sequence composition (Fig. 1) and (ii) a statistical measure of the effect of covariates (i.e., season, proximity to shore, and depth) on the network structure**
- Looking at Kegg pathways
- Viral encoded host genes
- Summer/winter
- Shallow/deep
- What pathways are present in the population in different conditions

Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses

Bonnie L. Hurwitz, Anton H. Westveld, Jennifer R. Brum, and Matthew B. Sullivan Authors Info & Affiliations

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved June 16, 2014 (received for review October 21, 2013)

July 7, 2014 | 111 (29) 10714-10719 | https://doi.org/10.1073/pnas.1319778111



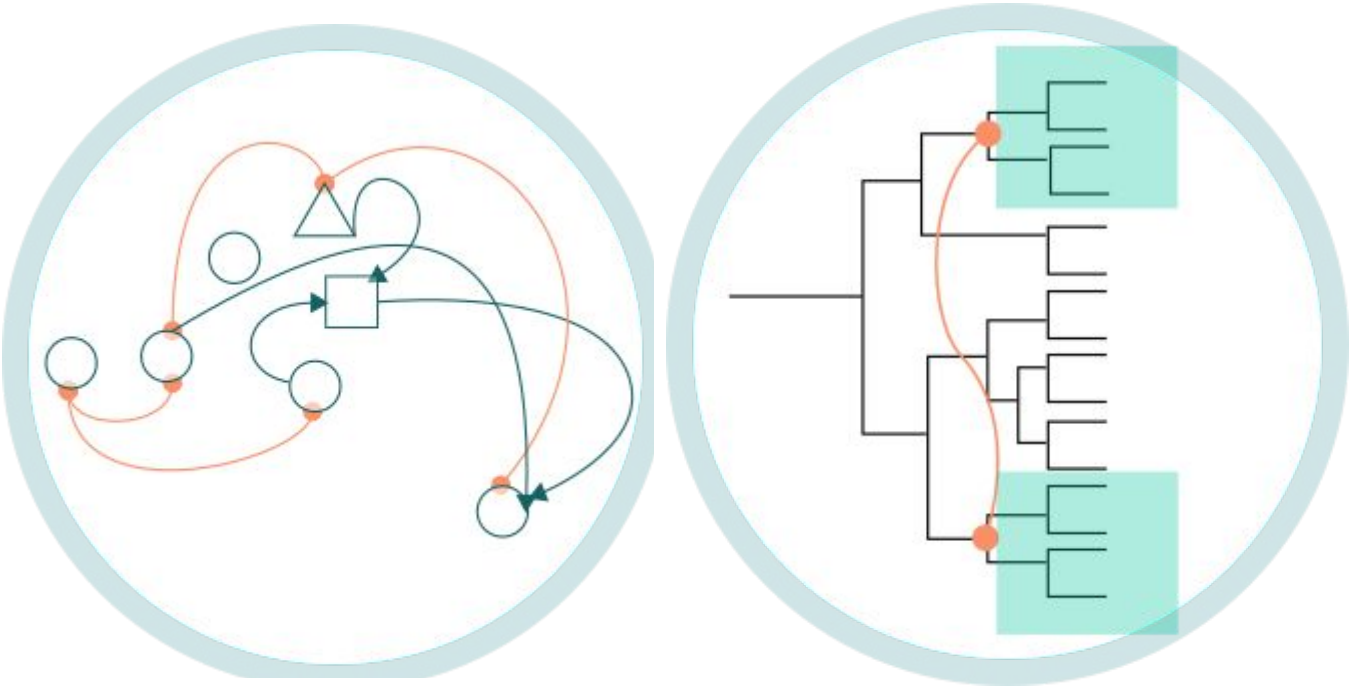
An evolutionary angle: When things emerged

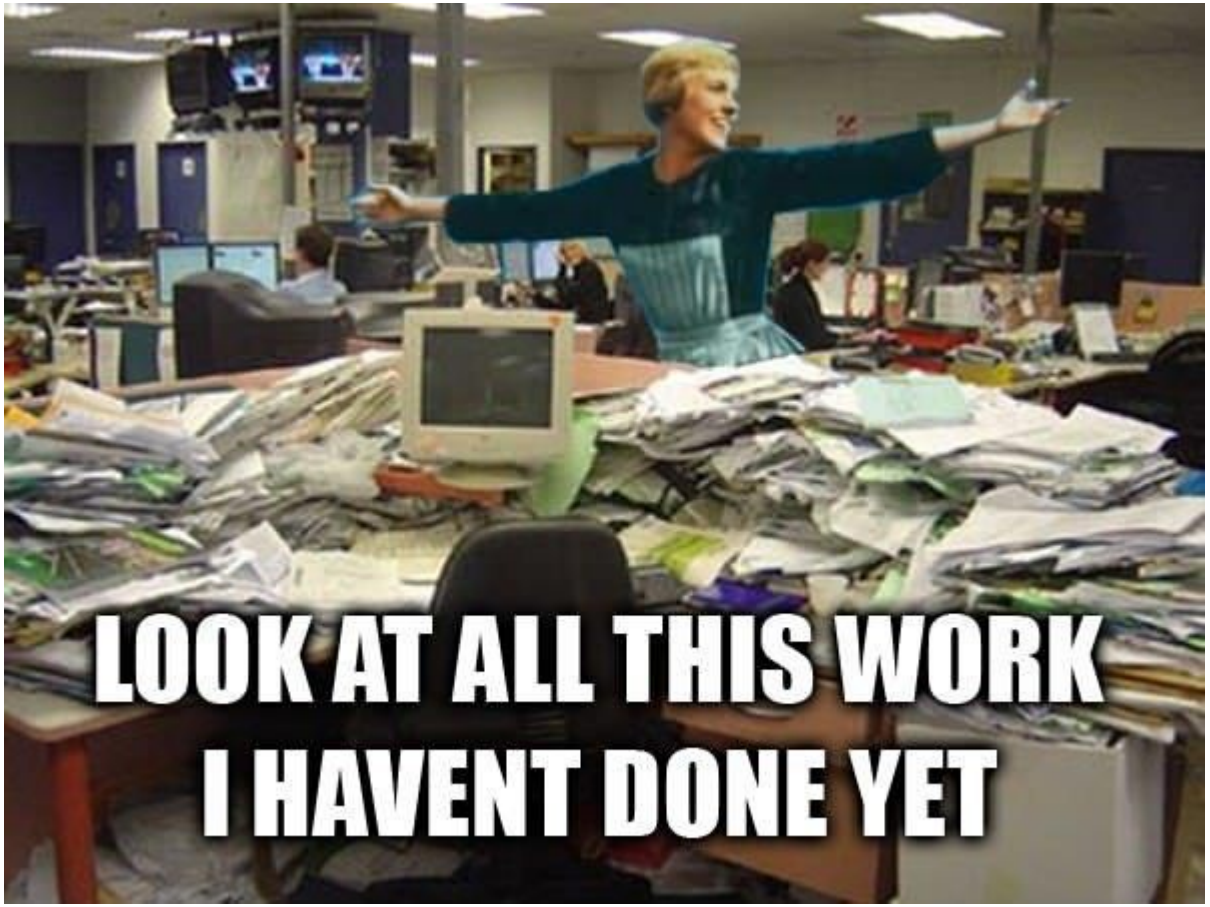
Appears unexplored so far.

When in the species phylogeny did trophic relationships emerge?

How has gene content changed as a consequence?

Are pangenomic entities evolving as a cluster of genes separately?




LOOK AT ALL THIS WORK
I HAVENT DONE YET