CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

iCM Institut de Ciències del Mar

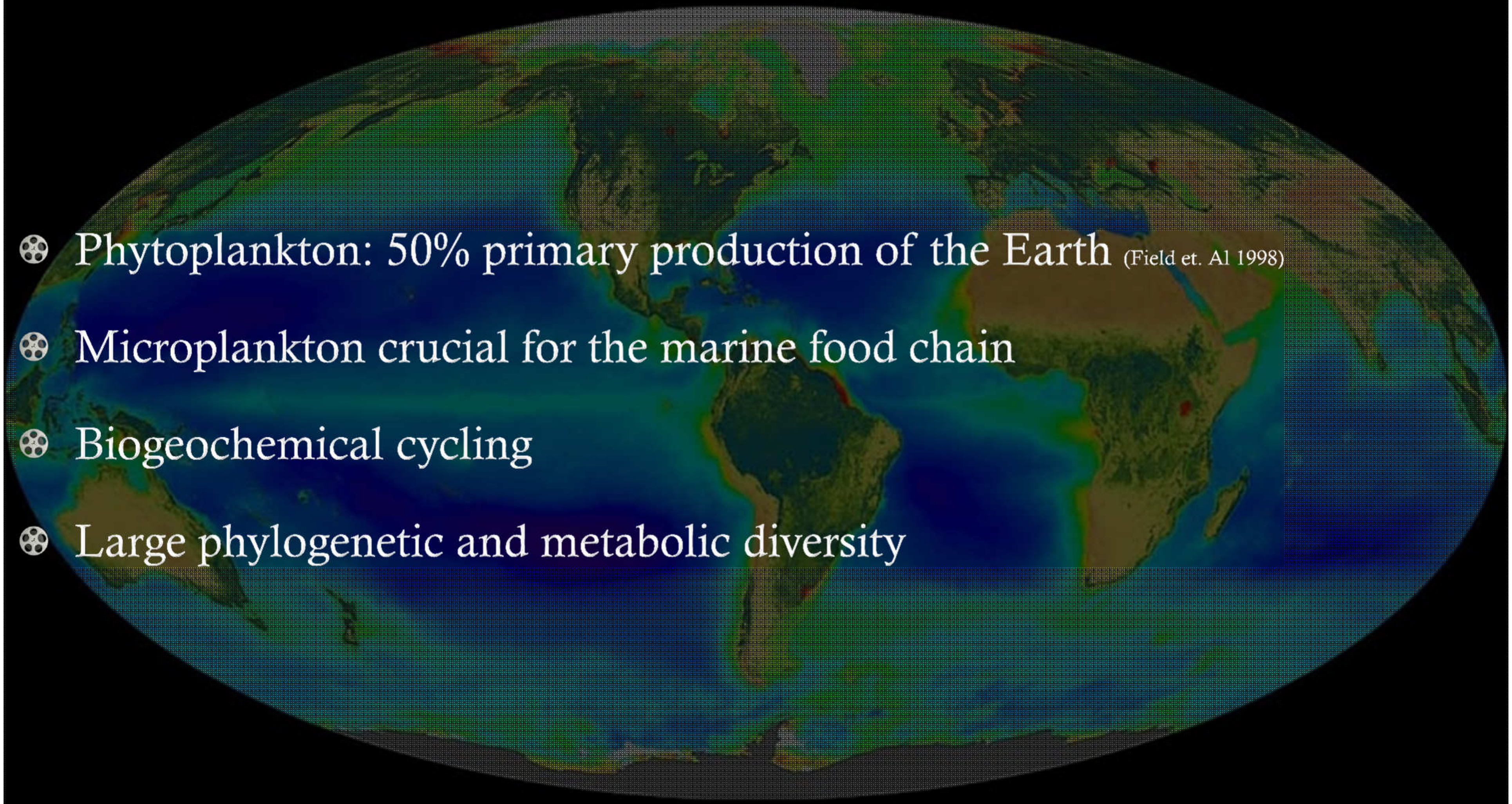
Microbial Ocean*omics* using High-Throughput DNA Sequencing

Ramiro Logares

Institute of Marine Sciences, CSIC, Barcelona

9th RES Users'Conference – 23 September 2015

Importance of microbes in the sunlit ocean

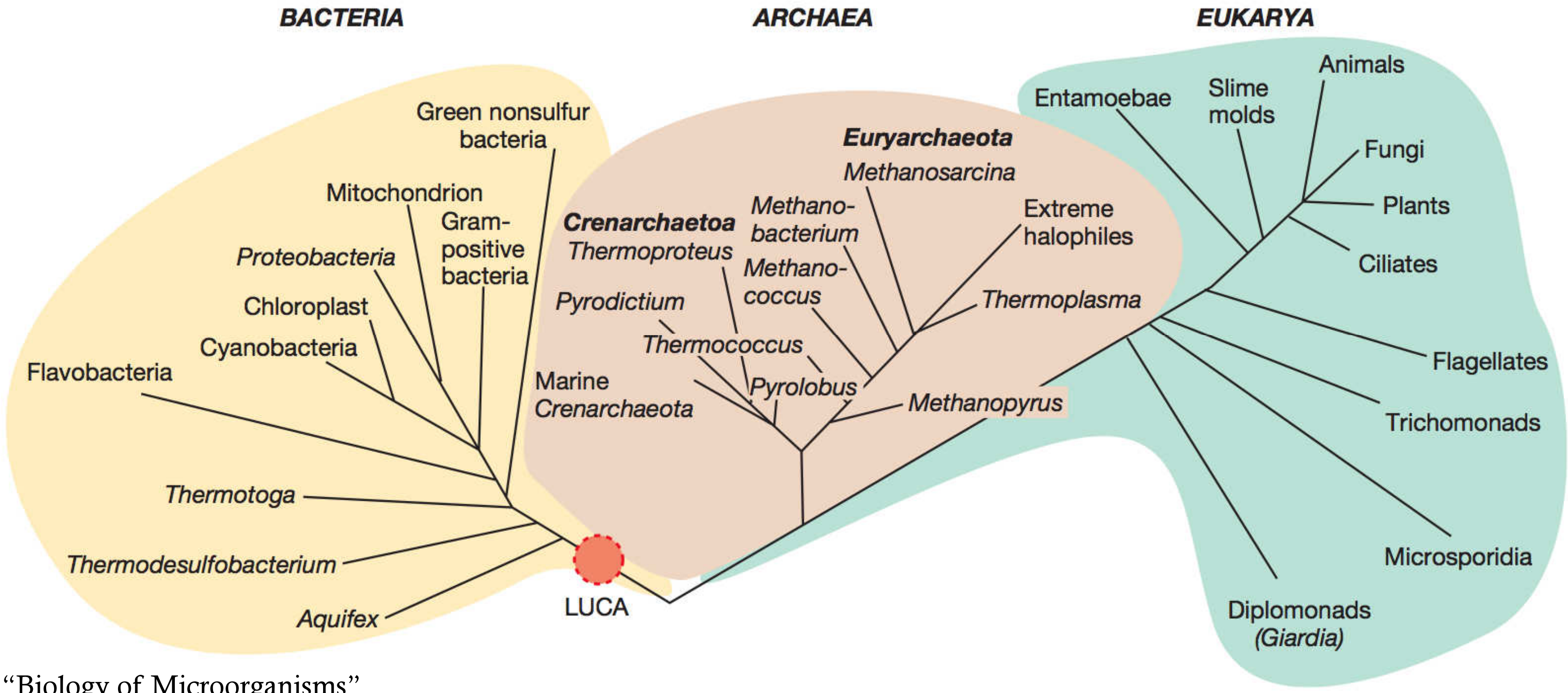


- Phytoplankton: 50% primary production of the Earth (Field et. Al 1998)
- Microplankton crucial for the marine food chain
- Biogeochemical cycling
- Large phylogenetic and metabolic diversity

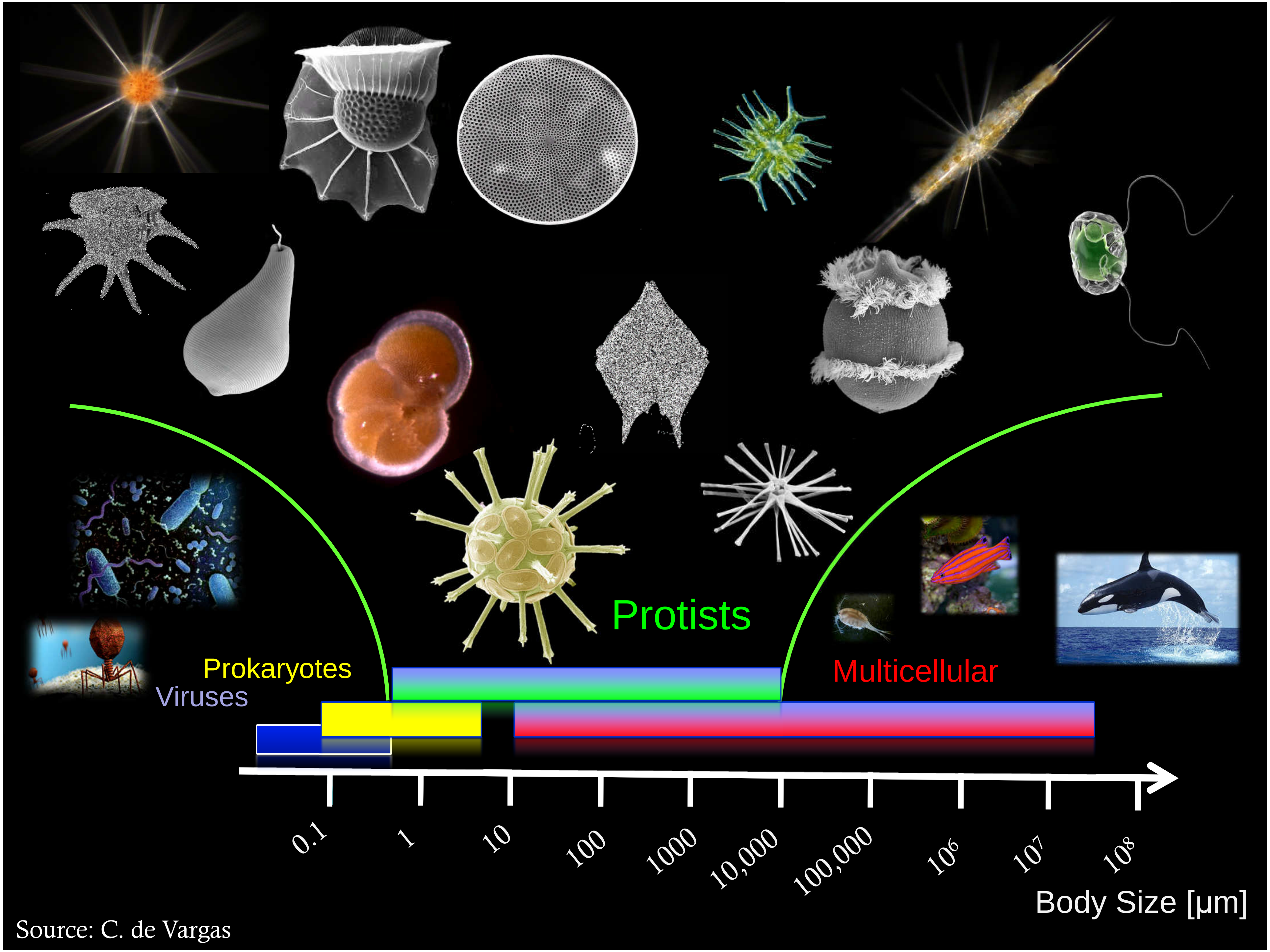
Chlorophyll concentration by SeaWiFS

September 1997 – August 2000

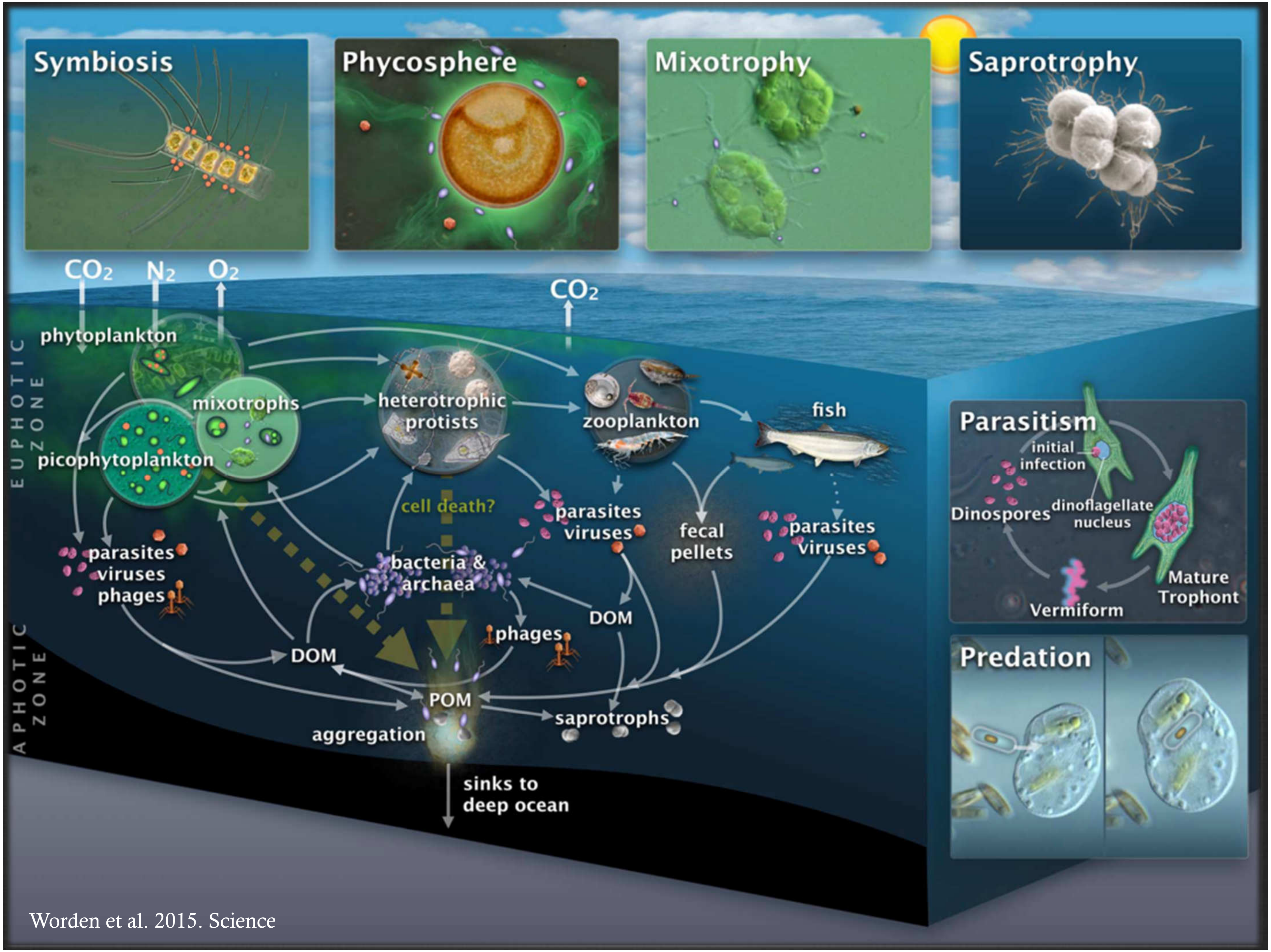
Microbial phylogenetic diversity



Brock "Biology of Microorganisms"

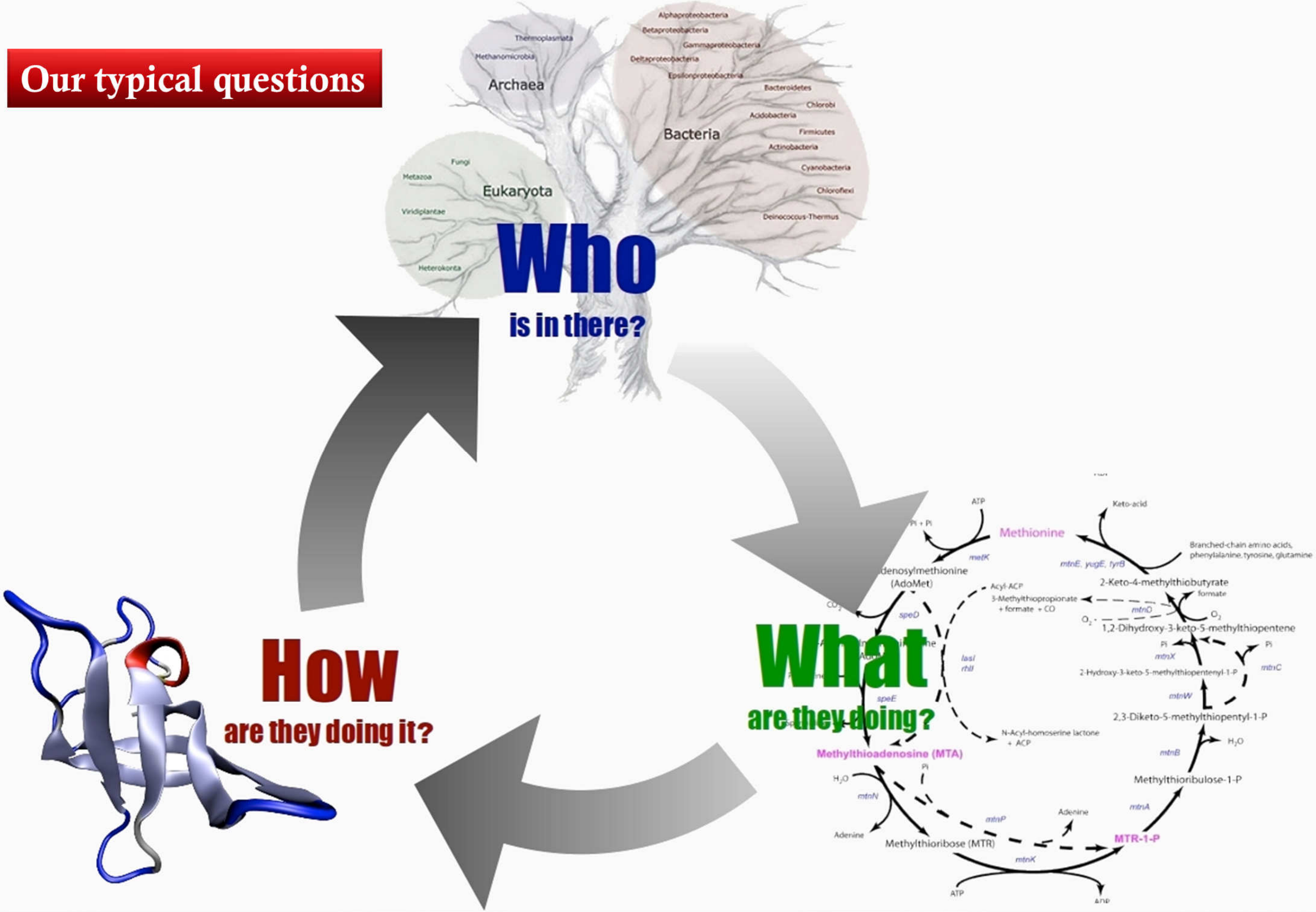


Source: C. de Vargas



Worden et al. 2015. Science

Our typical questions

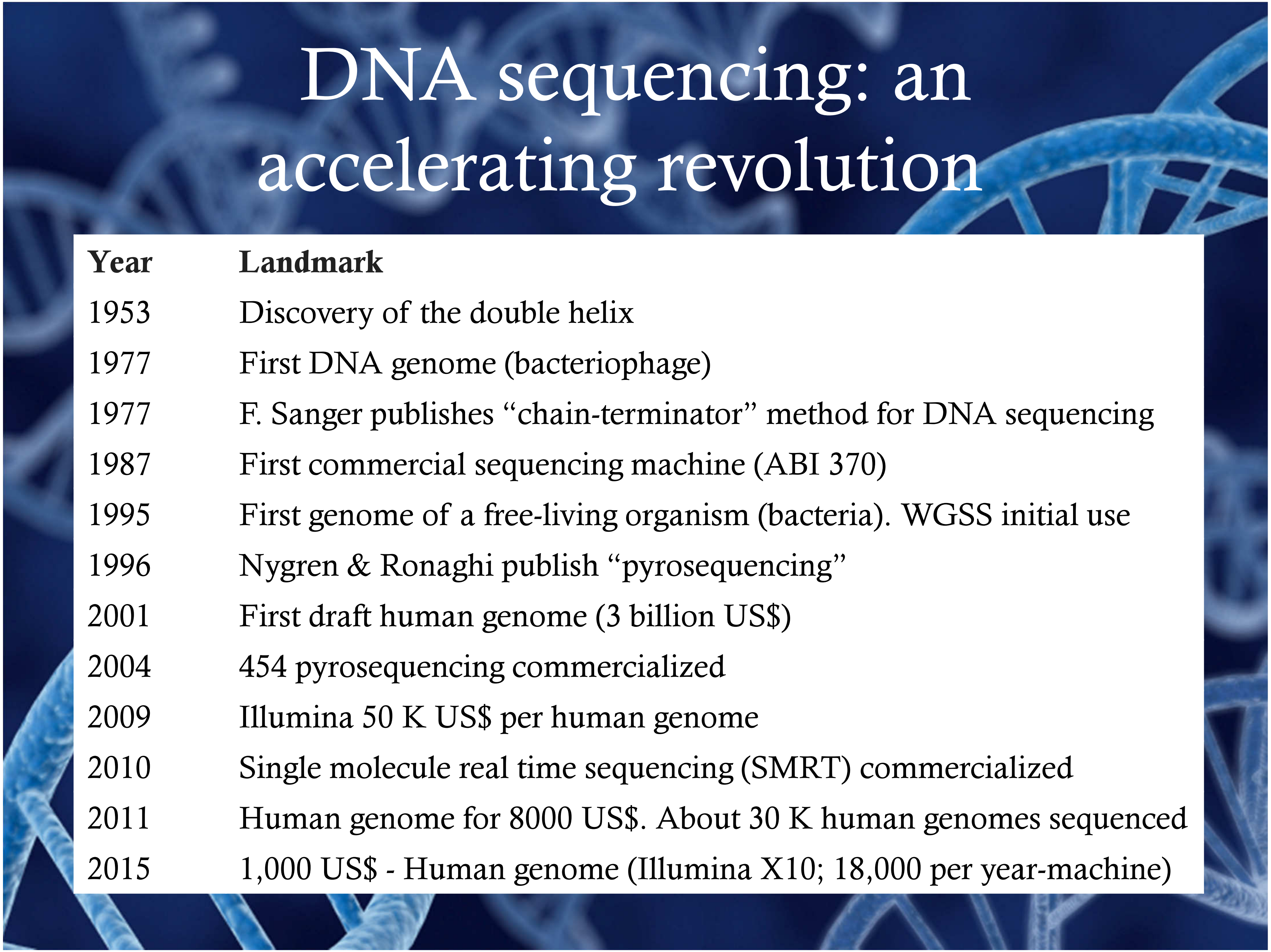


Who is in there?

What are they doing?

How are they doing it?

High throughput DNA sequencing technologies: *a new lens for viewing the microbial world*

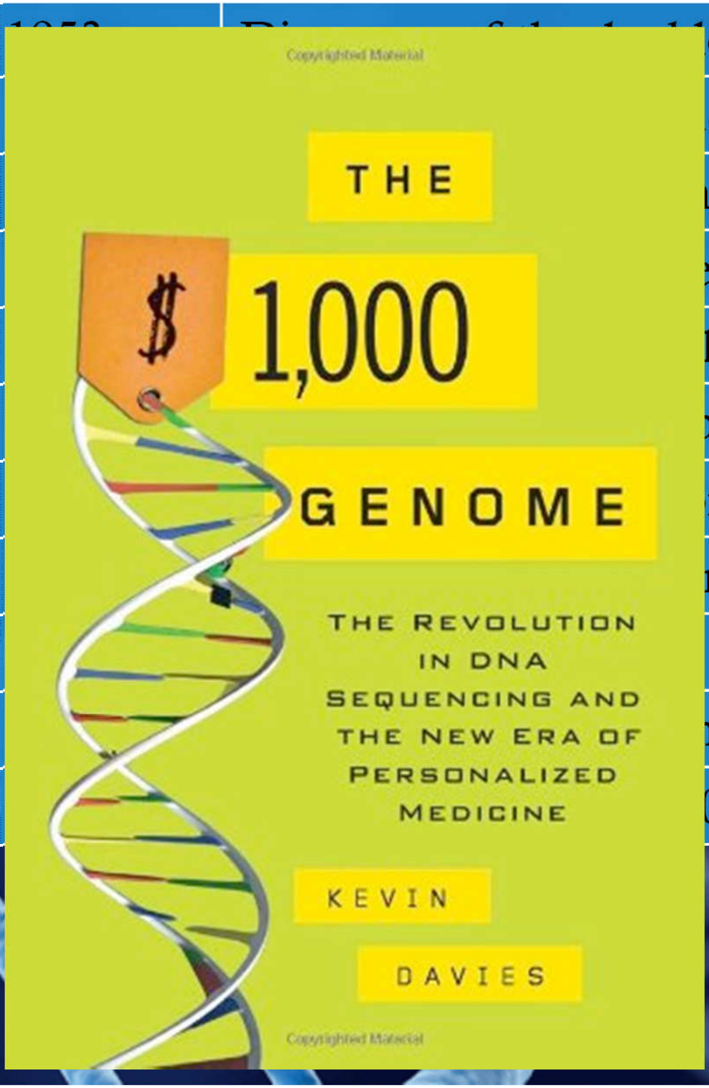


DNA sequencing: an accelerating revolution

Year	Landmark
1953	Discovery of the double helix
1977	First DNA genome (bacteriophage)
1977	F. Sanger publishes “chain-terminator” method for DNA sequencing
1987	First commercial sequencing machine (ABI 370)
1995	First genome of a free-living organism (bacteria). WGSS initial use
1996	Nygren & Ronaghi publish “pyrosequencing”
2001	First draft human genome (3 billion US$)
2004	454 pyrosequencing commercialized
2009	Illumina 50 K US$ per human genome
2010	Single molecule real time sequencing (SMRT) commercialized
2011	Human genome for 8000 US$. About 30 K human genomes sequenced
2015	1,000 US$ - Human genome (Illumina X10; 18,000 per year-machine)

DNA sequencin
revoluti

Year	Landmark
	e helix
	cteriophage
	ain-terminator” method for DNA sequencing
	ncing machine (370)
	living organi). WGSS initial use
	olish “pyr g”
	n e (3)
	m
	h ome
	ne sequencing (SMRT) commercialized
	00 US$. About 30 K human genomes sequenced



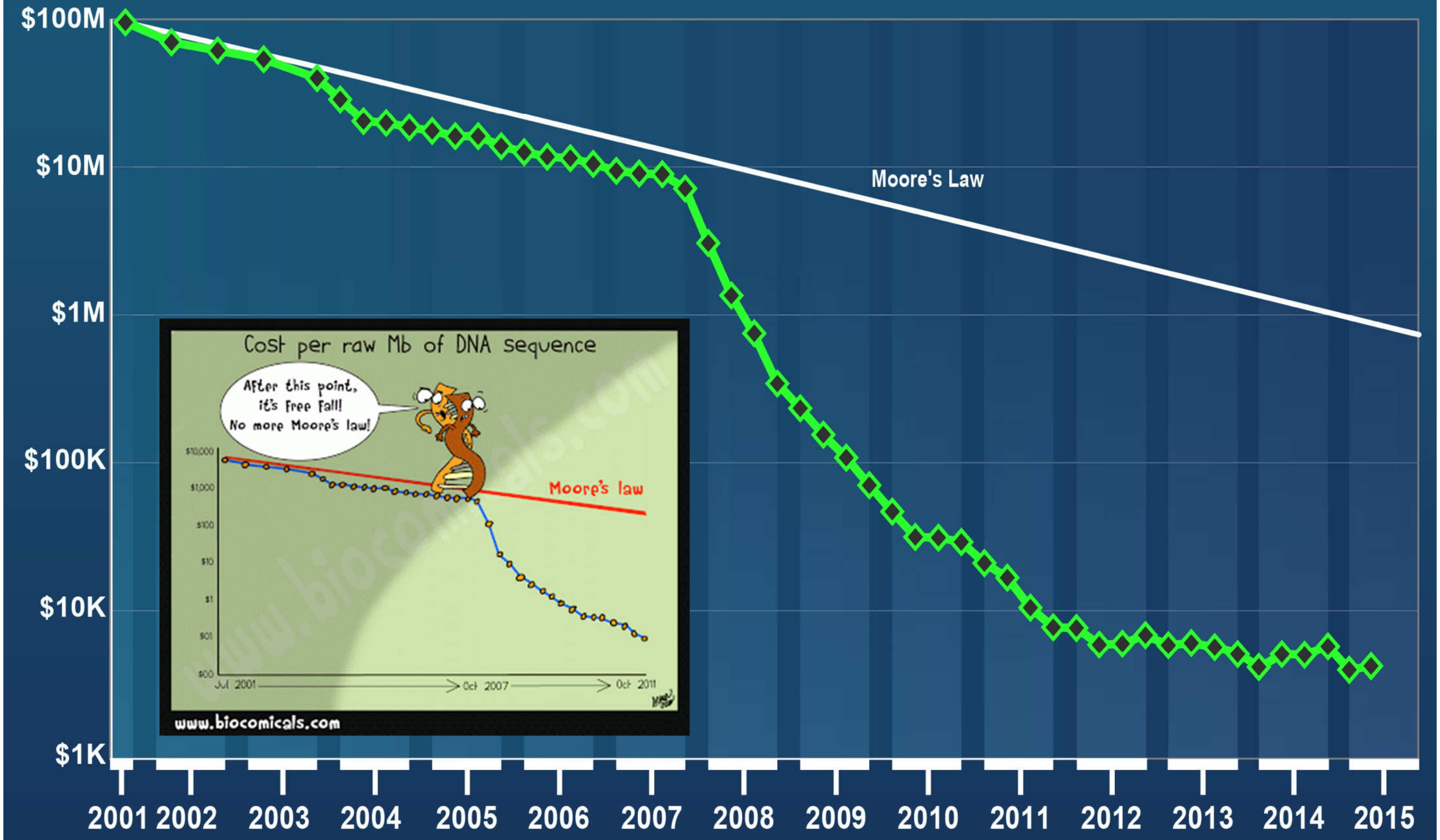
THE $1,000 GENOME

THE REVOLUTION IN DNA SEQUENCING AND THE NEW ERA OF PERSONALIZED MEDICINE

KEVIN DAVIES

62 years

Cost per Genome



Sequencing platforms evolution

1st Generation

Still used for smaller projects or when high quality is needed

Quality reference

2nd Generation

Widely used in most sequencing projects

3rd Generation

Not widely used yet, some devices still not in the market

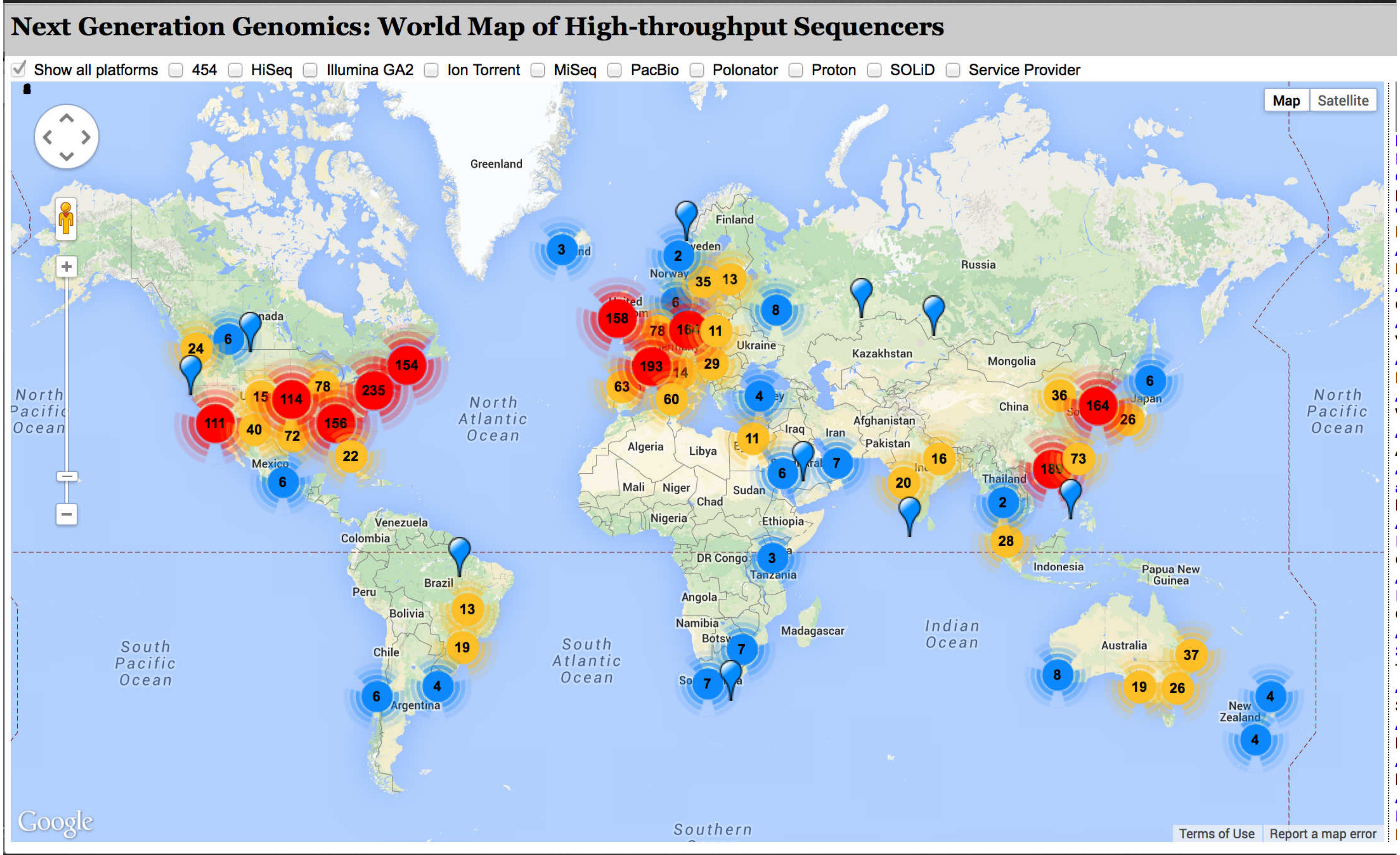
Oxford NANOPORE Technologies

SMRT seq

GridION

MinION

Signal: electricity

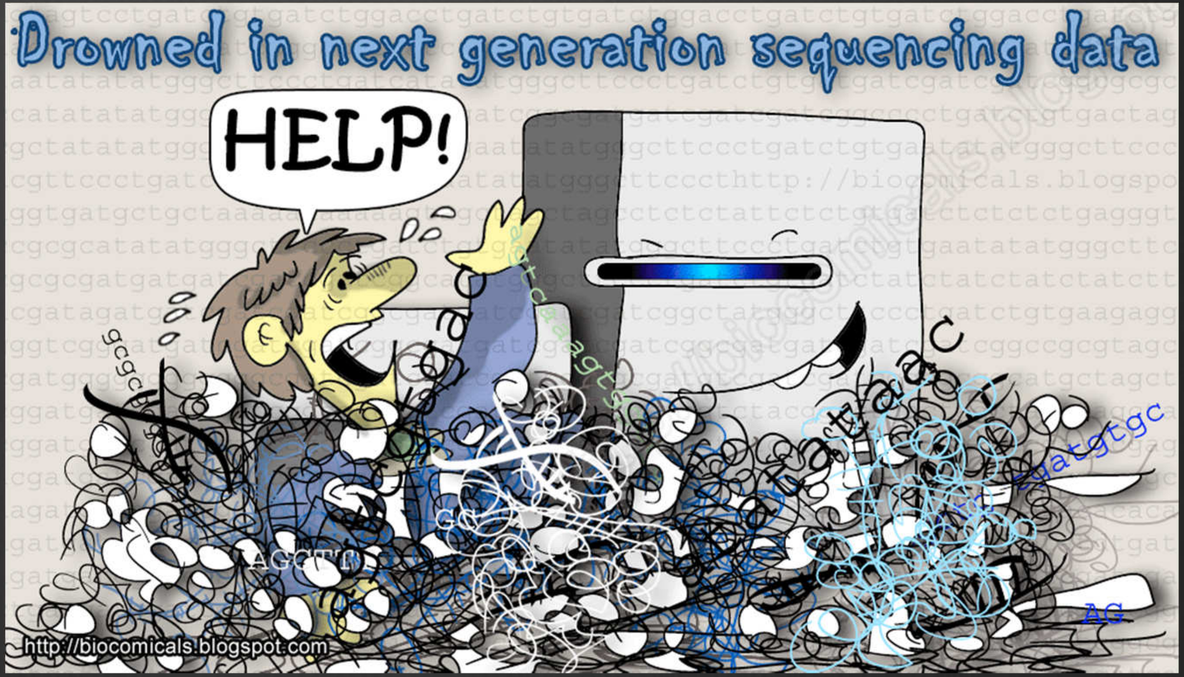


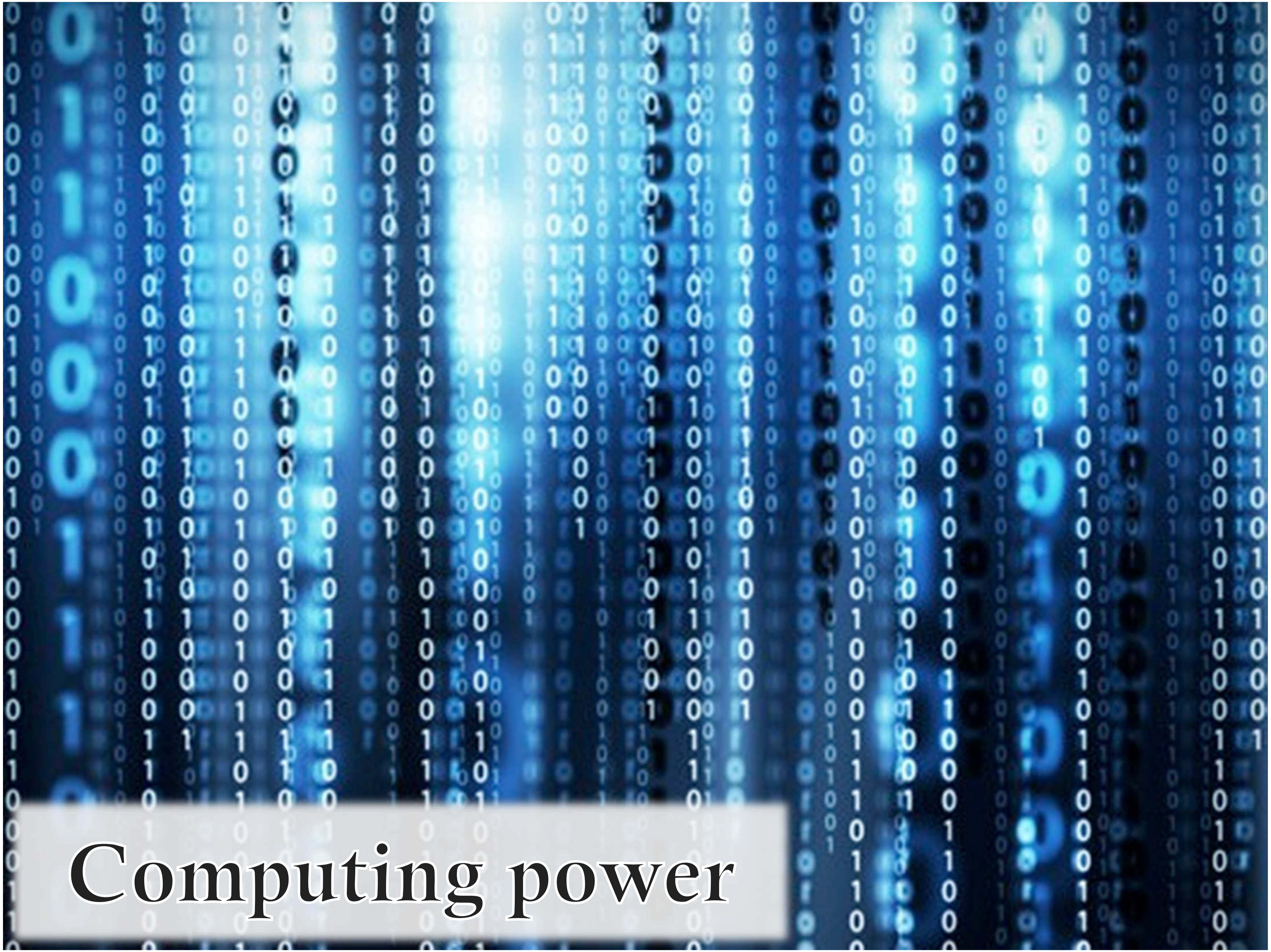
Next Generation Genomics: World Map of High-throughput Sequencers

☑ Show all platforms □ 454 □ HiSeq □ Illumina GA2 □ Ion Torrent □ MiSeq □ PacBio □ Polonator □ Proton □ SOLiD □ Service Provider

http://omicsmaps.com/hts/centres/imppc/

- Amount of HTS DNA data produced now in the world:
  - ≈ 7,389 functional HTS machines
  - ≈35 x 10 ^{15} bases / year == 35 PETAbases
  - 250,000 human genomes per year

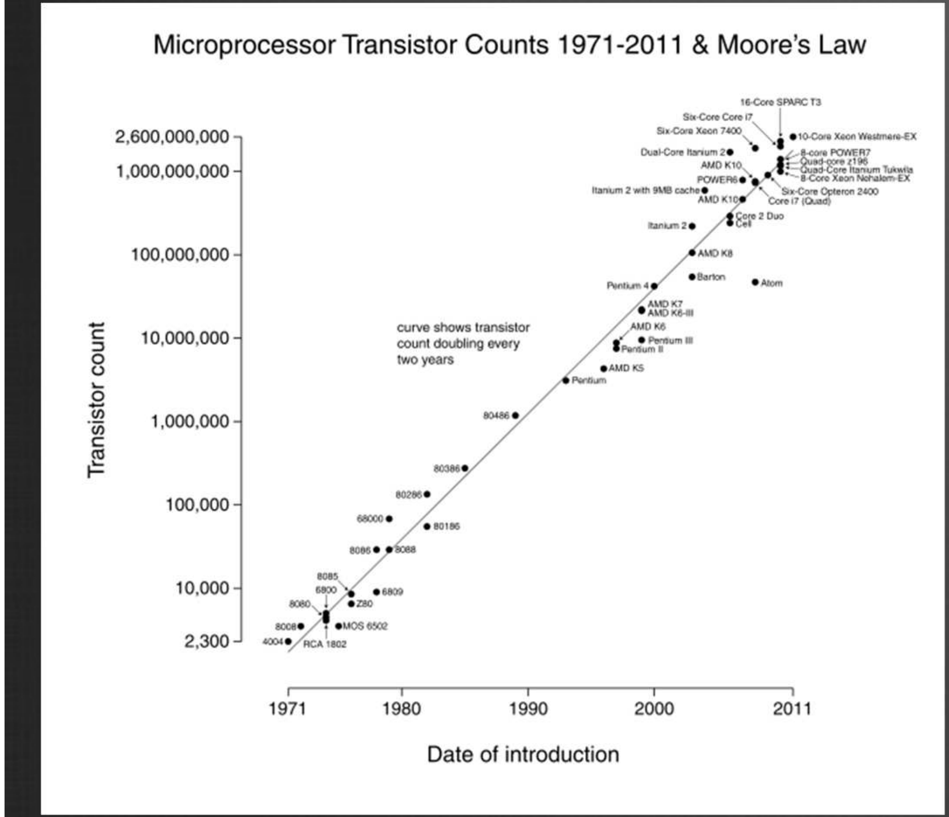

Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com
AG

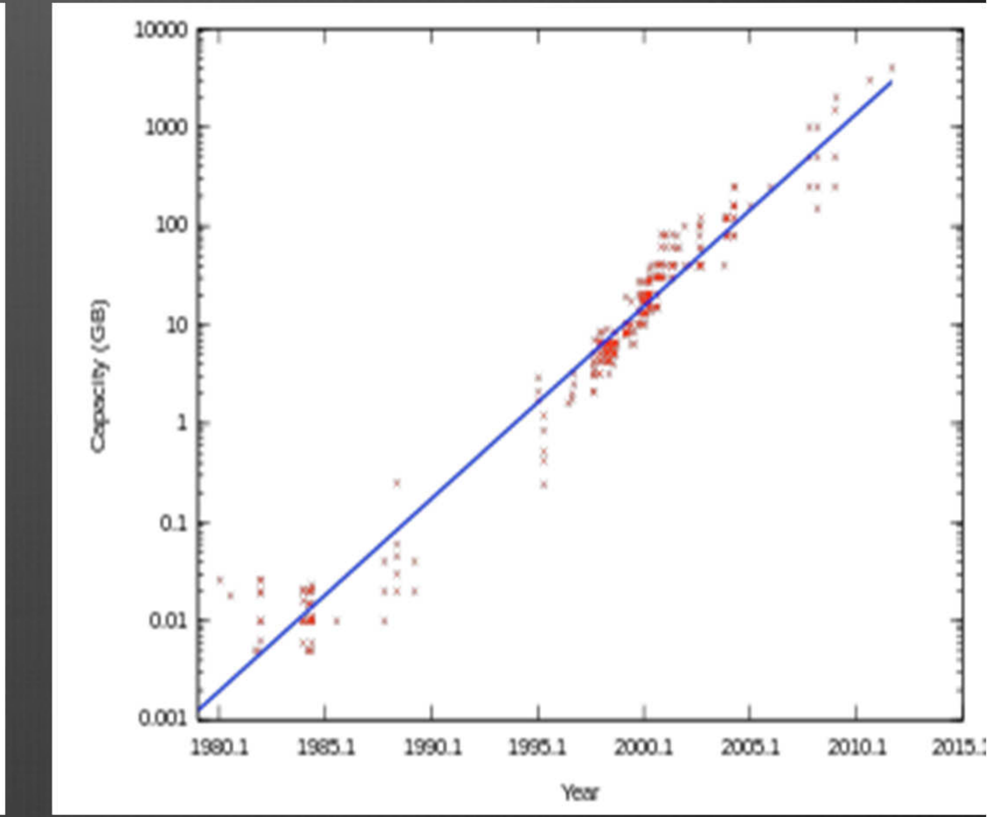


Computing power

Moore's Law

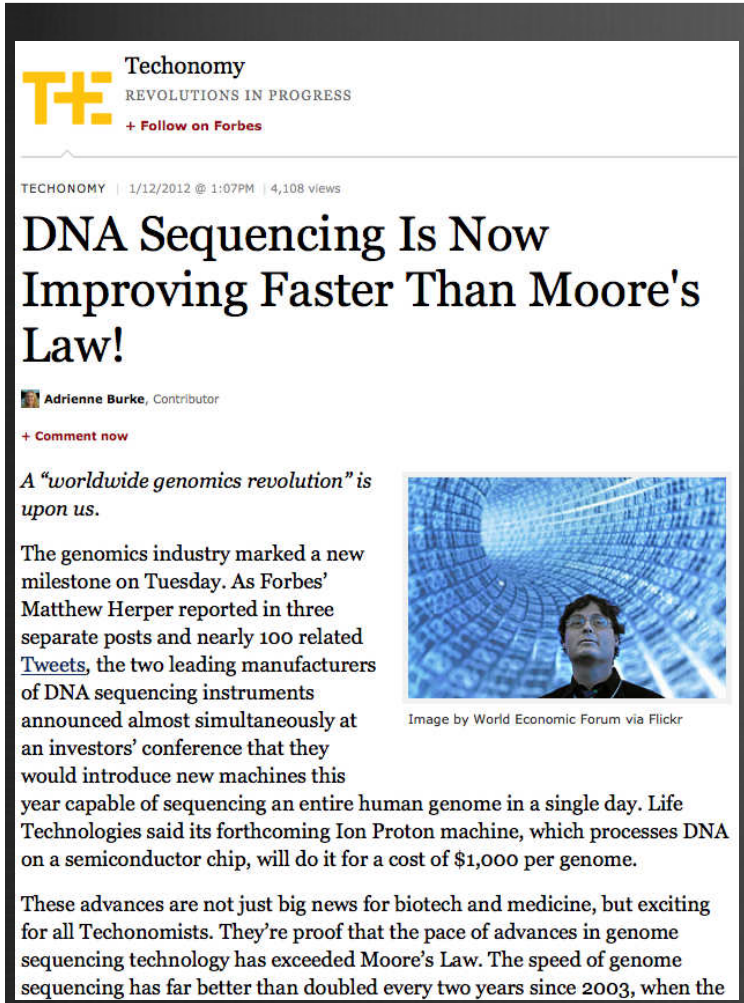
Kryder's Law





Computer power tend to double every two years

Storage capacity doubles annually



Techonomy
REVOLUTIONS IN PROGRESS
+ Follow on Forbes

TECHONOMY | 1/12/2012 @ 1:07PM | 4,108 views

DNA Sequencing Is Now Improving Faster Than Moore's Law!

Adrienne Burke, Contributor

+ Comment now

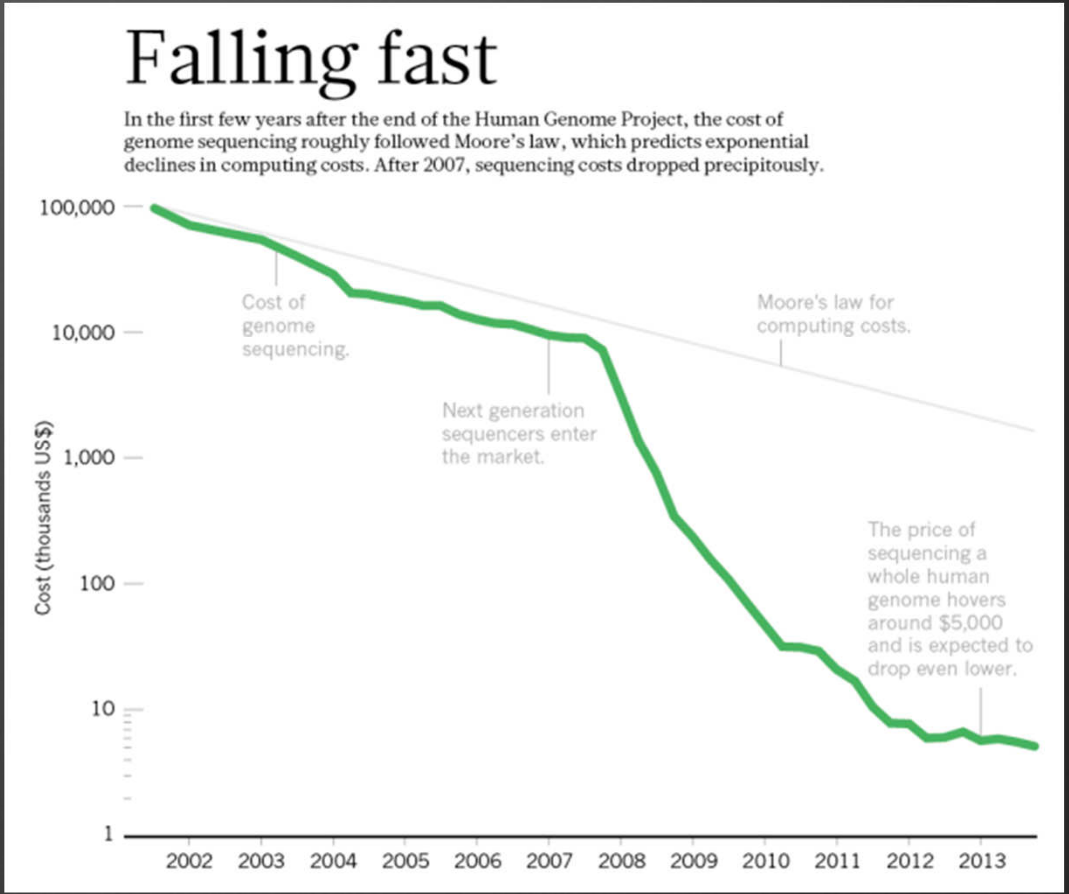
A "worldwide genomics revolution" is upon us.

The genomics industry marked a new milestone on Tuesday. As Forbes' Matthew Herper reported in three separate posts and nearly 100 related Tweets, the two leading manufacturers of DNA sequencing instruments announced almost simultaneously at an investors' conference that they would introduce new machines this year capable of sequencing an entire human genome in a single day. Life Technologies said its forthcoming Ion Proton machine, which processes DNA on a semiconductor chip, will do it for a cost of $1,000 per genome.

Image by World Economic Forum via Flickr

These advances are not just big news for biotech and medicine, but exciting for all Techonomists. They're proof that the pace of advances in genome sequencing technology has exceeded Moore's Law. The speed of genome sequencing has far better than doubled every two years since 2003, when the

Forbes Magazine, 2012



Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

Cost of genome sequencing.

Moore's law for computing costs.

Next generation sequencers enter the market.

The price of sequencing a whole human genome hovers around $5,000 and is expected to drop even lower.

Worldwide sequencing capacity is growing at about 2-3 times per year

- Only one HiSeq2500 produces about 3 TeraBytes of data per month

- Data processing costs should be considered
  - Electricity
  - Costs for data-admin, and reparation

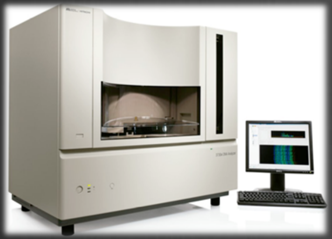
- Amortization (value decrease) of equipment (3yrs CSIC)

- Data storage cost is not trivial

- What to do with used data? And backups? (maybe cheaper re-sequencing than storing?)

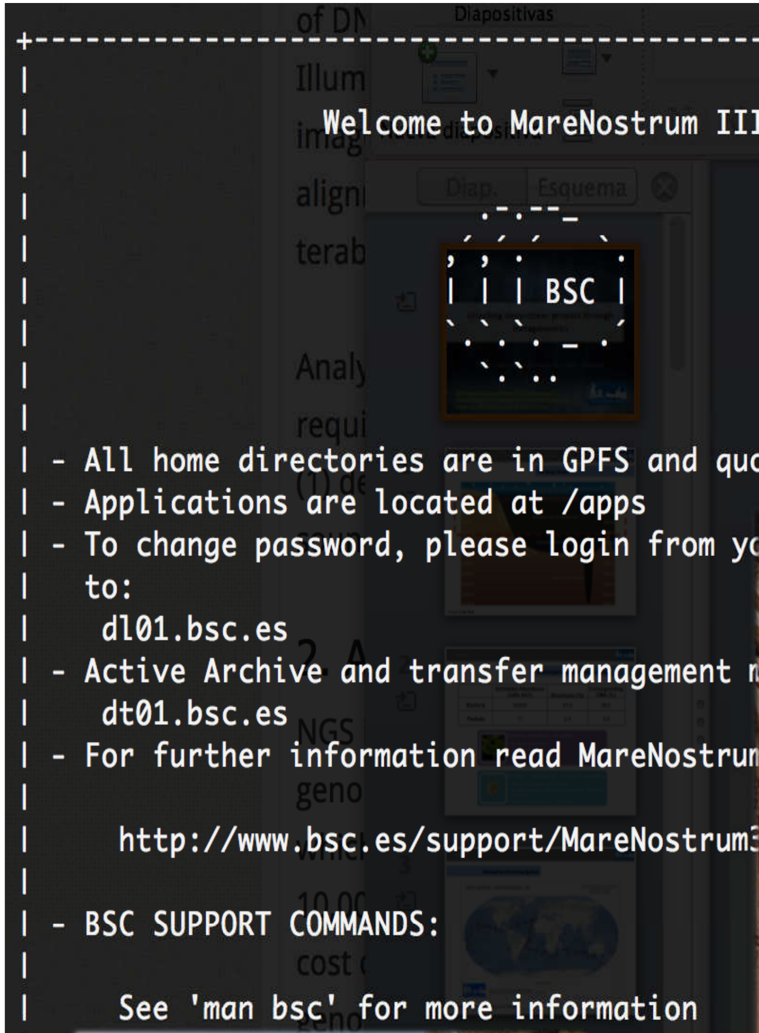
Data processing and computation

Minimum needed computer power

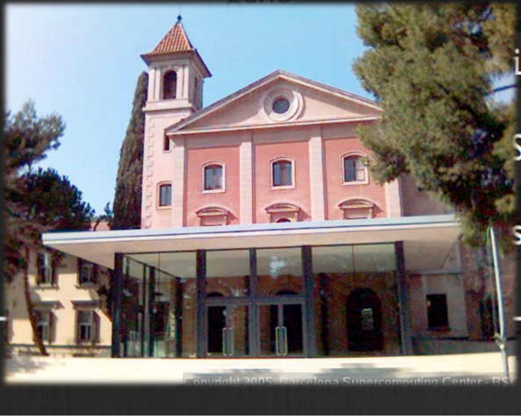


1st Generation	2nd Generation		3rd Generation
SANGER	454 Roche	Illumina HiSeq	PacBio
Cores= 1-2	16-32	>64 (128)	>16 (32)
Mem= 1-4 GB	32-64	>64 (128)	>32 (64)
Disk= 0.2Tb	>2 Tb	> 10 Tb	>1Tb

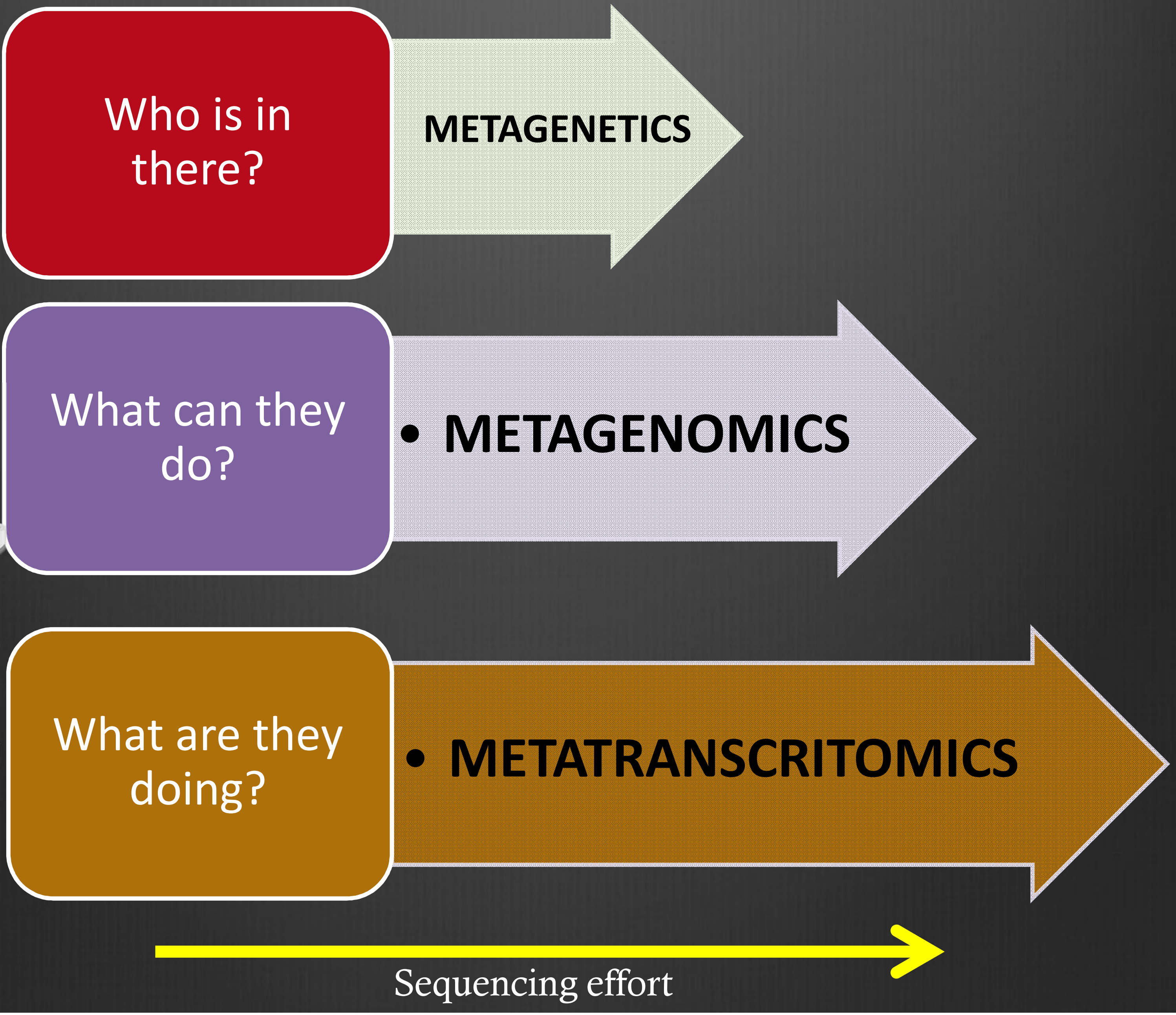
MareNostrum (Barcelona)



```
+-------------------------------------------------------------+
|                                                             |
|              Welcome to MareNostrum III                     |
|                                                             |
|                      .-.--_                                |
|                    ,´,´.´   `.                             |
|                    | | | BSC |                             |
|                    `.`.`. _ .´                             |
|                      `·.-.·´                               |
|                                                             |
| - All home directories are in GPFS and quotas are enabled  |
| - Applications are located at /apps
| - To change password, please login from yo
|   to:
|    dl01.bsc.es
| - Active Archive and transfer management m
|    dt01.bsc.es
| - For further information read MareNostrum
|
|     http://www.bsc.es/support/MareNostrum3
|
| - BSC SUPPORT COMMANDS:
|
|     See 'man bsc' for more information
|
| -                                ion:
|
|                                  SF/9.1.2
|
|                                  sc.es for c
|
+---                               --------------
```



Question in microbial community ecology



Who is in there?

METAGENETICS

What can they do?

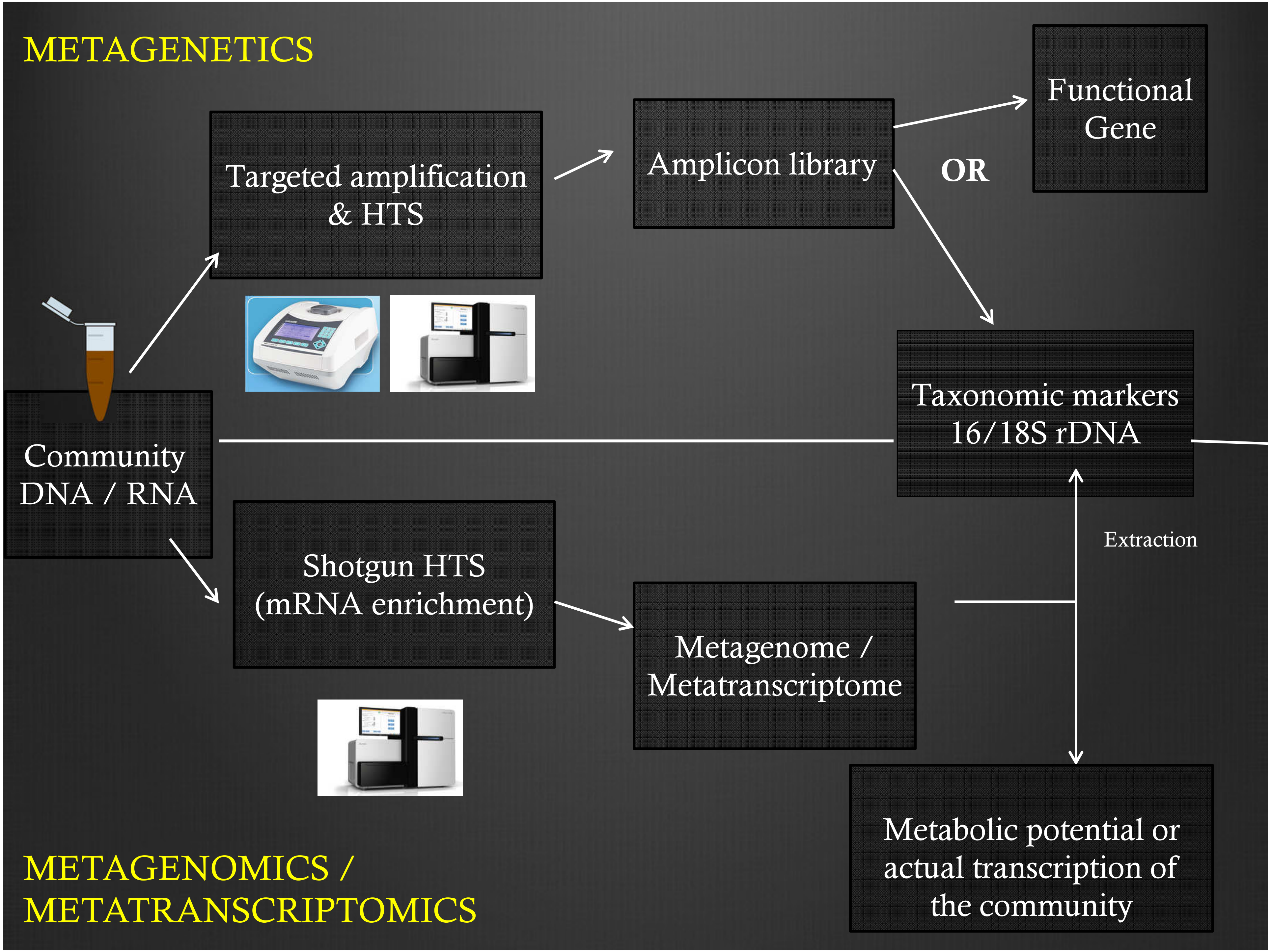
- METAGENOMICS

What are they doing?

- METATRANSCRITOMICS

Sequencing effort

METAGENETICS



Targeted amplification & HTS

Amplicon library

OR

Functional Gene

Taxonomic markers 16/18S rDNA

Community DNA / RNA

Shotgun HTS (mRNA enrichment)

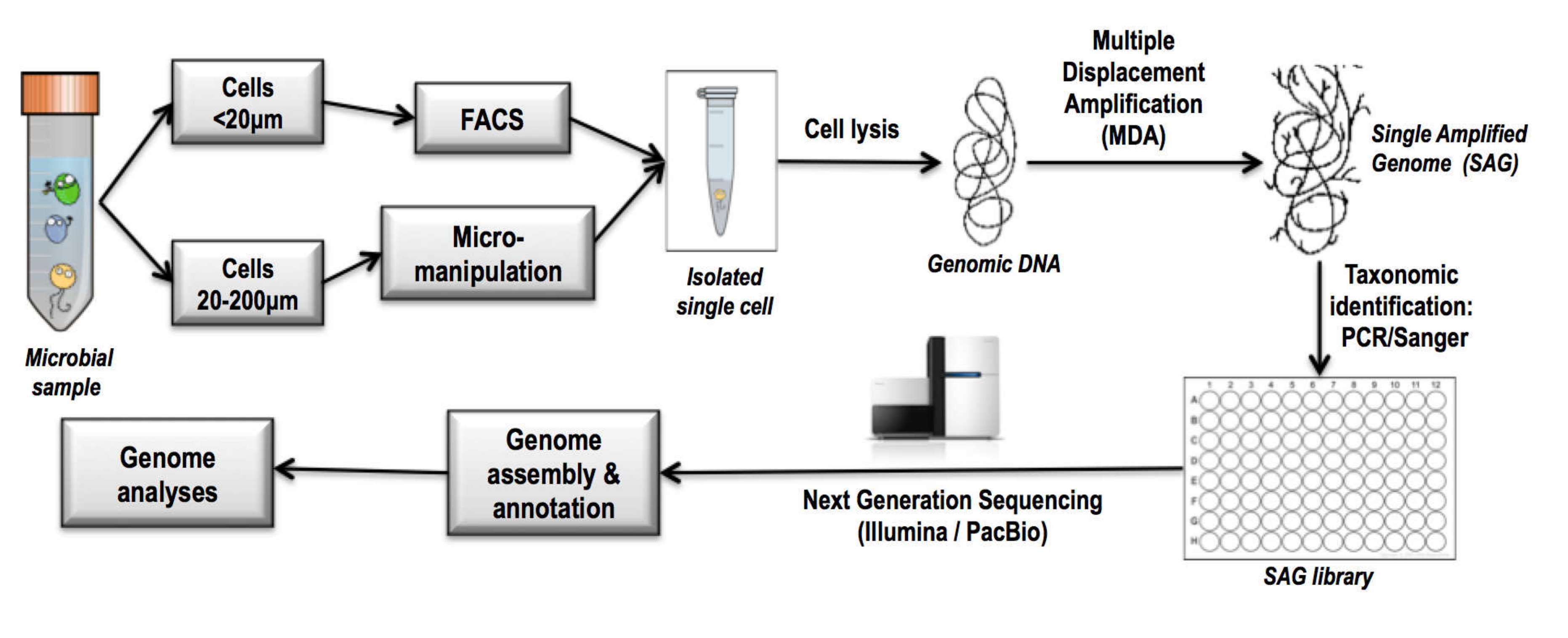
Metagenome / Metatranscriptome

Extraction

Metabolic potential or actual transcription of the community

METAGENOMICS / METATRANSCRIPTOMICS

Alternatively: Single Cell Genomics


Microbial sample
Cells <20µm
Cells 20-200µm
FACS
Micro-manipulation
Isolated single cell
Cell lysis
Genomic DNA
Multiple Displacement Amplification (MDA)
Single Amplified Genome (SAG)
Taxonomic identification: PCR/Sanger
SAG library
Next Generation Sequencing (Illumina / PacBio)
Genome assembly & annotation
Genome analyses

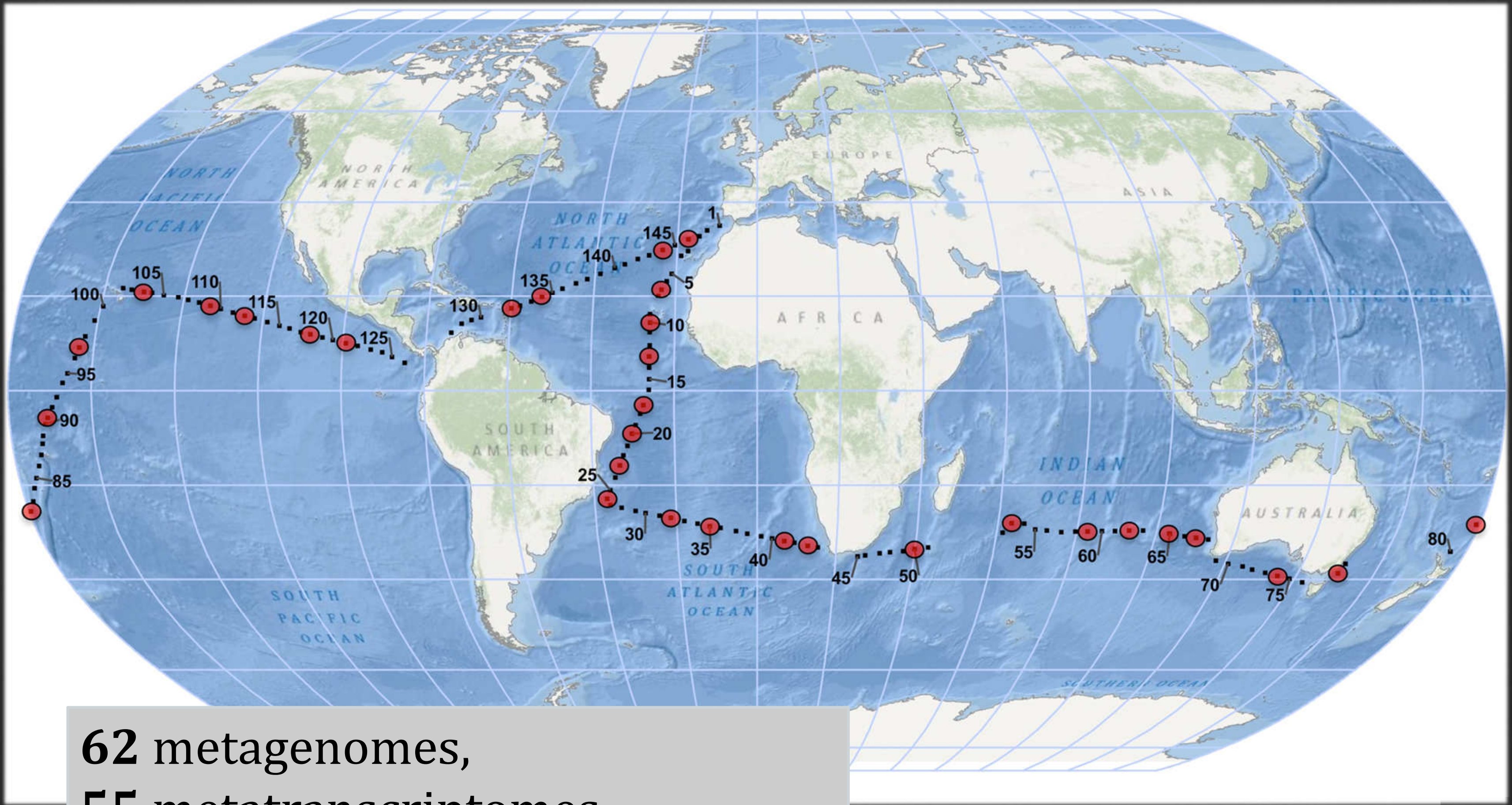
Some results using MareNostrum

Malaspina 2010 expedition

1) Deep-Malaspin*Omics* (□ 4,000m samples)

Sample characteristics: 120 L filtered through two fractions:

- ✓ 0.2-0.8 µm (free-living bacteria and archaea)
- ✓ 0.8-20 µm (protists and attached bacteria)



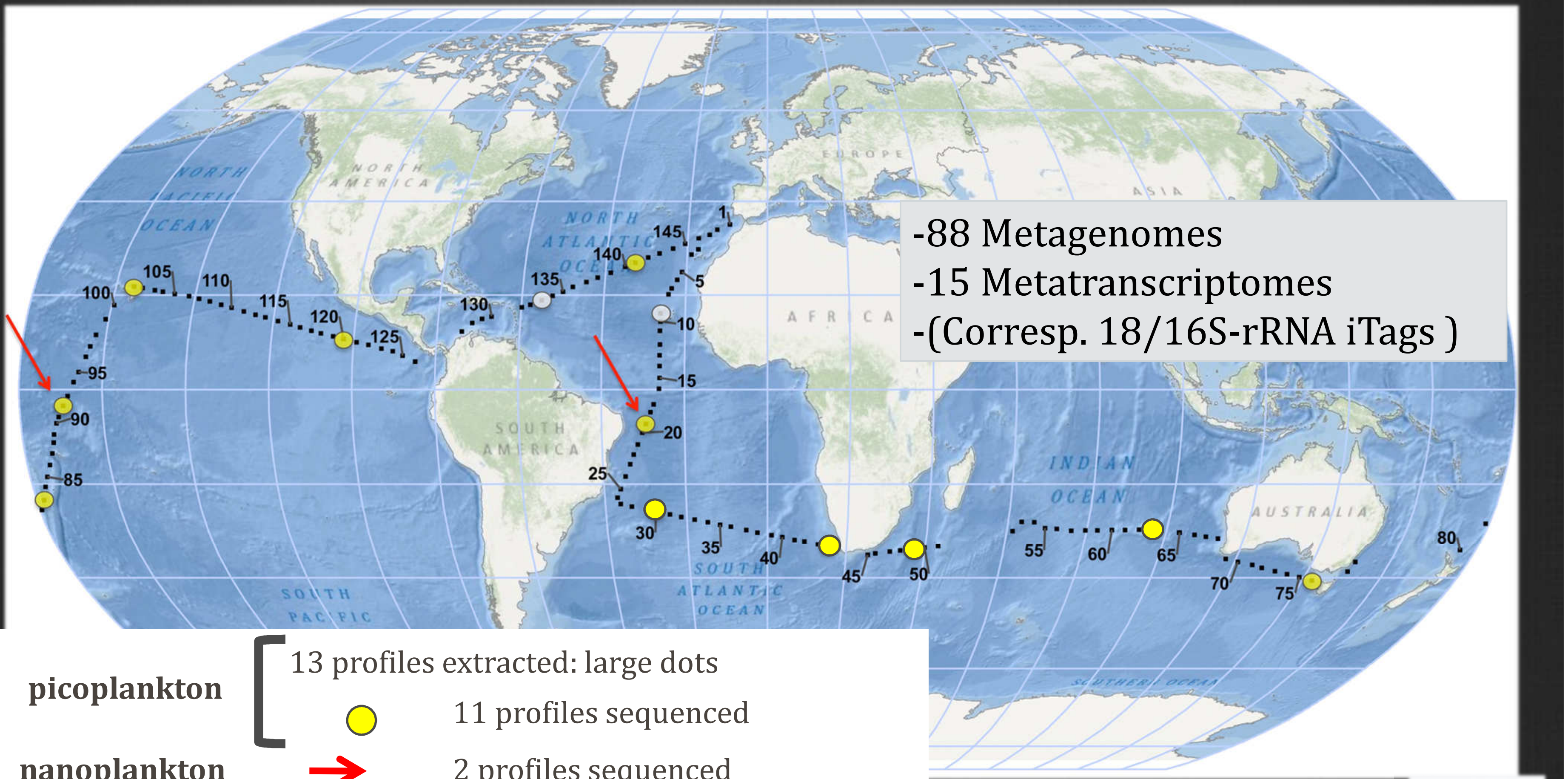
62 metagenomes,
55 metatranscriptomes
Corresp. 18/16S-rRNA i-454Tags

2) Malaspin*Omics* (0 to 4,000m samples)

Sample characteristics:

- ✓ 0.2-3 µm (picoplankton)
- ✓ 3-20 µm (nanoplankton and attached bacteria)

Vertical profiles: 7 depths from surface to 4000m



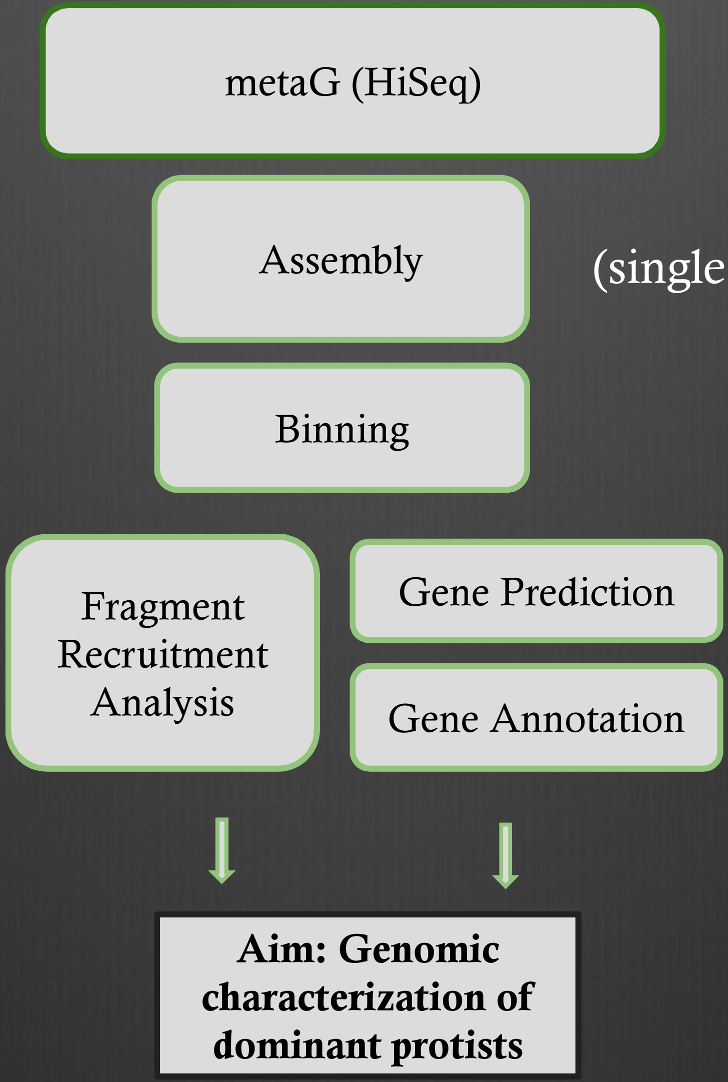
-88 Metagenomes
-15 Metatranscriptomes
-(Corresp. 18/16S-rRNA iTags)

picoplankton [13 profiles extracted: large dots

● 11 profiles sequenced

nanoplankton → 2 profiles sequenced

Metagenomes



metaG (HiSeq)

Assembly

(single / multiple metaG)

Binning

Fragment Recruitment Analysis

Gene Prediction

Gene Annotation

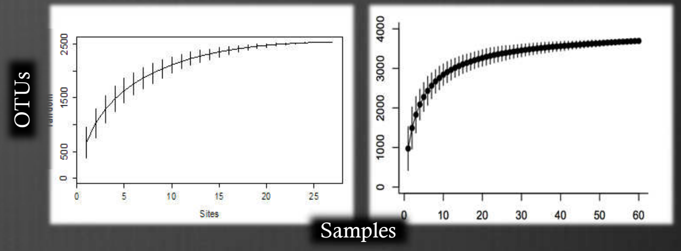
Aim: Genomic characterization of dominant protists

Metagenomes

4,000m samples

Assembly per sample ➔ eukaryotic few contigs

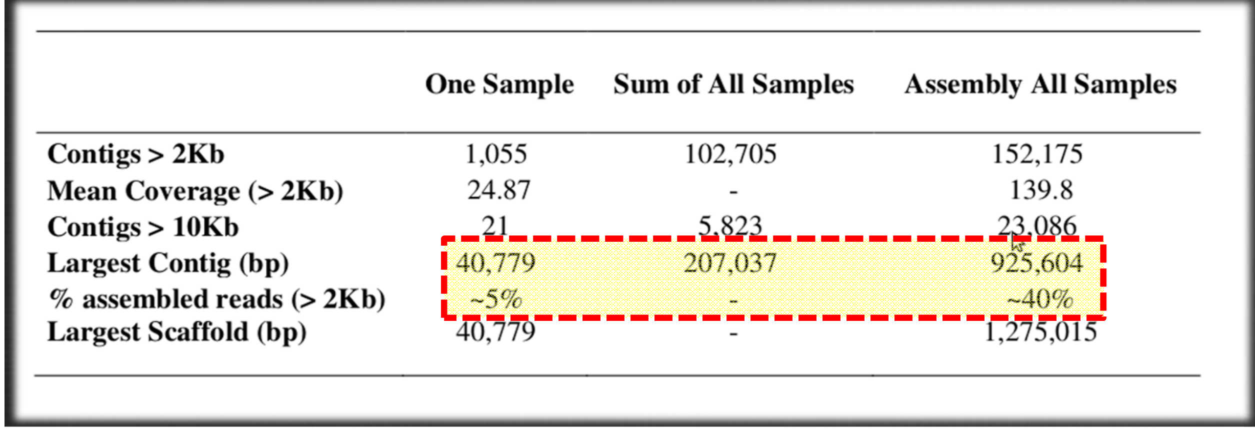
…but, we knew the same OTUs were found in several samples



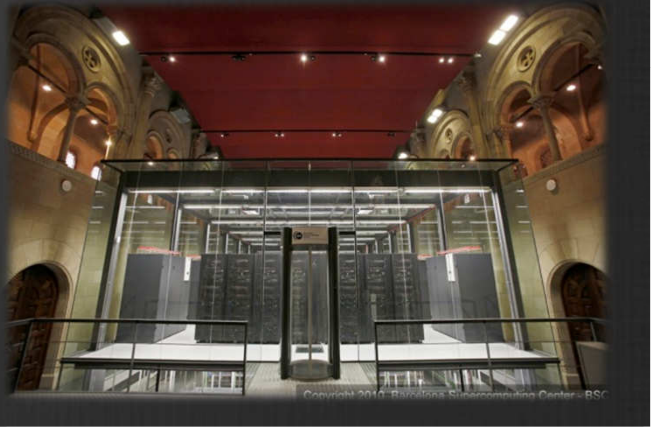
Co-Assembling all samples together (58 metagenomes) should generate longer contigs

Co-assembly of all samples generated longer contigs

MareNostrum Supercomputer
2,048 processors with
Ray assembler



	One Sample	Sum of All Samples	Assembly All Samples
Contigs > 2Kb	1,055	102,705	152,175
Mean Coverage (> 2Kb)	24.87	-	139.8
Contigs > 10Kb	21	5,823	23,086
Largest Contig (bp)	40,779	207,037	925,604
% assembled reads (> 2Kb)	~5%	-	~40%
Largest Scaffold (bp)	40,779	-	1,275,015



Metagenomes

1,500- 4,000m

Co-assembly of 1,500-4,000m samples:

- 58 | 4,000m (5Gb each)
- 29 | 1,500-4,000m (20-40 Gb each)

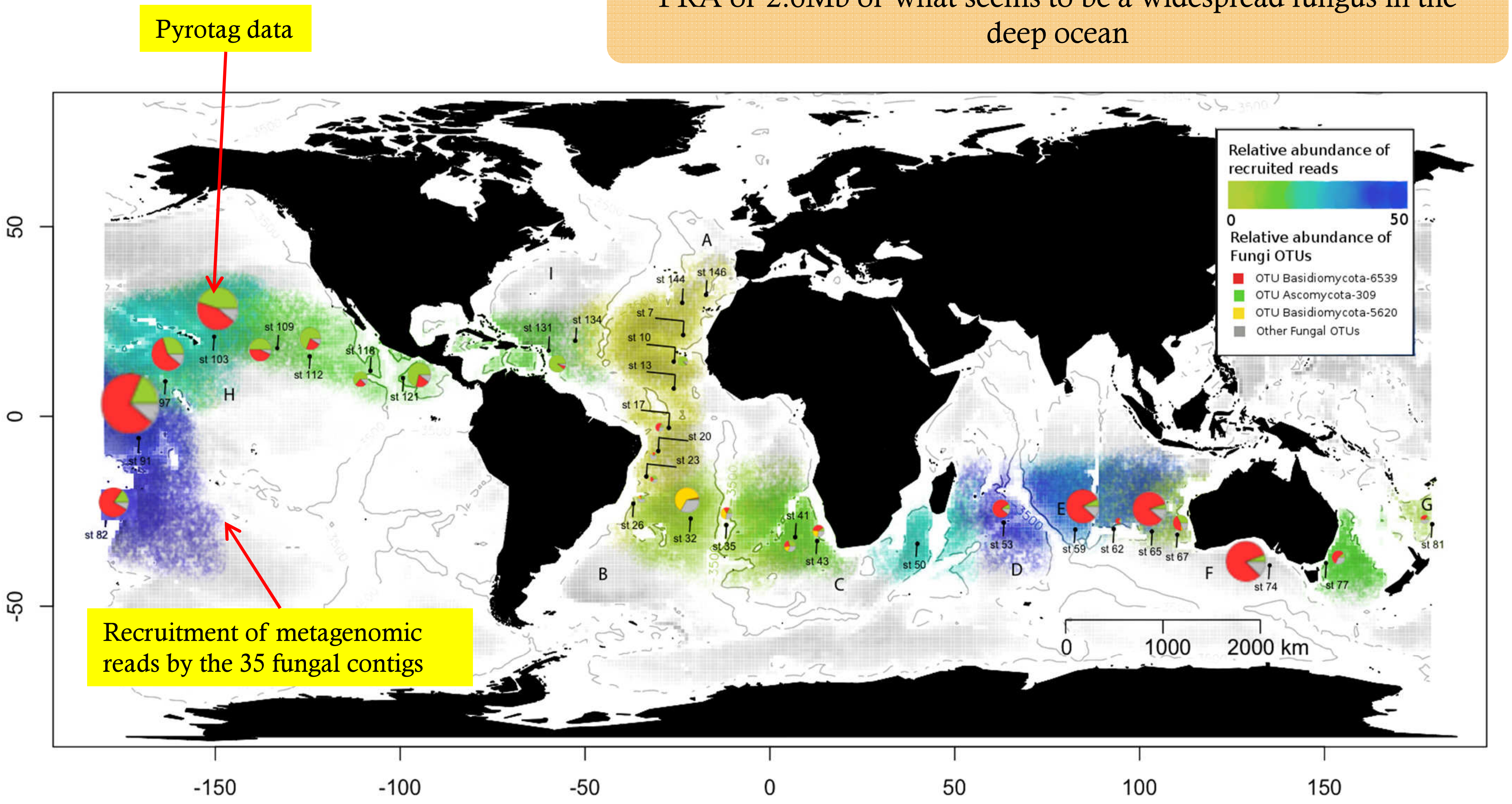
- Ray assembly with 2,048 threads @MN (18hs)

	4,000m	**1,500 - 4,000m**
# Contigs > 2kb	152,175	339,898
Largest contig (bp)	925,604	1,119,237

Metagenomes: deep ocean fungi

4,000m samples

FRA of 2.6Mb of what seems to be a widespread fungus in the deep ocean

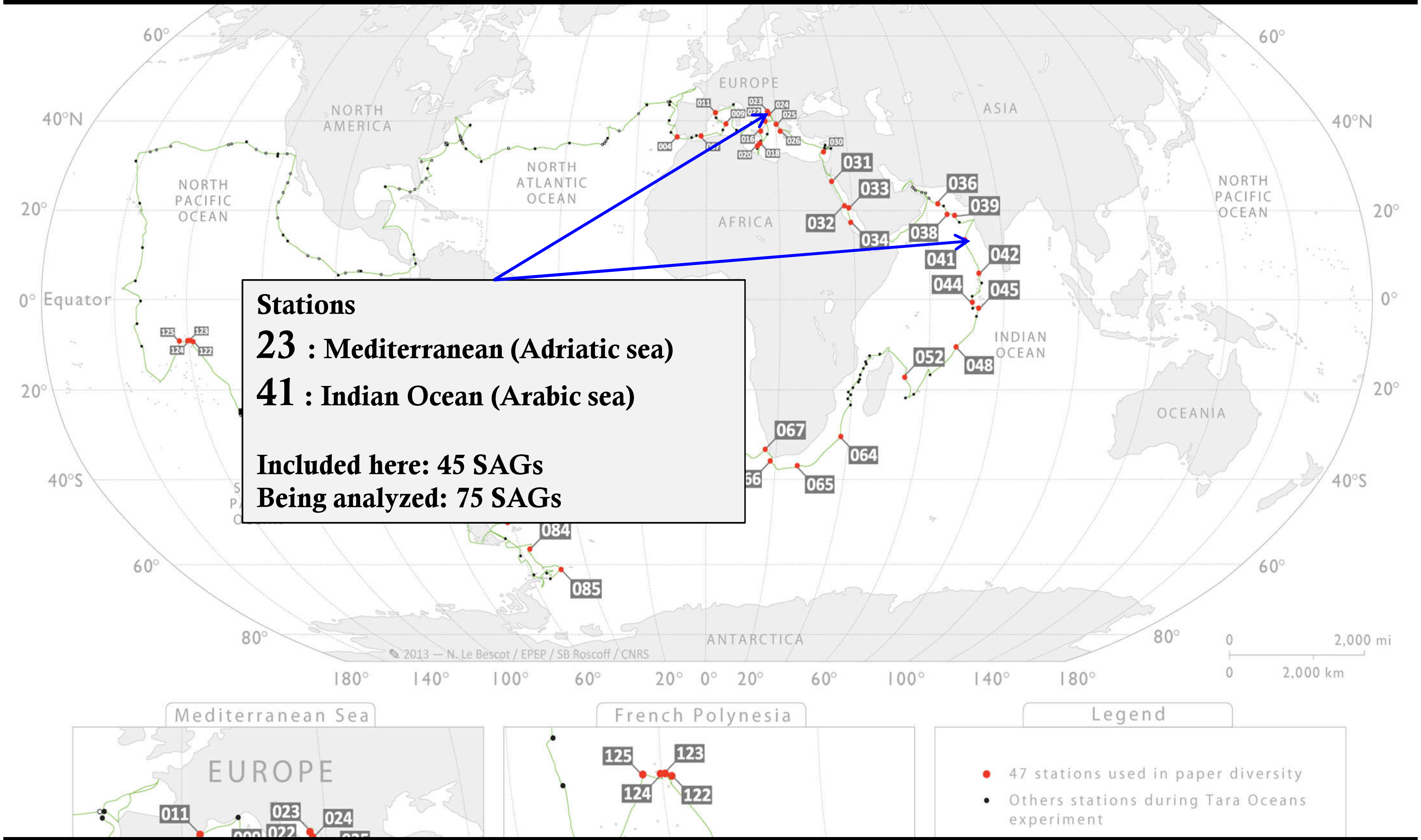


Pyrotag data

Recruitment of metagenomic reads by the 35 fungal contigs

TARA
OCEANS
tara

Analyzed Pico–euk SAGs from TARA



Stations

23 : Mediterranean (Adriatic sea)

41 : Indian Ocean (Arabic sea)

Included here: 45 SAGs
Being analyzed: 75 SAGs

A large collection of SAGs is being generated. So far, 903 SAGs were identified, with 568 affiliating to Stramenopiles

Co-assembly stats

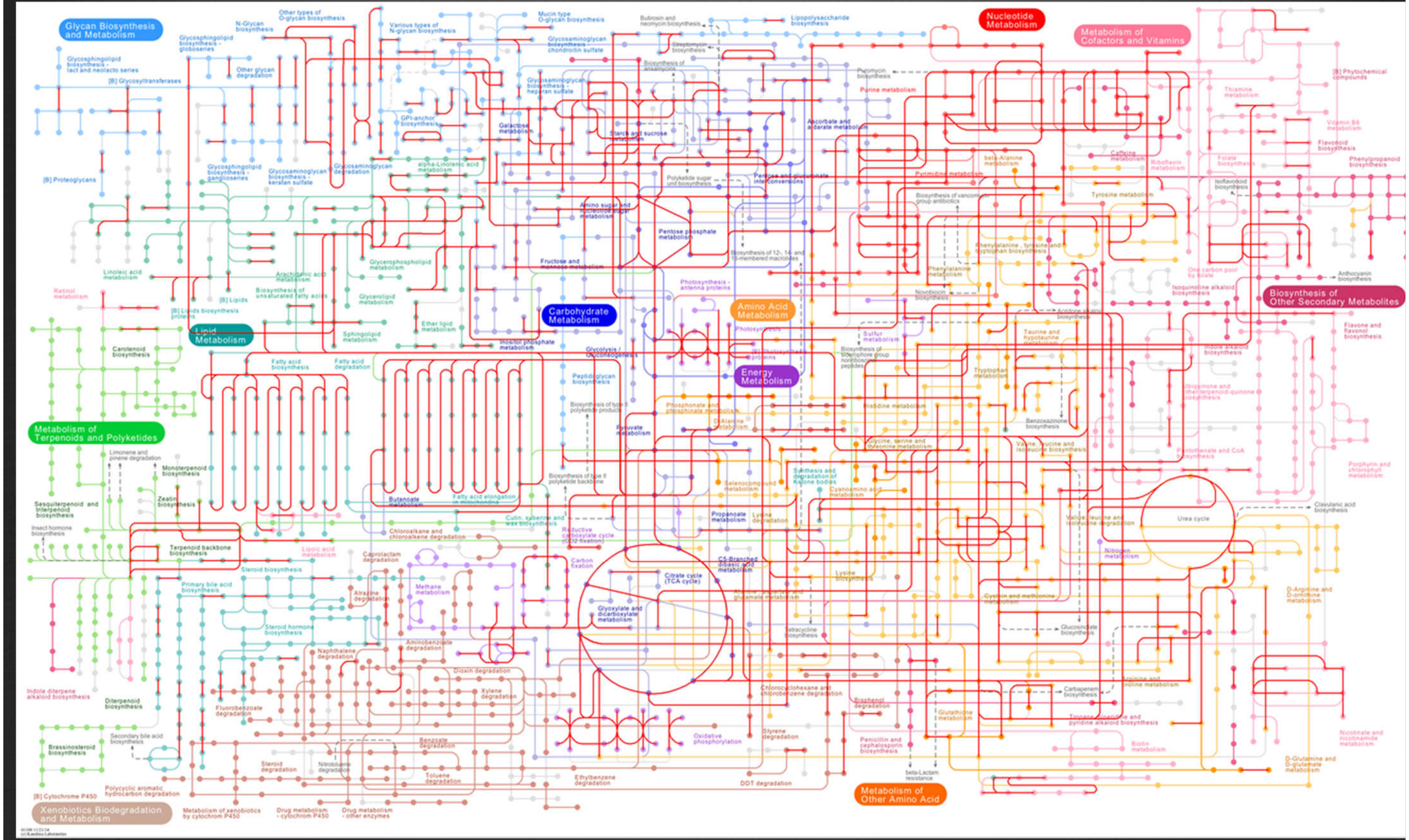
	% Genome recovery (CEGMA)	**Assembly size (Mb; contigs >1,000bp)**	**Contigs (> 1,000bp)**	**Max. contig**	**N50 (>1,000bp)**
Co-Assembly 14 SAGs MAST-4 clade A (SPAdes)	**89.1**	**47.5**	**14,564**	**57,905**	**4,563**
Co-Assembly 14 SAGs MAST-4 clade A (MegaHit)	80.6	42.5	15,158	51,080	3,475
MAST-4 clade A single SAG assembly (mean | SD)	*20.6 | 10.2*	*9.1 | 4.5*	*1,694 | 763*	*72,570 | 20,347*	*11,041 | 3,121*
Co-Assembly 9 SAGs MAST-4 clade E (SPAdes)	**68.5**	**32.3**	**5,677**	**104,912**	**9,991**
MAST-4 clade E single SAG assembly (mean | SD)	*14.3 | 5.5*	*6.2 | 2.4*	*1,098 | 350*	*63,915 |18,608*	*10,567 |1,920*

Continuing analyses with the co-assembly

1) Gene prediction [Augustus]

2) Annotation (KEGG, KOG, Pfam, eggNOG, OMRGC, MMETSP)

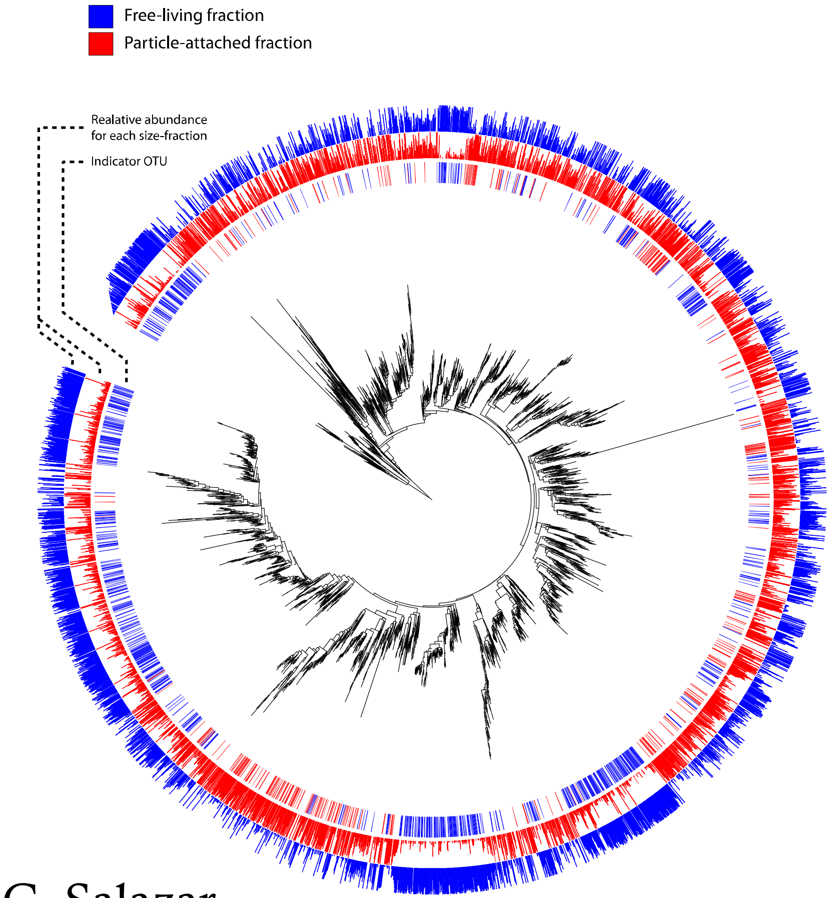
General metabolic pathways



Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Carbohydrate Metabolism
Amino Acid Metabolism
Lipid Metabolism
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Biosynthesis of Other Secondary Metabolites
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Urea cycle
Citrate cycle (TCA cycle)
Pentose phosphate metabolism
Oxidative phosphorylation
Purine metabolism
Pyrimidine metabolism

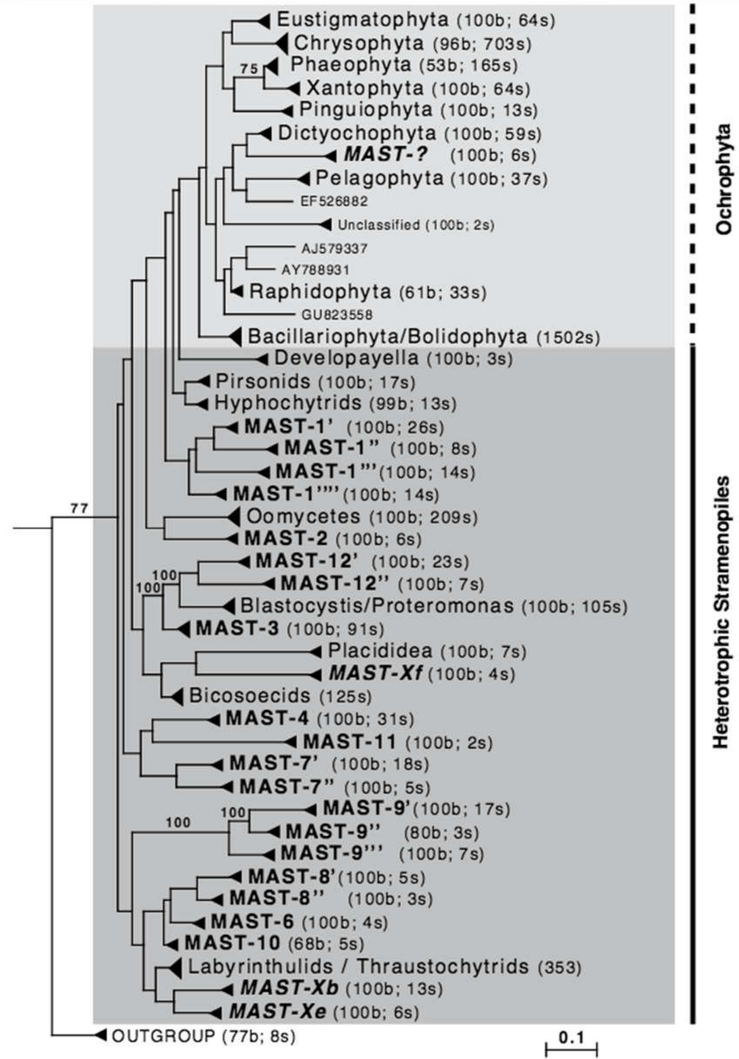
In Red: mapped MAST4 pathways/functions according to KO

Large phylogenetic analyses



G. Salazar

- Deep ocean Bacteria [Malaspina]
- 3,500 sequences (16S rRNA)
- 60 samples
- About 1,000 processes

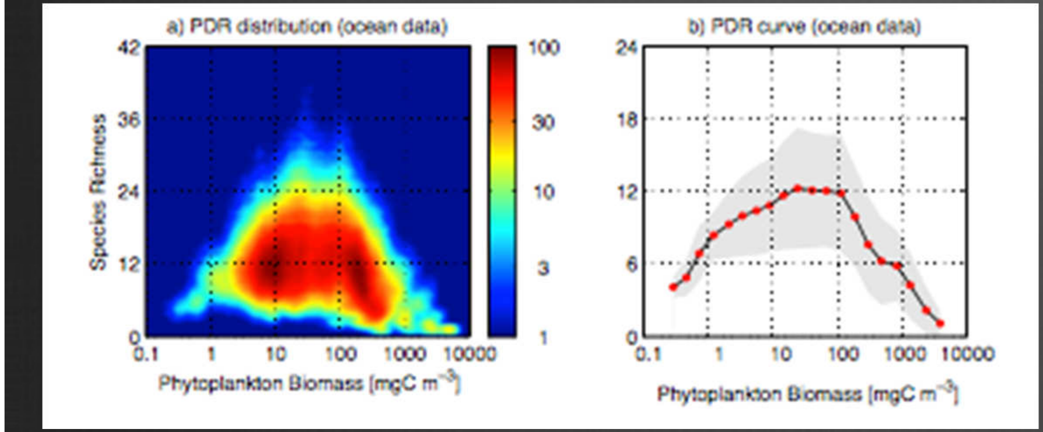


- Stramenopiles (protists)
- 3,835 sequences (>1,100bp) [18S]
- About 1,000 processes

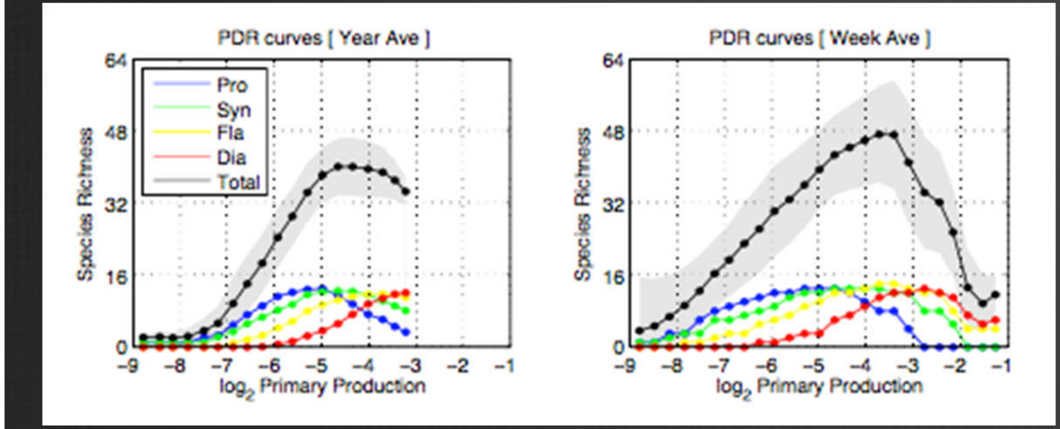
Simulations

Primary production vs. richness

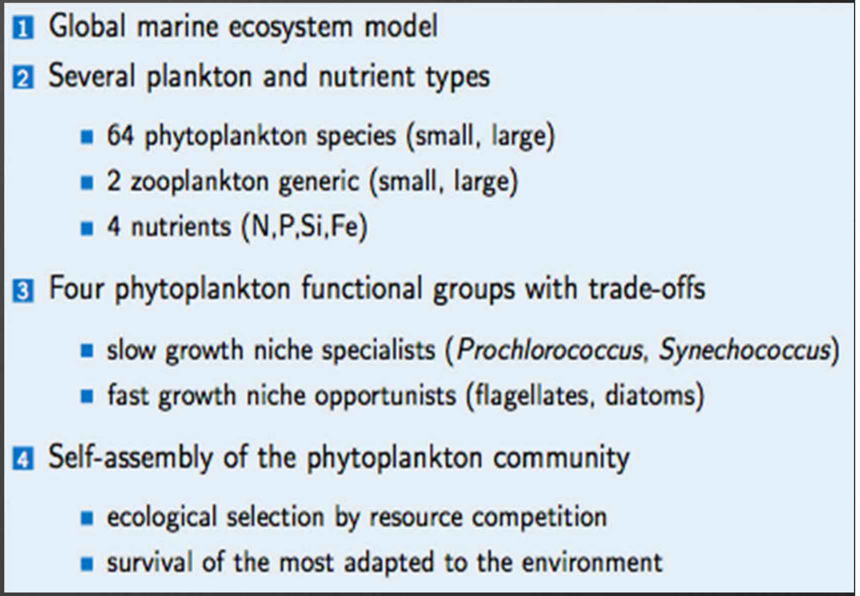
Real data



Simulation



MIT ecological selection model



1 Global marine ecosystem model

2 Several plankton and nutrient types

- 64 phytoplankton species (small, large)
- 2 zooplankton generic (small, large)
- 4 nutrients (N,P,Si,Fe)

3 Four phytoplankton functional groups with trade-offs

- slow growth niche specialists (*Prochlorococcus*, *Synechococcus*)
- fast growth niche opportunists (flagellates, diatoms)

4 Self-assembly of the phytoplankton community

- ecological selection by resource competition
- survival of the most adapted to the environment

Vallina et al., 2014. Nat. Comm

Summary of results with RES support since 2011

- 11 published papers
- 2 in revision
- Contributing mostly
  - Metagenomics
  - Genomics
  - Phylogenetics
  - Modelling

Conclusions: Microbial ecology

- Massive amounts of DNA data need powerful computers as well as programs that can deal with them
- Future developments require further integration with high-performance computers and quantitative methods
- Analysis of large datasets will likely unveil patterns of genomic functioning as well as interactions between marine microbes

ICM Institut de Ciències del Mar

microbial Malaspina@*ICM*

MALASPINA 2010

MASSIMO PERNICE

GUILLEM SALAZAR

SILVIA G. ACINAS

RAMON MASSANA

JULIA PERERA BEL

PABLO SANCHEZ

PEP GASOL

CARLES PEDRÓS-ALIÓ

JOSE A. MARTIN CANO

FRANCISCO CORNEJO

MARTA SEBASTIAN

CATERINA RODRÍGUEZ

Institut de Ciències del Mar

CSIC
Consejo Superior de Investigaciones Científicas

TARA OCEANS
by
tara EXPEDITIONS
www.taraexpeditions.org

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cnag

GENOSCOPE
Centre National de Séquençage

JGI
Joint Genome Institute
Department of Energy

Bigelow Laboratory for Ocean Sciences

CNRS UPMC INSU
Station Biologique Roscoff

Gobierno de España
Ministerio de Economía y Competitividad

PACIFIC BIOSCIENCES®

- Single Molecule Real Time (SMRT) sequencing

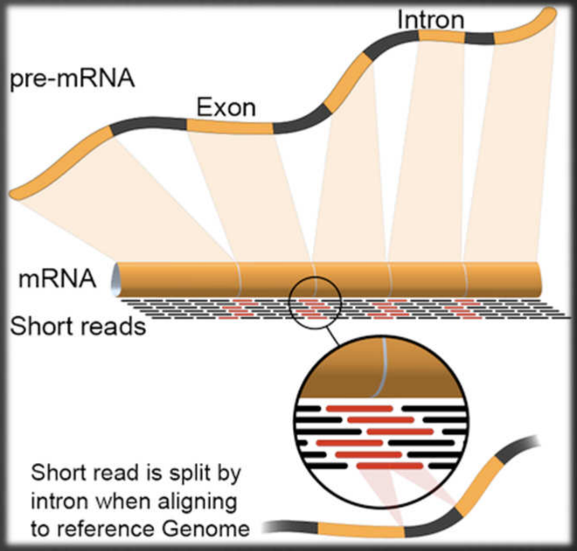
- Average read length: 4,200-8,500 bp (longest read 30Kbp)
  - P4/C2: shorter reads, higher accuracy
  - P5/C3: longer reads, lower accuracy
  - 200-300 Mbp from each SMRT cell for 15-20kb insert size libraries
  - 100-150 Mbp for >20Kbp libraries
  - No multiplexing in genomics libraries (multiplexing in amplicons)
  - Library preparation 400-1200€
  - About 350 € per SMRT cell
  - Signal: colors

NSC
NORWEGIAN SEQUENCING CENTRE

What if you are not interested in the whole community but in one species?

If you have a clonal culture:

Genomics and/or RNA-Seq


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

If you don't have a clonal culture:

Single-Cell genomics

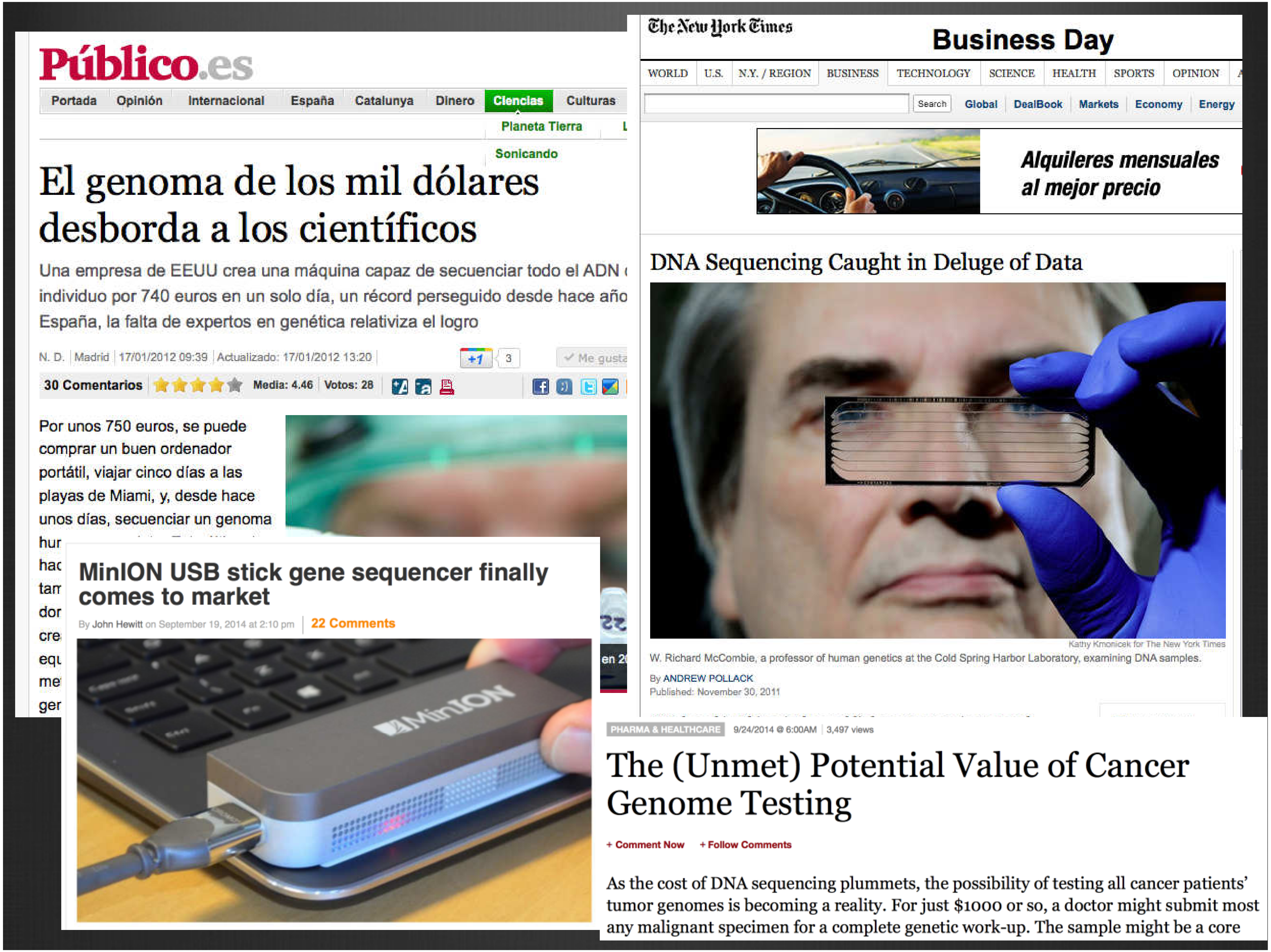
illumina

- $25\text{x}10^6$ reads
- 15Gb/run
- 2x300bp

- $400\text{x}10^6$ reads
- 120Gb/run
- 2x150bp

- $4\text{x}10^9$ reads
- 1000 Gb/run
- 2x125bp

Signal: colors
$1\text{Gb}=1\text{x}10^9$ bases



El genoma de los mil dólares desborda a los científicos

Una empresa de EEUU crea una máquina capaz de secuenciar todo el ADN individuo por 740 euros en un solo día, un récord perseguido desde hace año España, la falta de expertos en genética relativiza el logro

N. D. | Madrid | 17/01/2012 09:39 | Actualizado: 17/01/2012 13:20

30 Comentarios ★★★★☆ Media: 4.46 | Votos: 28

Por unos 750 euros, se puede comprar un buen ordenador portátil, viajar cinco días a las playas de Miami, y, desde hace unos días, secuenciar un genoma hur hac tam dor cre equ me ger

DNA Sequencing Caught in Deluge of Data

Kathy Kmonicek for The New York Times

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples.

By ANDREW POLLACK

Published: November 30, 2011

MinION USB stick gene sequencer finally comes to market

By John Hewitt on September 19, 2014 at 2:10 pm | 22 Comments

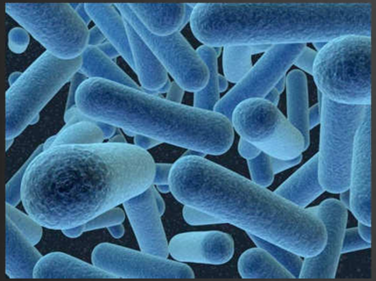
PHARMA & HEALTHCARE 9/24/2014 @ 6:00AM | 3,497 views

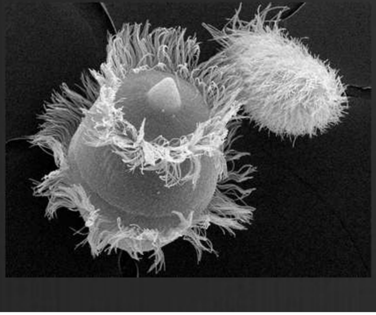
The (Unmet) Potential Value of Cancer Genome Testing

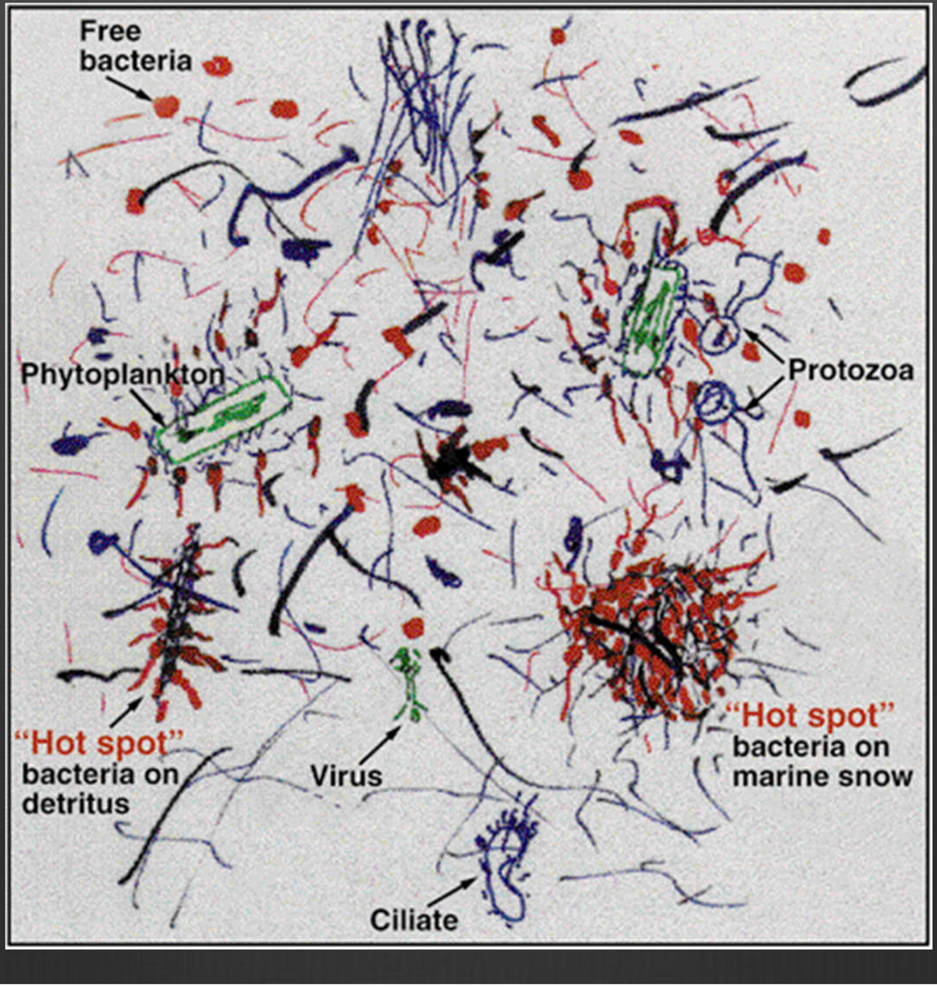
As the cost of DNA sequencing plummets, the possibility of testing all cancer patients' tumor genomes is becoming a reality. For just $1000 or so, a doctor might submit most any malignant specimen for a complete genetic work-up. The sample might be a core

Oceanic microbial community

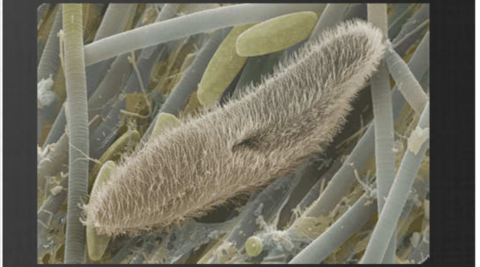
- Includes all species occurring at a particular site and their abundances

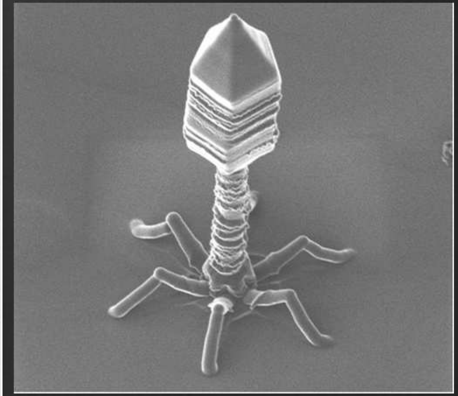




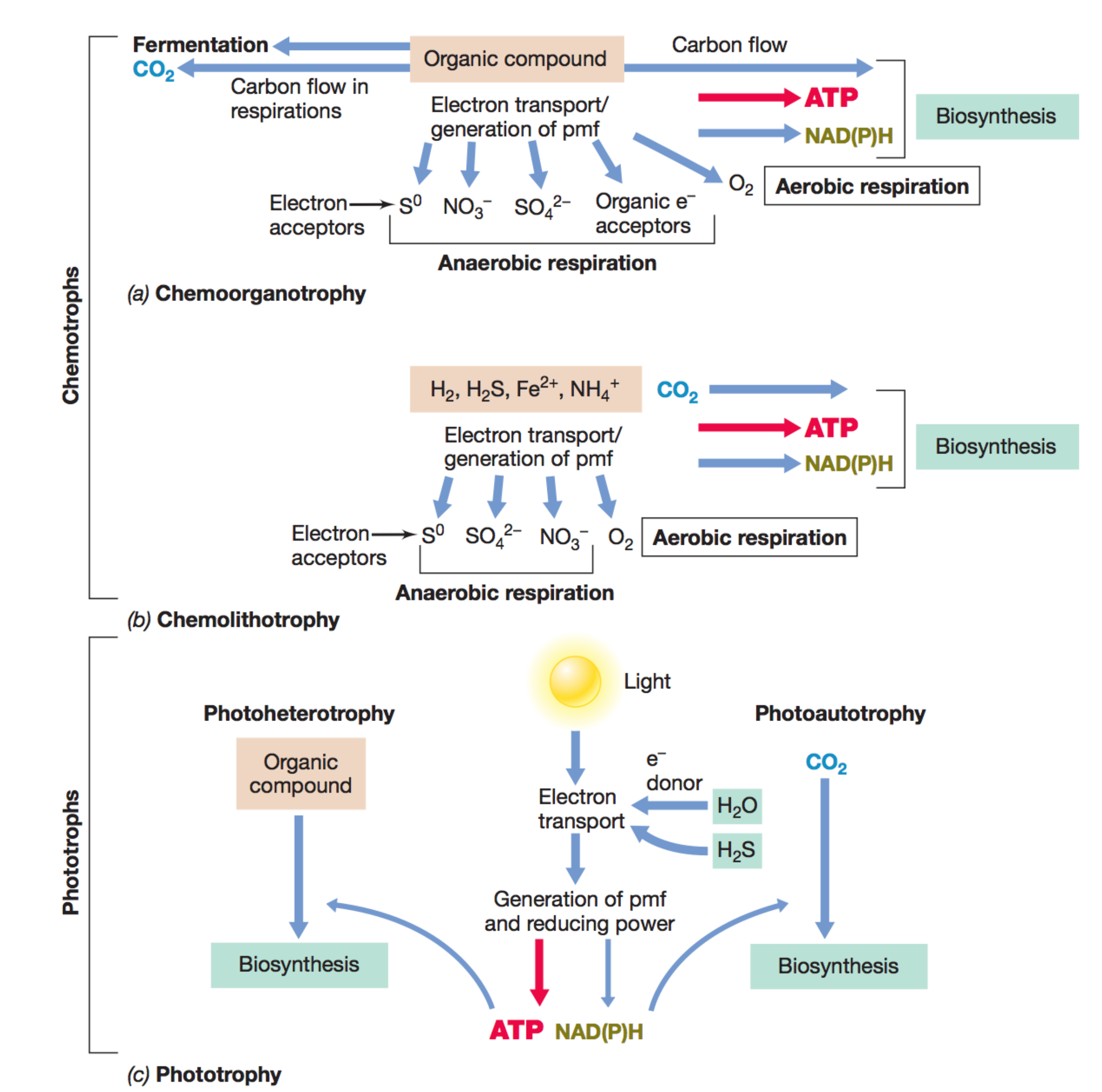


Free bacteria
Phytoplankton
Protozoa
"Hot spot" bacteria on detritus
Virus
"Hot spot" bacteria on marine snow
Ciliate





Multiple metabolisms



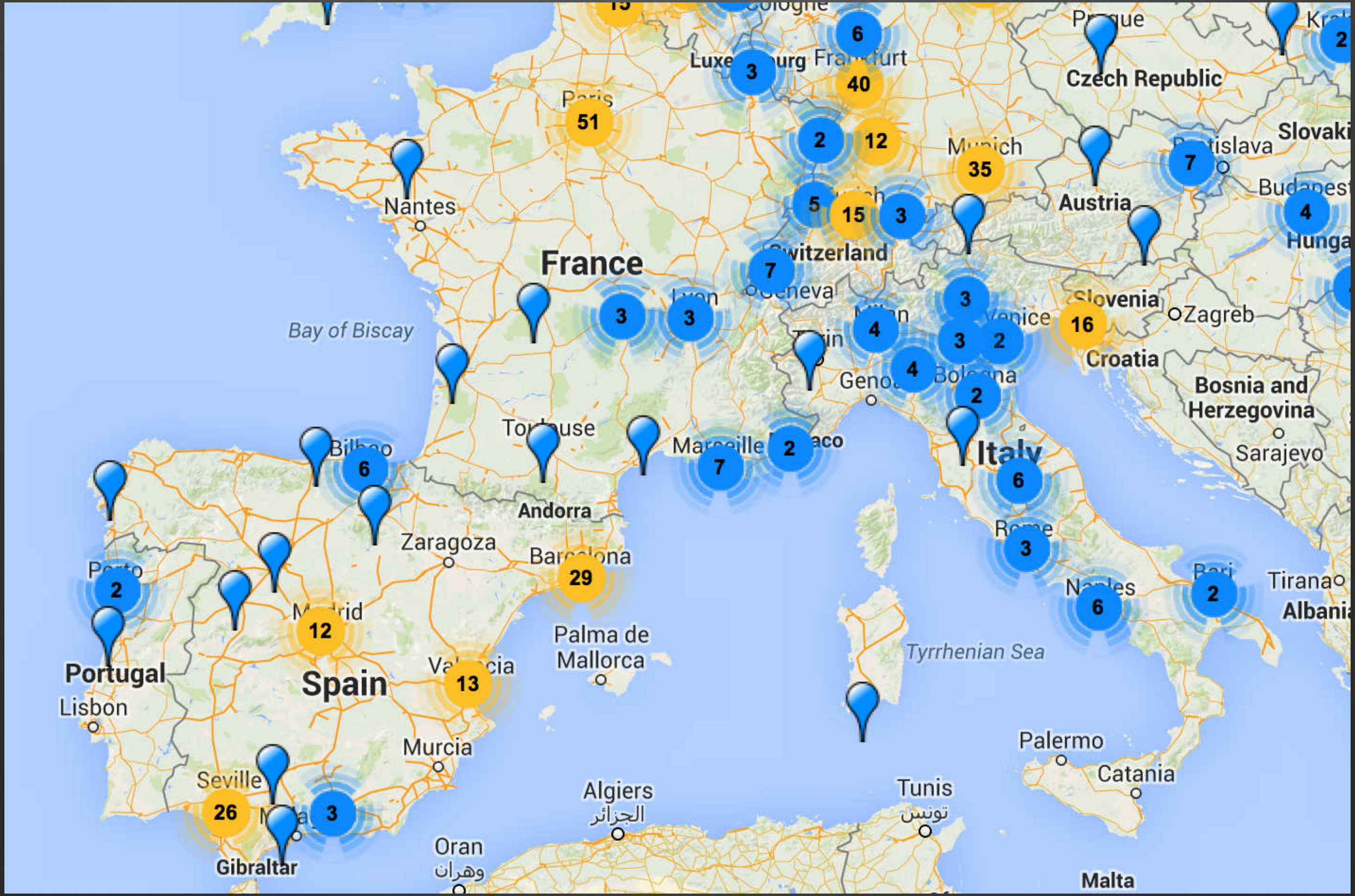
Brock “Biology of Microorganisms”

Comparison of platforms

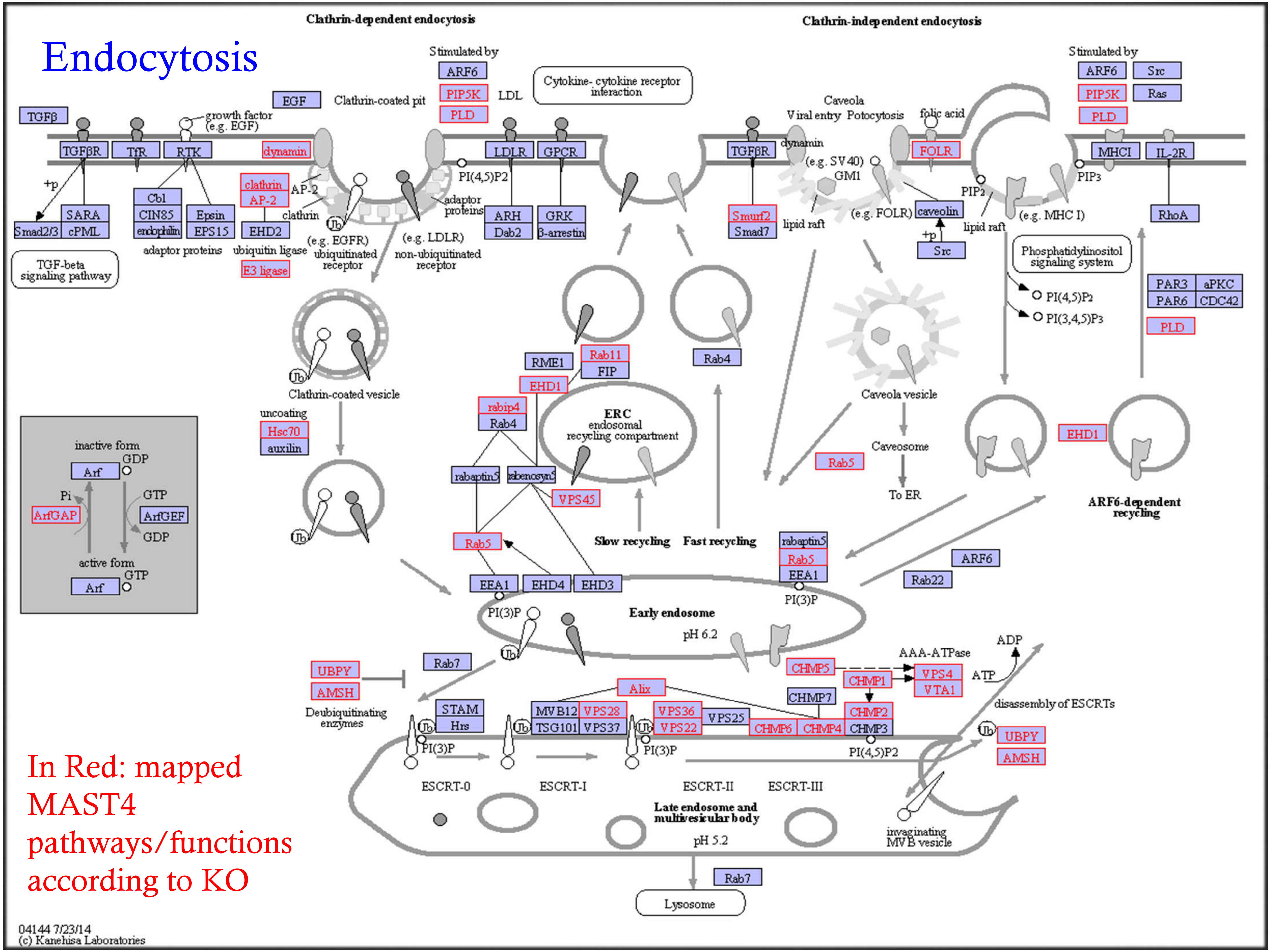
	Run time	Mreads/run	Read length	Mb/run	€/Mb
Sanger (3730xl)	2h	0.000096	400-900	0.06	1500
454 FLX Titanium	10h	1	400	400	15
454 FLX+	18-20 h	1	700	900	9
Ion Torrent	2h	80	400	32,000	1
PacBio	0.5-2h	0.005	4-8 K	300 (SMRT)	0.33-1
Illumina MiSeq	55h	25	2x300	15,000	0.1
Illumina GAIIx	14 days	320	2x150	96,000	0.12
Illumina HiSeq2500	1-11 days	4000	2x125	1,000,000	0.05

1Gb=1x10^9 bases

Glenn 2011 updated in
http://www.molecularecologist.com/next-gen-table-2-2014/

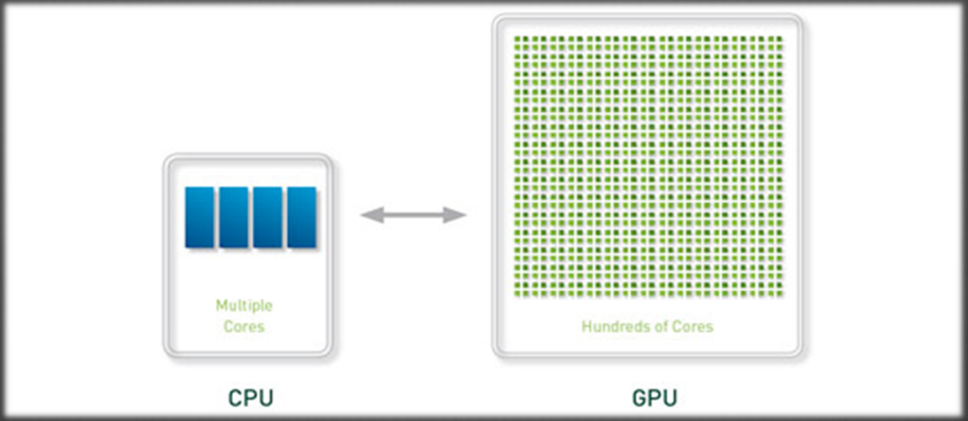

Cologne
6
Luxembourg
3
Frankfurt
40
Prague
Czech Republic
2
Paris
51
Nantes
2
12
Munich
35
Austria
Bratislava
7
Slovakia
Budapest
4
Hungary
5
15
3
Switzerland
7
Geneva
France
Lyon
3
3
Bay of Biscay
3
Slovenia
Zagreb
Venice
16
4
Turin
3
2
Croatia
Milan
Genoa
4
Bologna
2
Bosnia and Herzegovina
Sarajevo
Toulouse
Marseille
Monaco
7
2
Italy
6
Bilbao
6
Andorra
Zaragoza
Barcelona
29
Rome
3
Porto
2
Naples
6
Bari
2
Tirana
Albania
Madrid
12
Palma de Mallorca
Tyrrhenian Sea
Portugal
Spain
Valencia
13
Lisbon
Murcia
Palermo
Catania
Seville
26
3
Algiers
الجزائر
Tunis
تونس
Gibraltar
Oran
وهران
Malta

Endocytosis



In Red: mapped MAST4 pathways/functions according to KO

GPU (graphics processing unit) computing



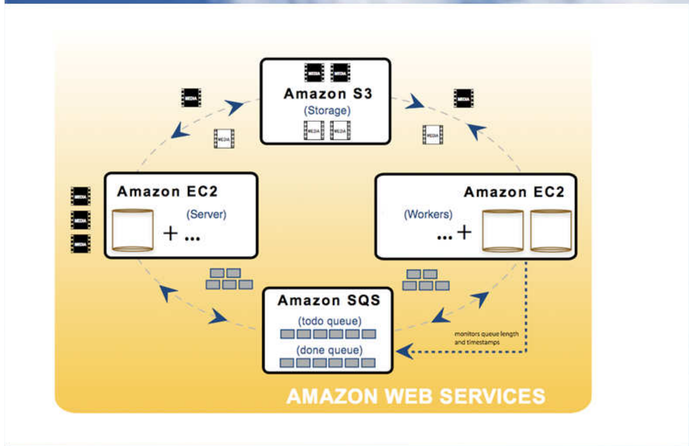
Serial part of an application runs on a CPU and the computationally-intensive part runs on a GPU

GPU Pipeline for HTS sequencing
Centro de Investigación Príncipe Felipe

http://docs.bioinfo.cipf.es/projects/ngs-gpu-pipeline/wiki

Cloud Computing…

- Purchase needed computer power
- Scalable (few to thousands of processors)
- No maintenance costs

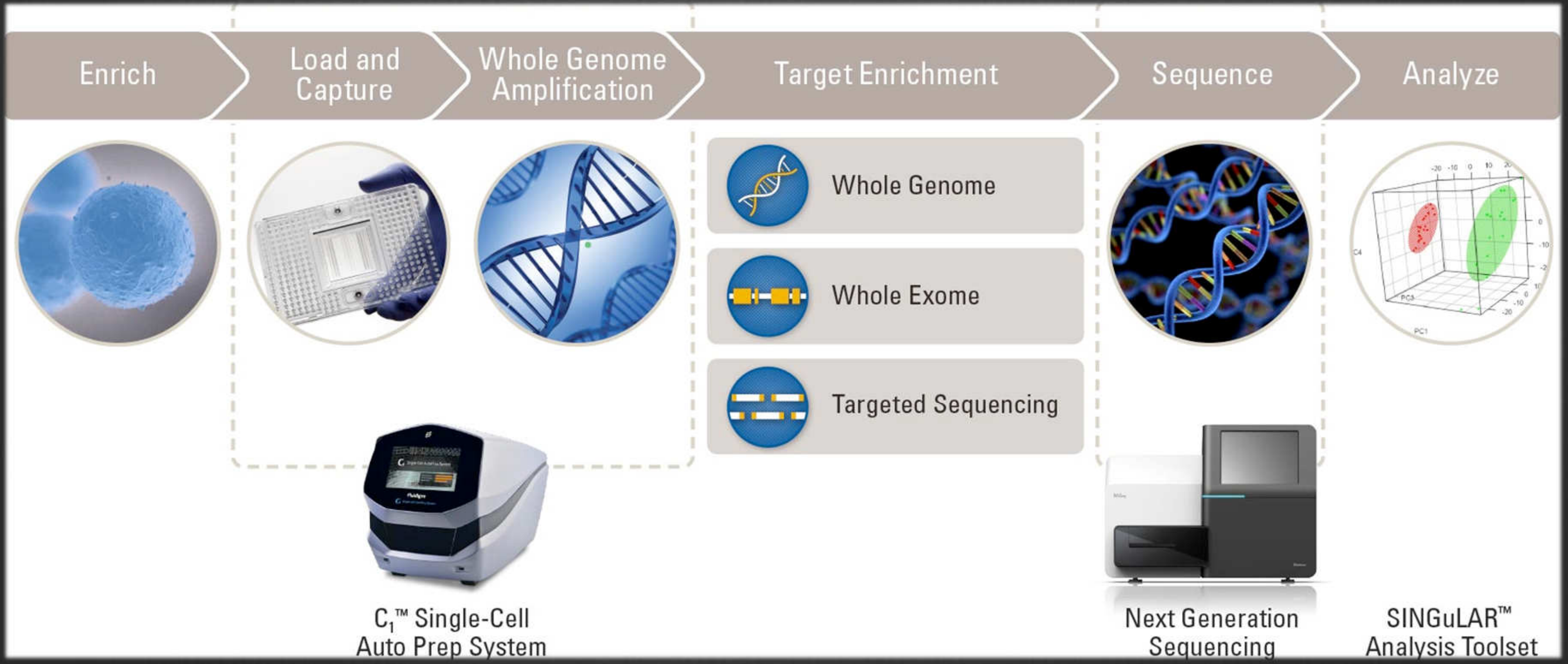


- GALAXY (g2.bx.psu.edu)
- N3phele: HTS analyses at the EC2 with QIIME

REVEAL HIDDEN VARIATION

C$_1$™ Single-Cell Targeted DNA Sequencing

The fastest and easier workflow to discover and screen mutations in target genes from individual cells.



Enrich

Load and Capture

Whole Genome Amplification

Target Enrichment

- Whole Genome
- Whole Exome
- Targeted Sequencing

Sequence

Analyze

C$_1$™ Single-Cell Auto Prep System

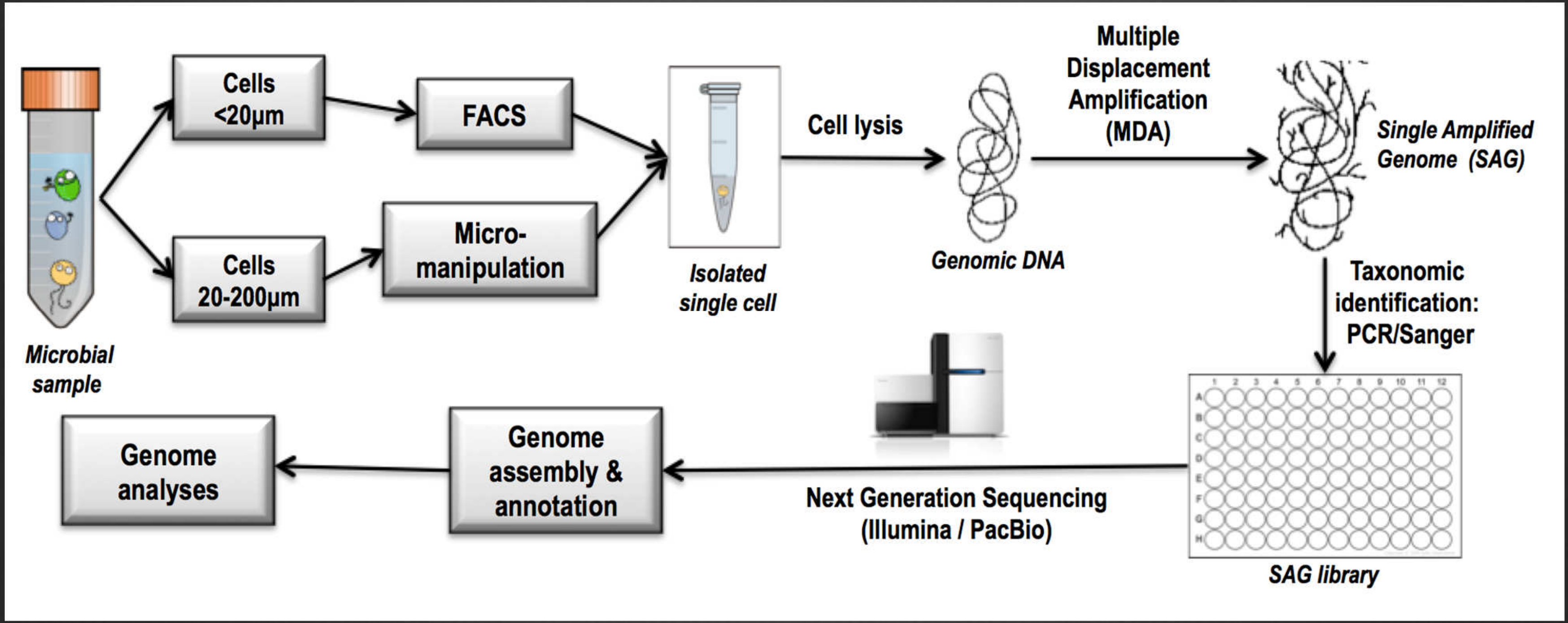
Next Generation Sequencing

SINGuLAR™ Analysis Toolset

- Molecular biology + computers + stats

- The next 20 years of genome research
- M. Schatz (2015)
- http://biorxiv.org/content/early/2015/06/02/020289

General SAG construction strategy


Cells <20µm
FACS
Cells 20-200µm
Micro-manipulation
Microbial sample
Isolated single cell
Cell lysis
Genomic DNA
Multiple Displacement Amplification (MDA)
Single Amplified Genome (SAG)
Taxonomic identification: PCR/Sanger
SAG library
Next Generation Sequencing (Illumina / PacBio)
Genome assembly & annotation
Genome analyses

GENOSCOPE Centre National de Séquençage

Bigelow | Laboratory for Ocean Sciences