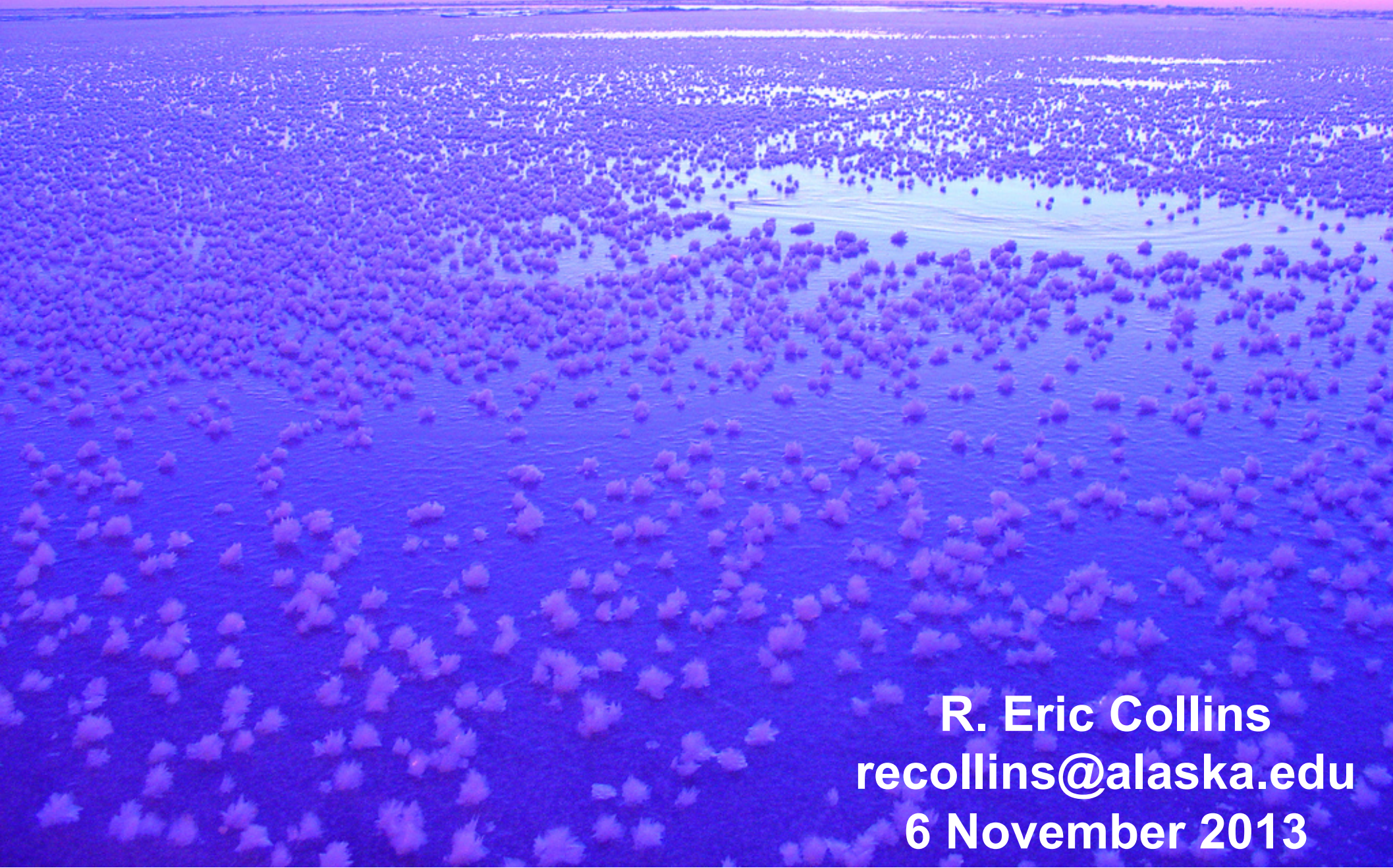
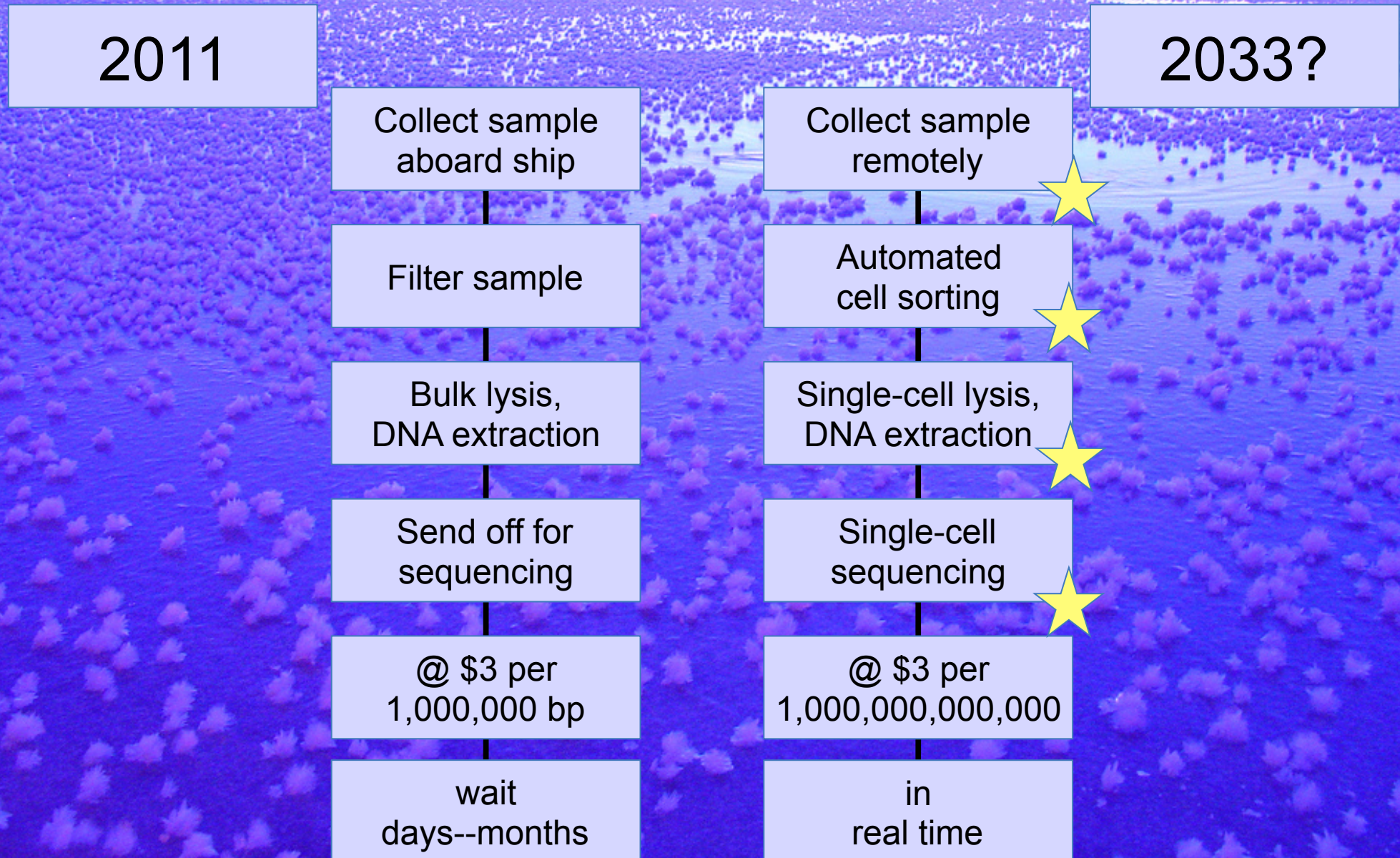


Polar Marine Science in the Genomics Era



R. Eric Collins
recollins@alaska.edu
6 November 2013

Polar Marine Science in the Genomics Era



Polar Marine Science in the Genomics Era

2013

Collect sample
aboard ship

Filter sample

Bulk lysis,
DNA extraction

Send off for
sequencing

@ \$3 per
100,000,000 bp

wait
days--months

2033?

Collect sample
remotely

Automated
cell sorting

Single-cell lysis,
DNA extraction

Single-cell
sequencing

@ \$3 per
1,000,000,000,000

in
real time



Polar Marine Science in the Genomics Era

Genomics Era highlights

- Global Ocean Survey
- Moore Foundation Marine Microbes
- Joint Genome Institute
- Earth Microbiome Project

Comparative Genomics projects

- genome modeling: birth, death, and the pangenome
- hima: database of cold-active microbe genomes
- ddmp: database of cold-active microbe phenotypic traits
- SRB: sulfur isotope fractionation

Upcoming Projects

- Compatible Solutes

λ Global Ocean Survey



The 2003 Global Ocean Survey: The Genomics Era in marine science is born



Craig Venter is on a boat!

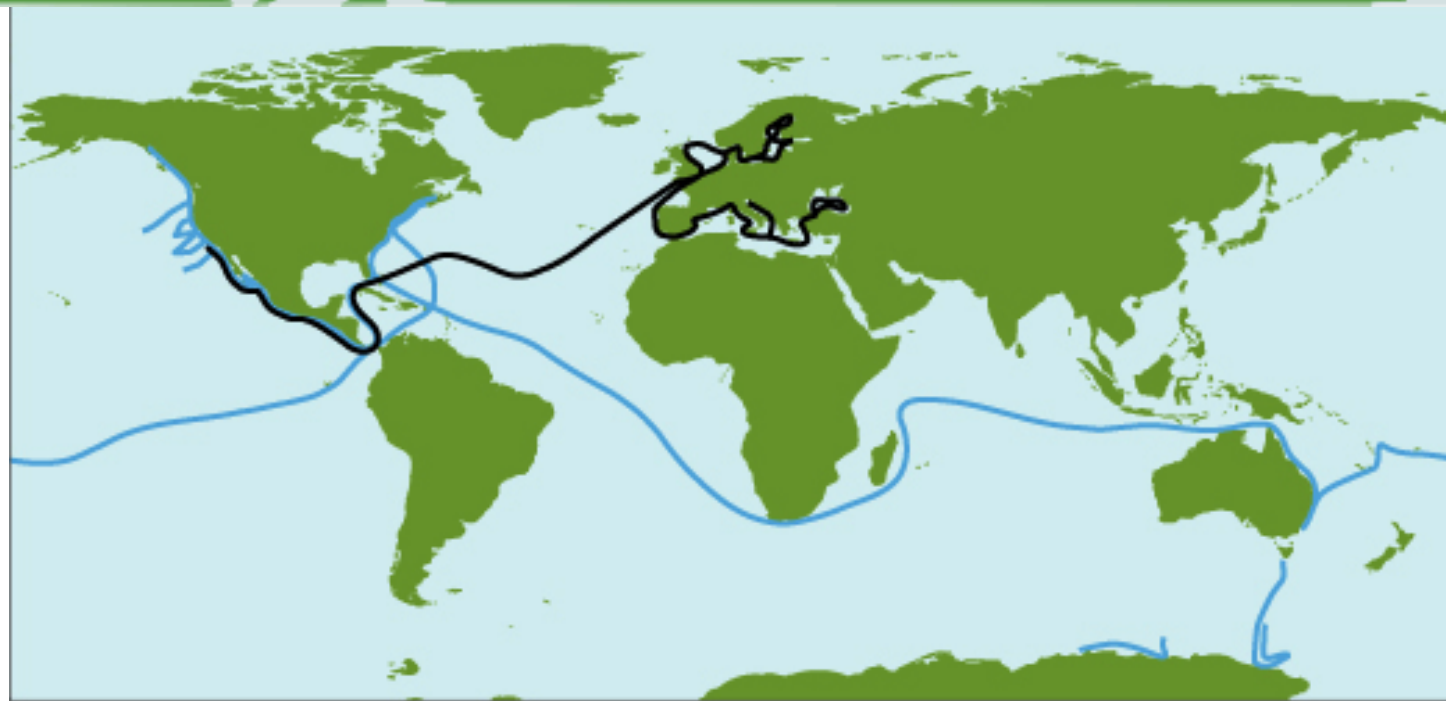


But Charles Darwin beat him to it.





Beagle
1831-1836

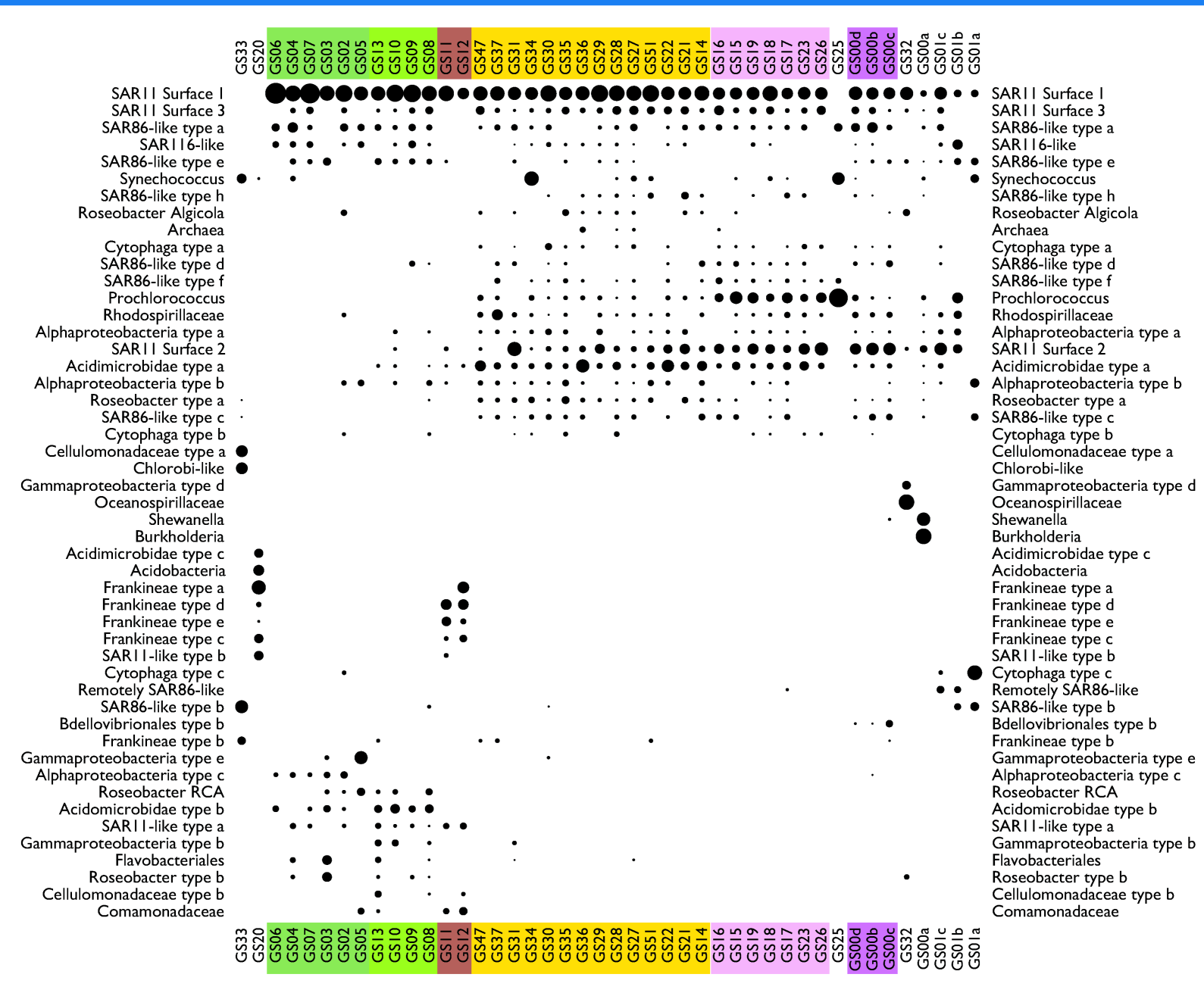


Sorcerer II
2003-2010

— 2003 – 2008 Routes — 2009 – 2010 Route

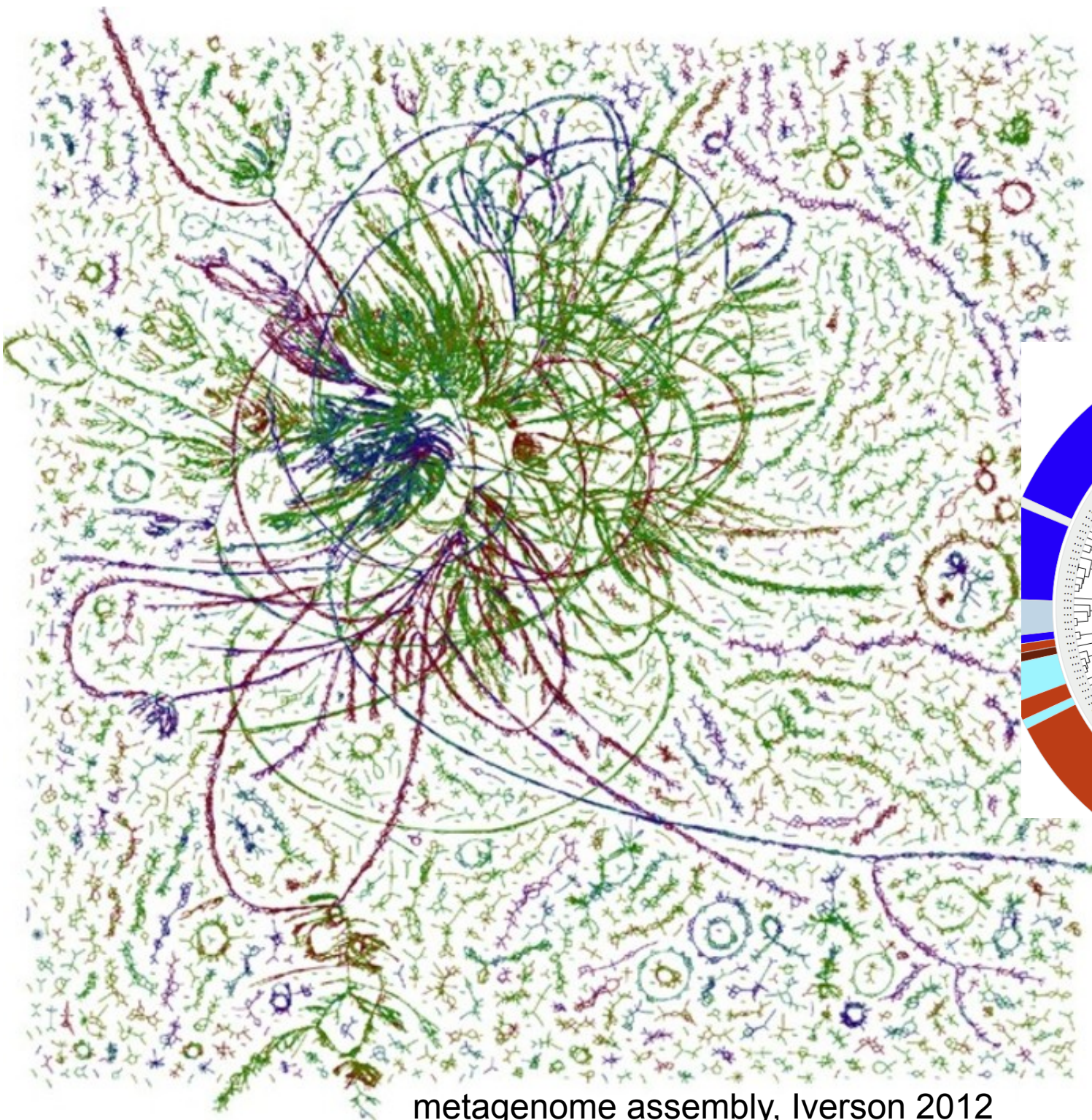
Metagenomics: GOS oceanic diversity

▼ PRESENCE AND ABUNDANCE OF DOMINANT RIBOTYPES

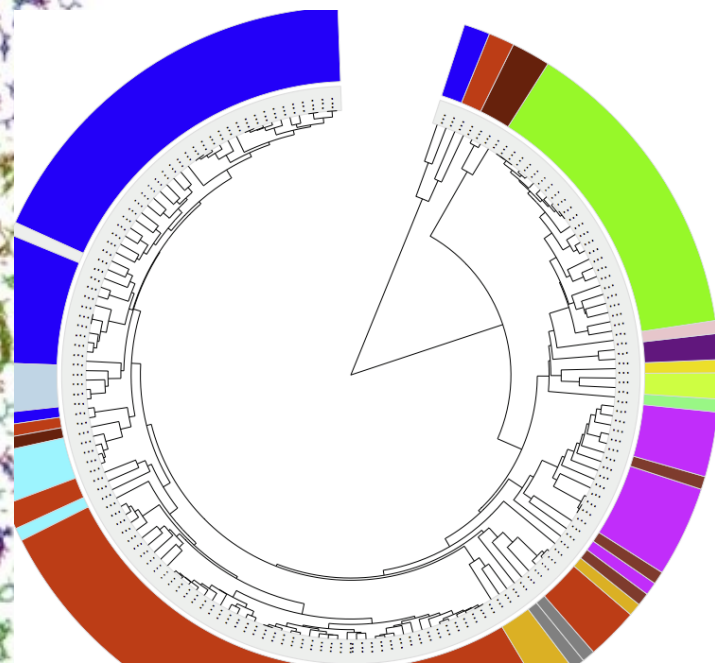


λ Moore Foundation

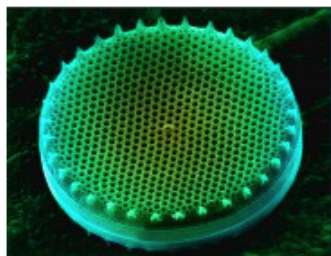




182 bacterial
strains



metagenome assembly, Iverson 2012



Marine Microbial Eukaryote Transcriptome Sequencing Project

NCGR's MMETSP Website | List of Publicly Available Datasets

Home :: List of Publicly Available Datasets

MMETSP ID ▲	Genus ◆	Species ◆	Strain ◆	Principle Investigator(s) ◆
MMETSP0004_2	Ochromonas	sp.	CCMP1393	David Caron
MMETSP0005	Ochromonas	sp.	CCMP1393	David Caron
MMETSP0006_2	Prymnesium	parvum	Texoma1	David Caron
MMETSP0007	Prymnesium	parvum	Texoma1	David Caron
MMETSP0008_2	Prymnesium	parvum	Texoma1	David Caron
MMETSP0010_2	Corethron	hystrix		E. Virginia Armbrust
MMETSP0011_2	Rhodorus	marinus	CCMP769	E. Virginia Armbrust
MMETSP0013_2	Skeletonema	costatum	RA080513-05	E. Virginia Armbrust
MMETSP0014_2	Nitzschia	sp.	RCC80	E. Virginia Armbrust
MMETSP0015_2	Odontella	aurita	isolate 1302-5	E. Virginia Armbrust
MMETSP0017_2	Cylindrotheca	closterium	KMMCC:B-181	E. Virginia Armbrust
MMETSP0019_2	Dinobryon	sp.	UTEXLB2267	David Caron
MMETSP0020_2	Dinobryon	sp.	UTEXLB2267	David Caron
MMETSP0027	Karenia	brevis	CCMP2229	Geoff Sinclair, Boris Wawrik
MMETSP0029	Karenia	brevis	CCMP2229	Geoff Sinclair, Boris Wawrik
MMETSP0030	Karenia	brevis	CCMP2229	Geoff Sinclair, Boris Wawrik
MMETSP0031	Karenia	brevis	CCMP2229	Geoff Sinclair, Boris Wawrik
MMETSP0033_2	Dolichomastix	tenuilepis	CCMP3274	Alexandra Z. Worden
MMETSP0034_2	Nephroselmis	pyriformis	CCMP717	Alexandra Z. Worden
MMETSP0038_2	Cryptomonas	paramecium	CCAP977/2a	John Archibald
MMETSP0039_2	Eutreptiella	gymnastica	NIES-381	John Archibald
MMETSP0040_2	Lotharella	oceanica	CCMP622	John Archibald
MMETSP0041_2	Lotharella	globosa	LEX01	John Archibald
MMETSP0042_2	Lotharella	amoebiformis	CCMP2058	John Archibald
MMETSP0043_2	Hemiselmis	andersenii	CCMP644	John Archibald
MMETSP0045_2	Bigelowiella	natans	CCMP2755	John Archibald
MMETSP0046_2	Guillardia	theta	CCMP2712	John Archibald
MMETSP0052_2	Polytomella	parva	SAG 63-3 (a clonal isolate of SAG 63-3)	David Smith
MMETSP0058	Tetraselmis	sp.	CCMP879	Aurora Nedelcu

686 eukaryotic
transcriptomes

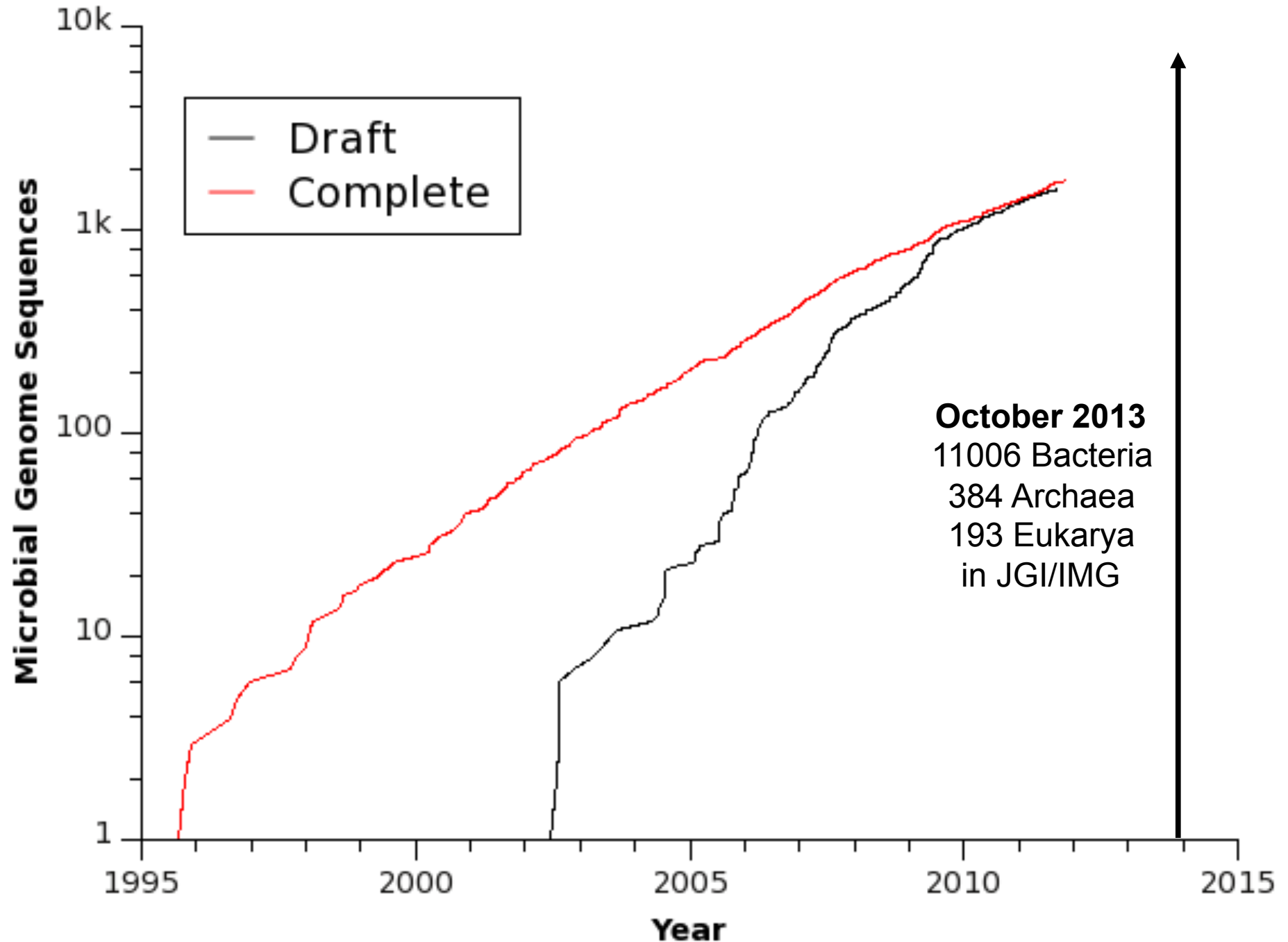
Metatranscriptomics

	Mid-Bloom				Post-Bloom				Combined data		
	DNA-High CO ₂	mRNA-High CO ₂	DNA-Present Day	mRNA-Present Day	DNA-High CO ₂	mRNA-High CO ₂	DNA-Present Day	mRNA-Present Day	All DNA	All mRNA	All samples
Total size (Mbp)	47,289,282	30,567,377	30,991,689	38,021,523	59,316,369	21,805,955	68,187,679	26,982,195	205,784,939		
117,377,050	323,161,989										
Total No. of reads	209,073	131,089	134,915	162,871	344,216	96,201	304,020	116,192	992,224	506,353	1,498,577
Average length (bp)	226	233	229	233	172	226	224	232	207	231	215
% of rRNA genes ^a	0.33	0.16	0.31	0.1	0.17	0.03	0.24	0.03	0.25	0.08	0.16
Absolute number of unique nucleotide sequence clusters ^b	170,580	65,717	112,459	67,283	257,375	9,349	232,729	10,703	630,159	133,447	723,050
Normalized number of clusters ^c	86,791	50,320	84,096	43,213	86,996	9,349	87,112	9,281	n/a	n/a	n/a
Total number of reads in top cluster	12	988	23	932	19	2,019	20	2,421	36	4860	4866
Clustering: 1 sequence	141,340	56,126	94,386	55,691	200,569	7,397	183,028	8,106	437,149	107,593	494,604
2–9 sequences	29,232	8,824	18,072	10,367	56,729	1,182	49,681	1,677	191,822	23,224	224,198
10–99 sequences	8	652	1	1,032	77	552	20	673	1188	2,011	3,639
100+ sequences	0	115	0	193	0	218	0	247	0	619	609
SEED Subsystem hits	130,567	75,884	120,141	83,076	161,789	16,545	175,477	18,315	n/a	n/a	n/a
Total pORFs ^d	419,565	284,665	279,061	345,502	532,373	237,187	637,896	289,951	1,868,895	1,157,305	3,026,200
Unique pORFs at 95% ^e	358,705	140,763	242,317	147,697	435,876	31,104	515,266	36,113	1,340,241	299,898	1,571,348
Protein clusters ^f	321,839	120,220	223,888	121,547	382,762	16,641	452,754	19,259	1,083,644	238,655	1,228,601
Protein clusters ^g with similarity to:	11	296	3	468	29	516	31	589	695	1,131	2,029
PFAM ^h	9	19	1	23	14	56	13	63	379	76	571
TIGRfam ⁱ	9	0	1	0	12	0	7	0	366	1	476
COG ^j	10	2	1	0	18	0	17	0	431	6	572
Number of novel Protein clusters ^k	0	276	2	444	9	447	12	509	202	1026	1287

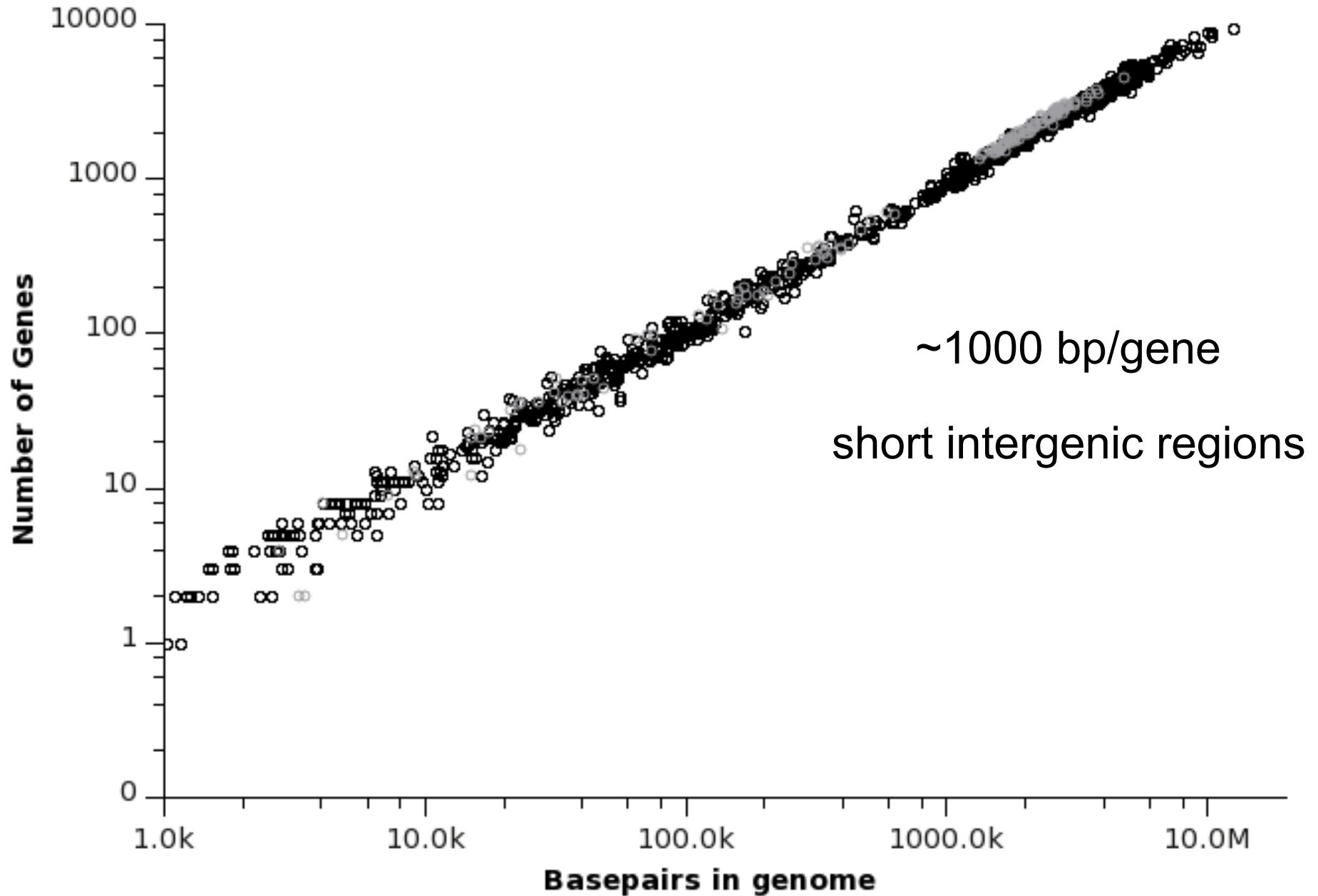
λ JGI/IMG



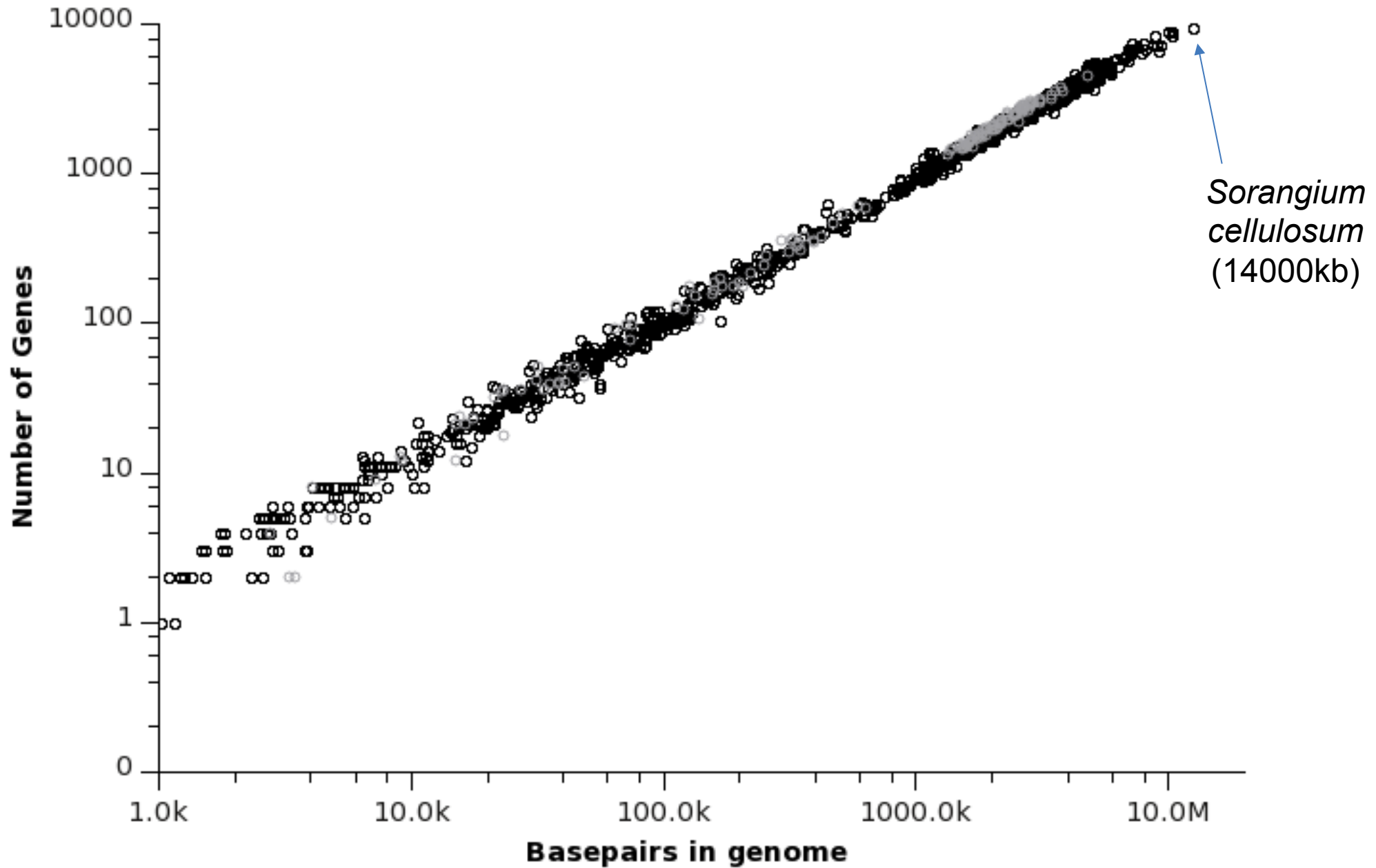
Genome Sequencing Explosion



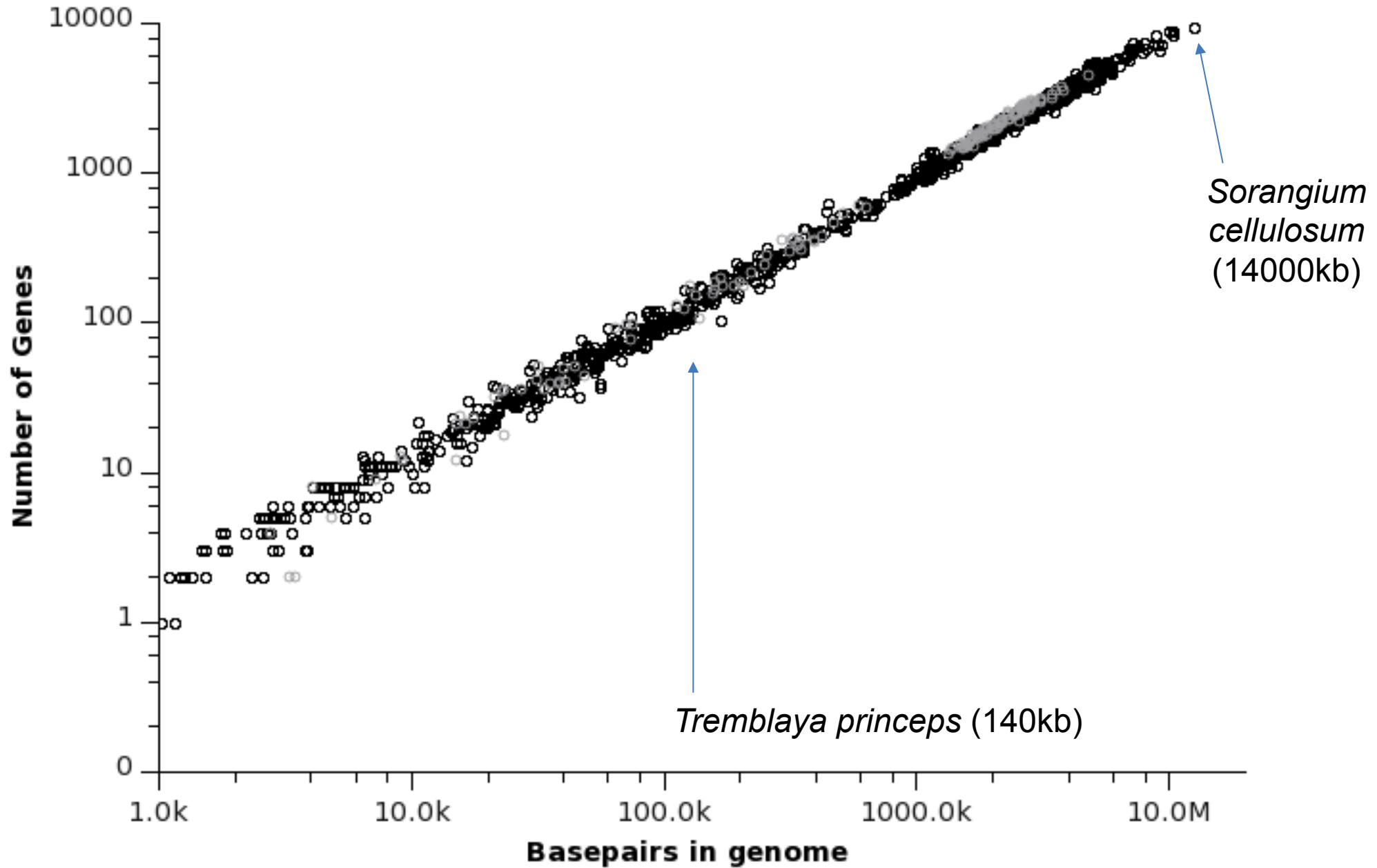
Streamlining in 1424 genomes



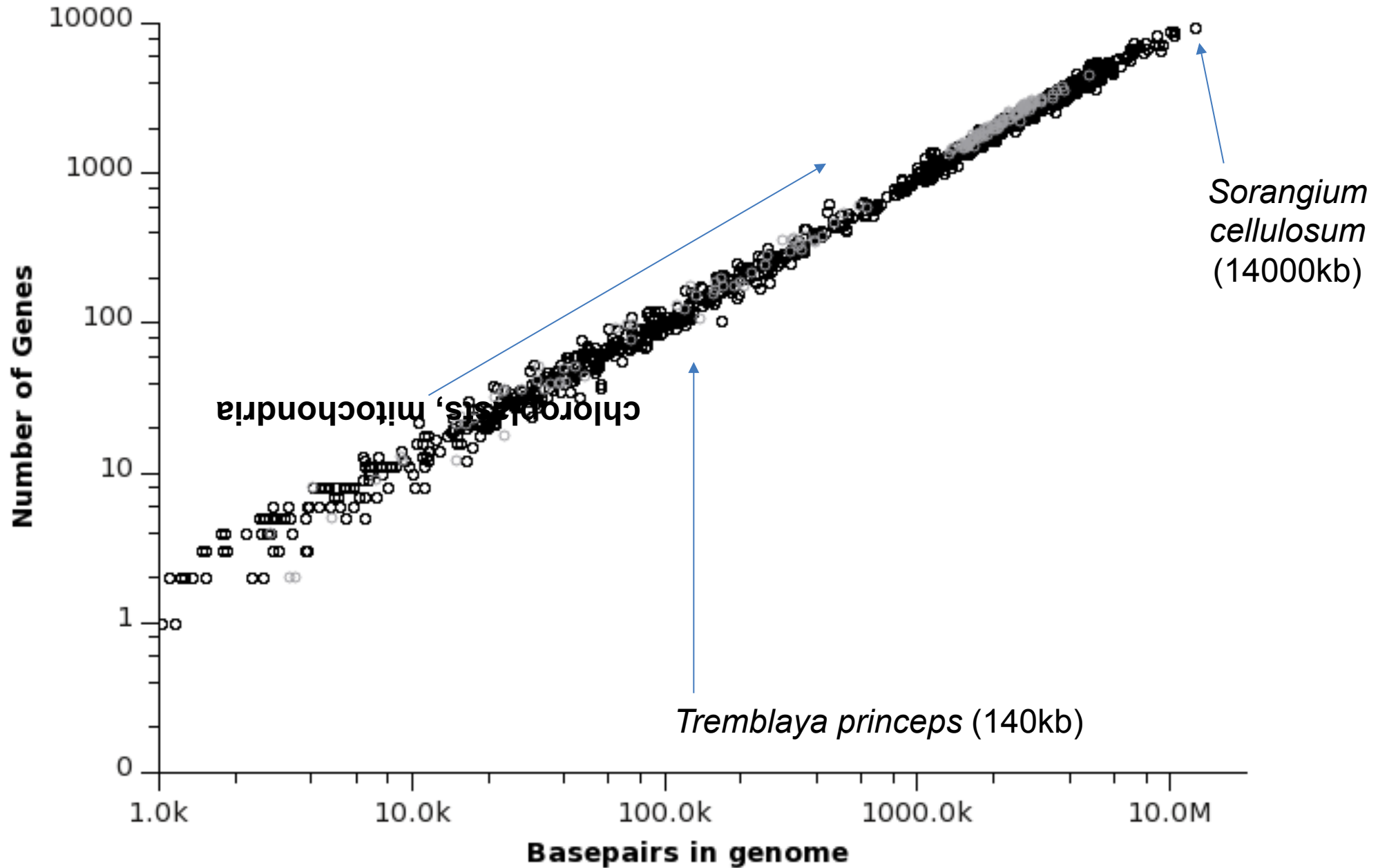
Largest bacterium



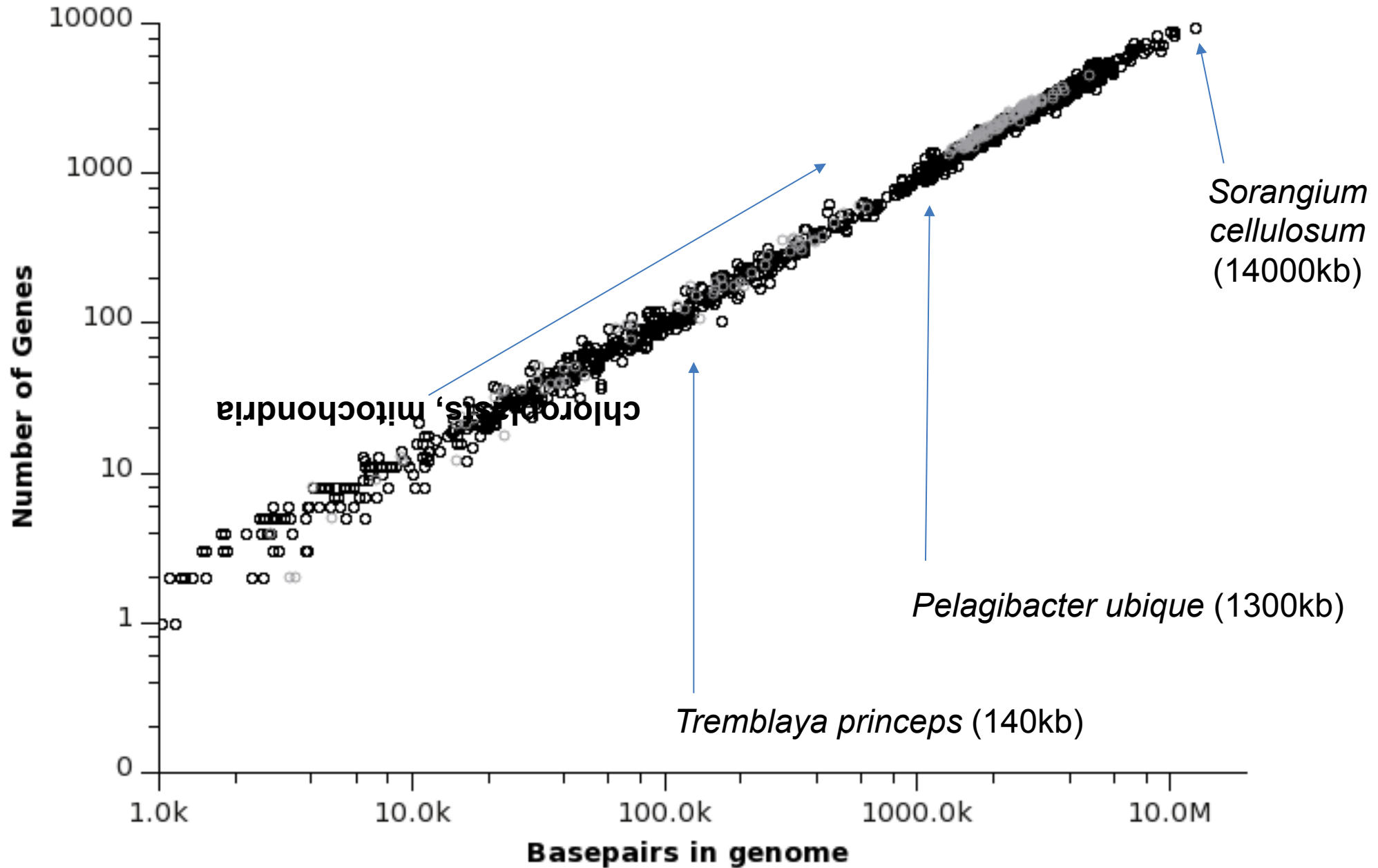
Smallest bacterium



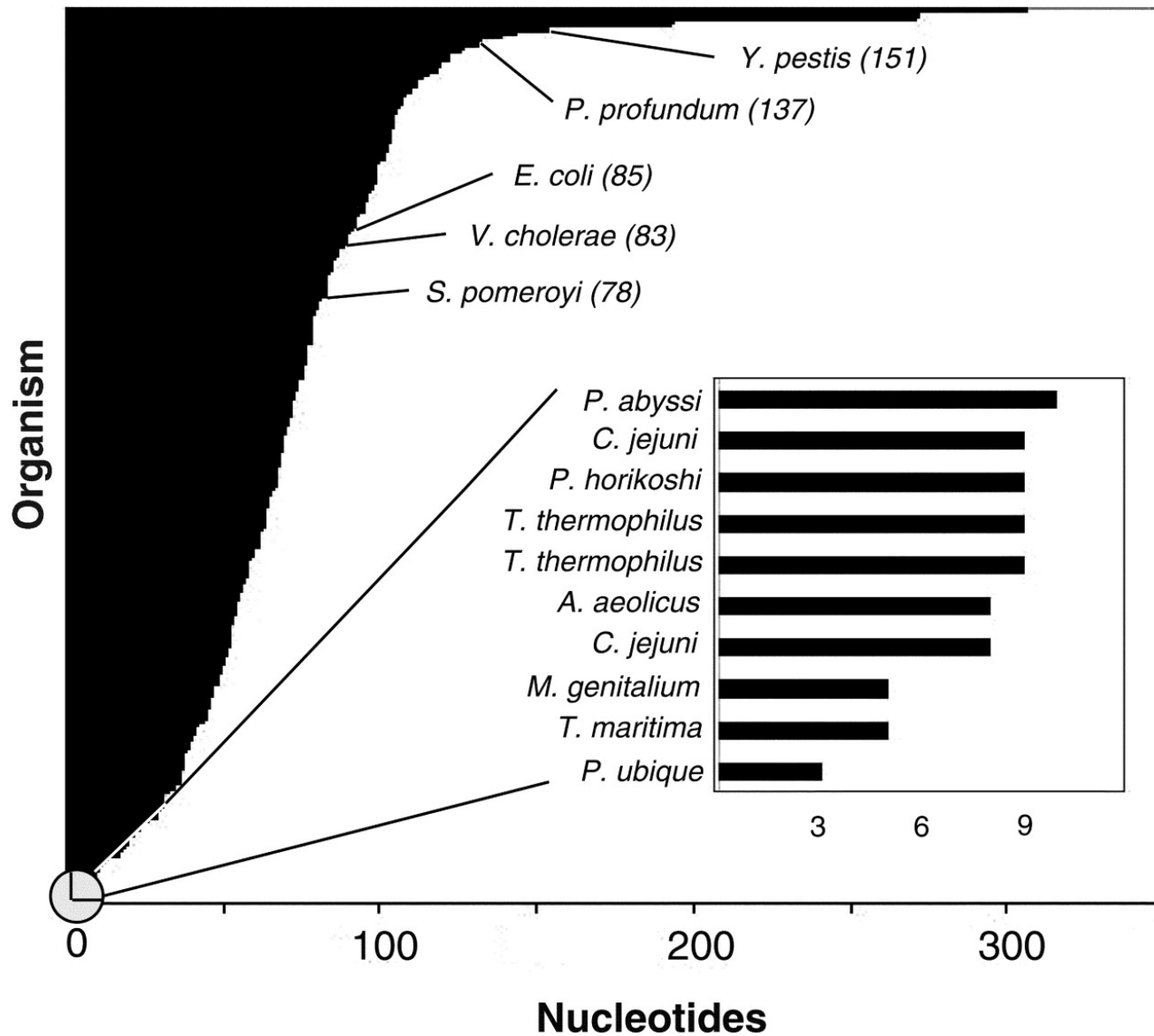
Organelles overlap endosymbionts



Smallest free-living bacterium

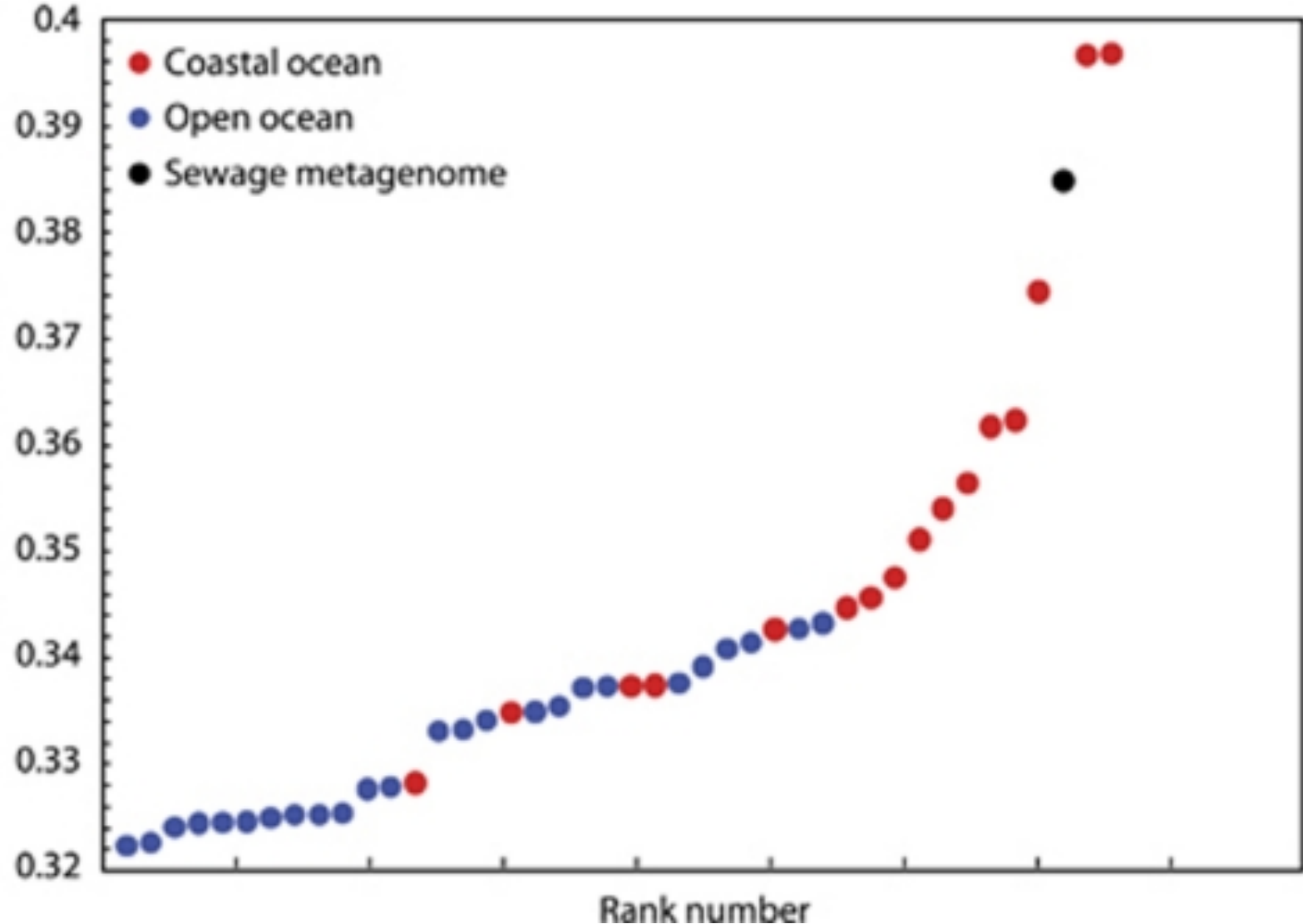


Minimization of intergenic spacer lengths



Minimization of nutrient requirements

Nitrogen residues per amino acid side chain



Gene family size distributions show that gene deletion is biased over duplication

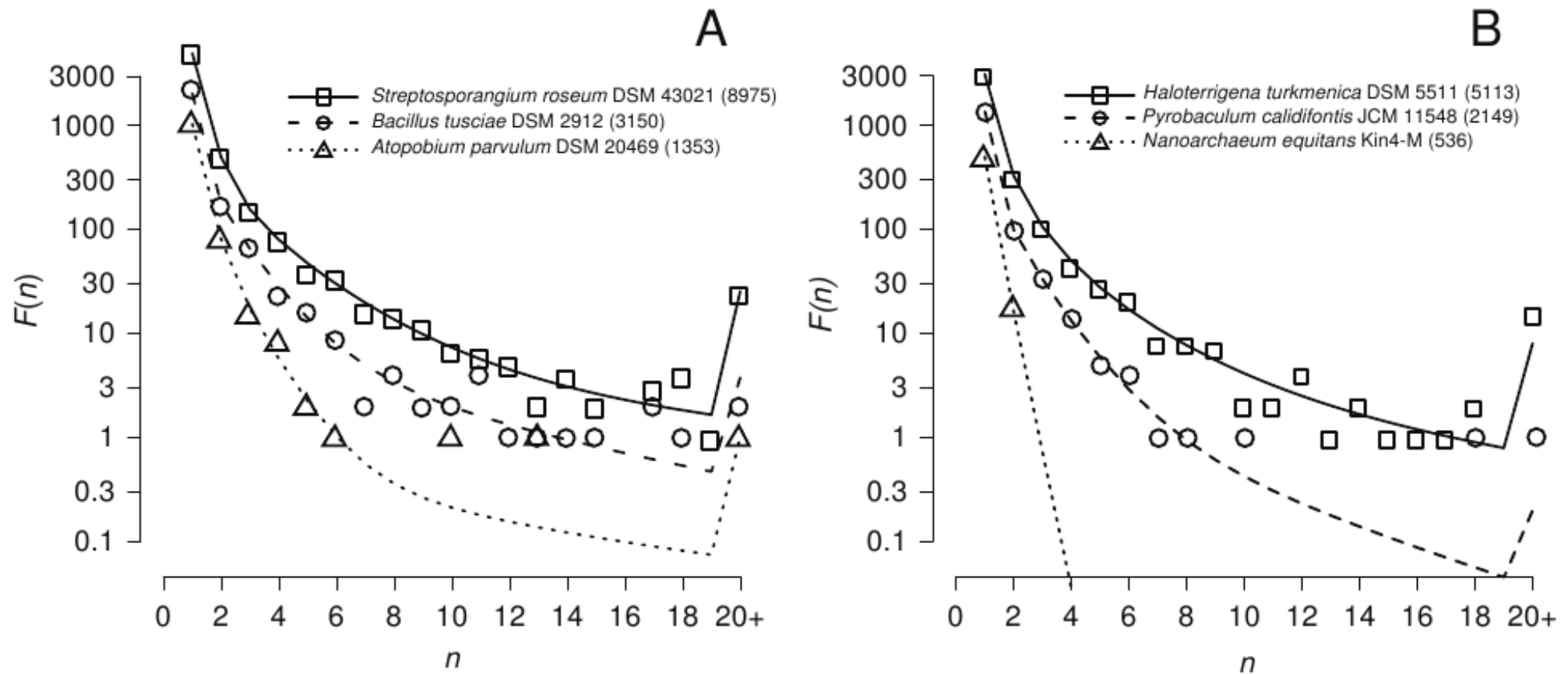
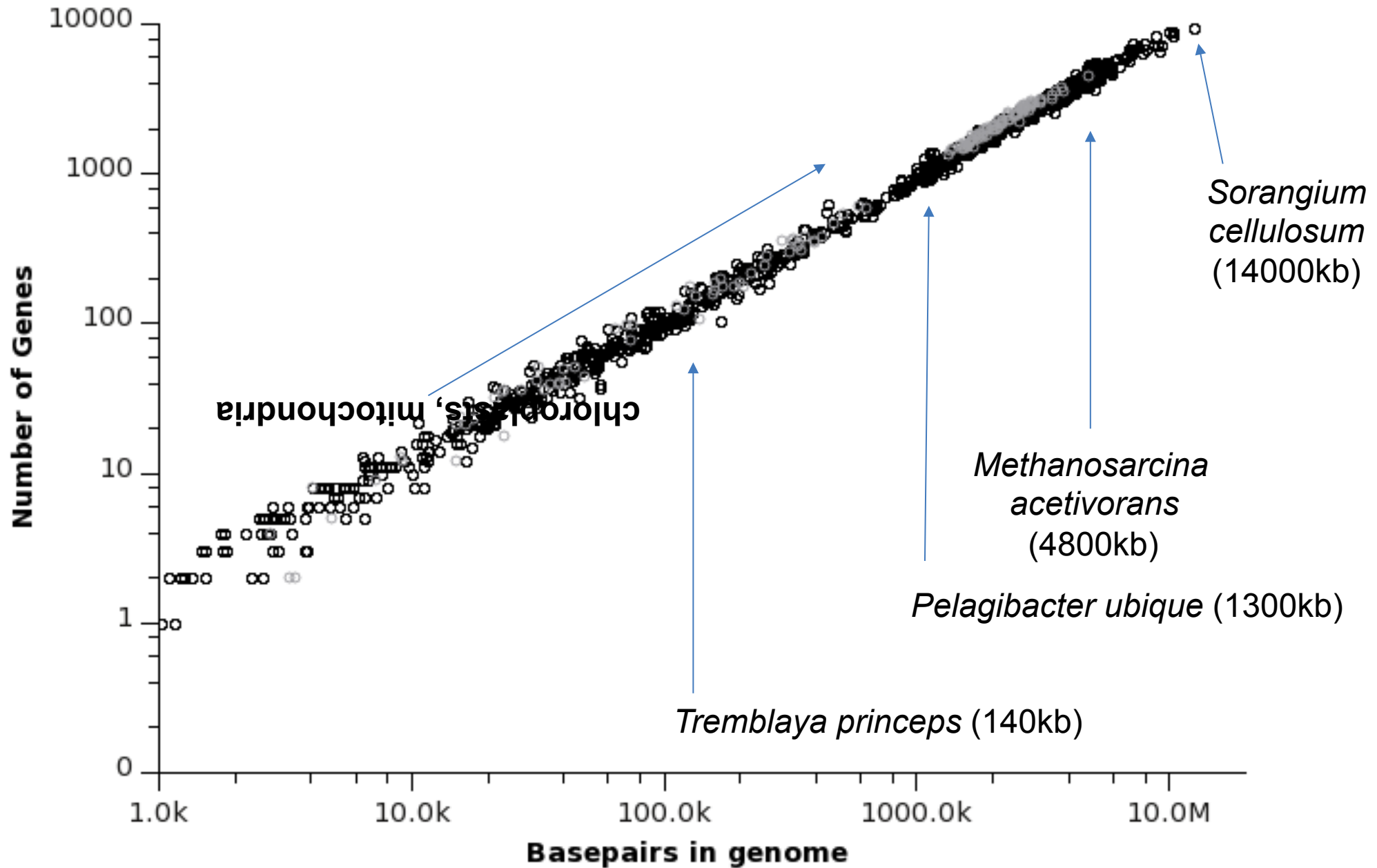
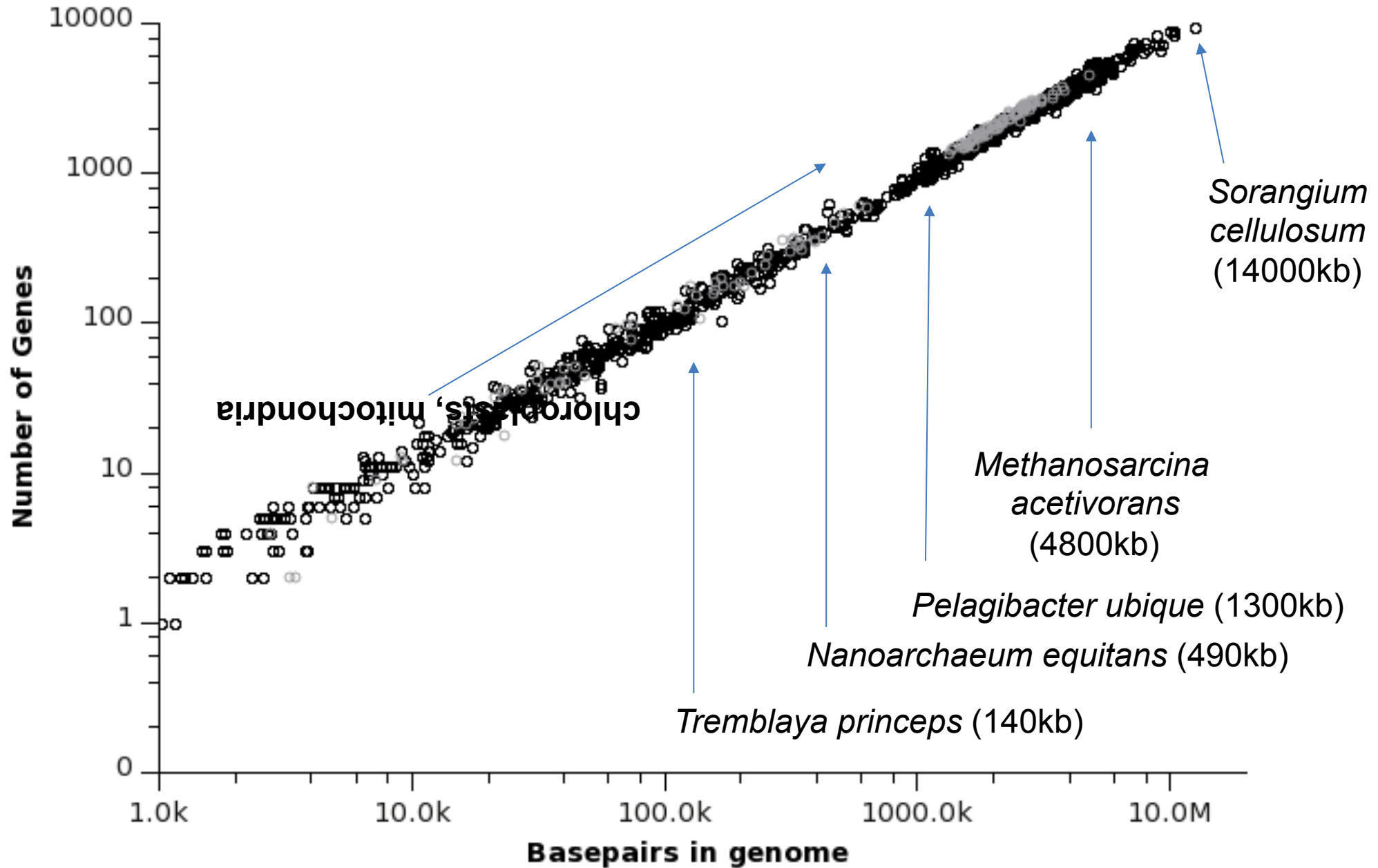


Figure 7 Maximum Likelihood fits of model BD13 to the family size distribution $F(n)$ in genomes of (A) 3 Bacteria and (B) 3 Archaea, clustered at $E_{cut} = 1e-20$ and $f_{min} = 0.7$. The selected taxa are the smallest, median, and largest genomes in the clustered datasets; genome size, as number of protein coding sequences, is shown in parentheses following each taxa name.

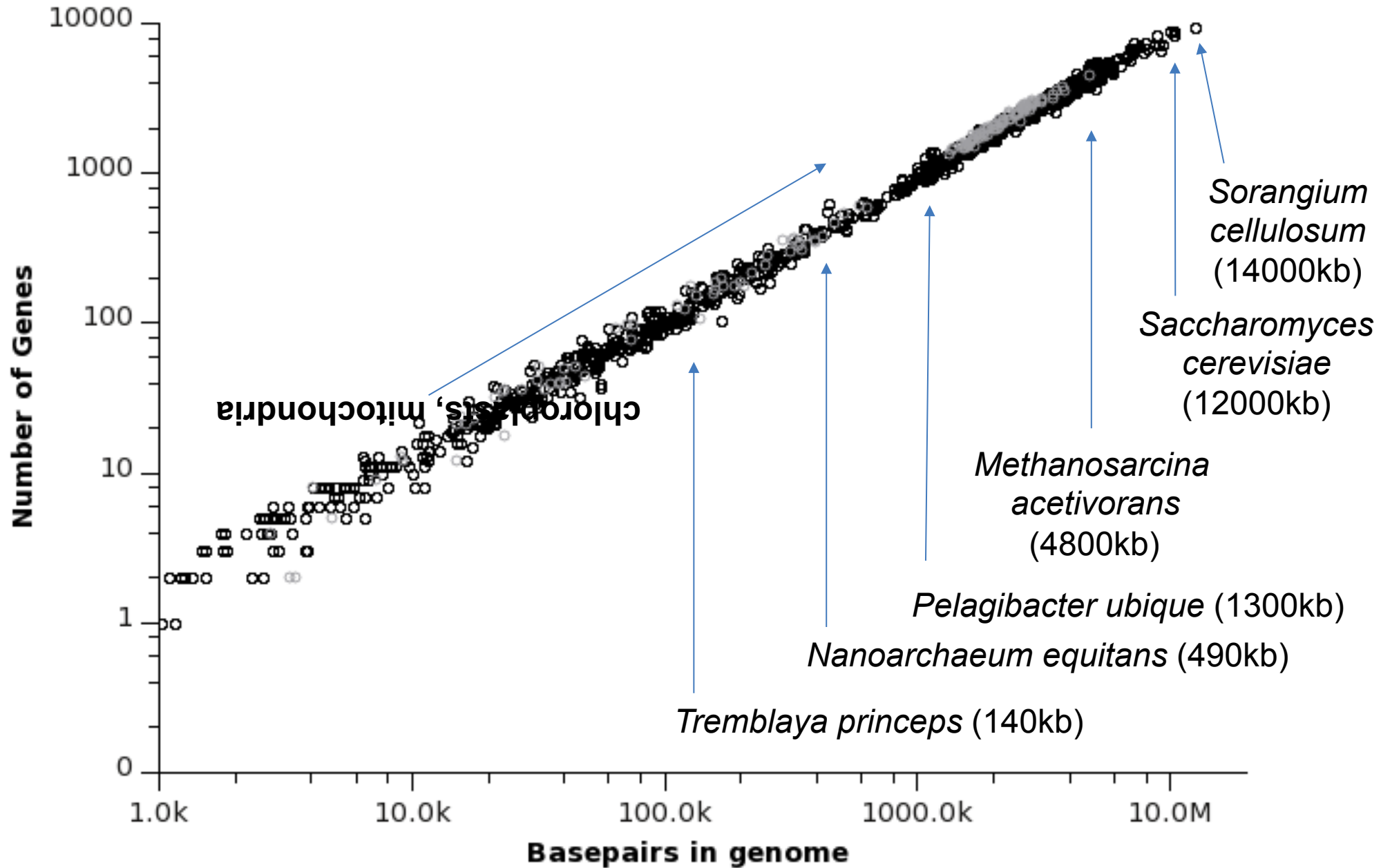
Largest archaeon



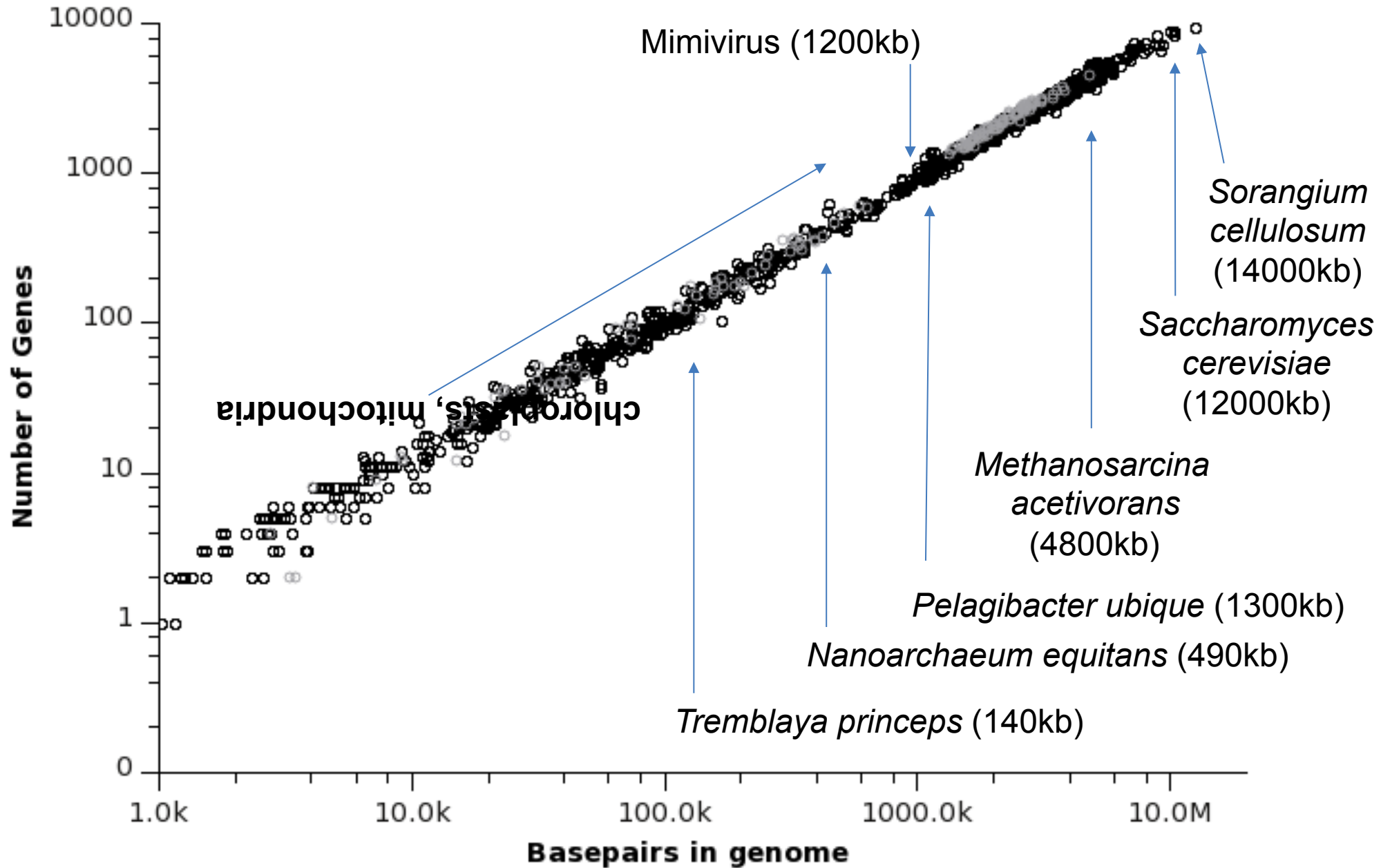
Smallest archaeon



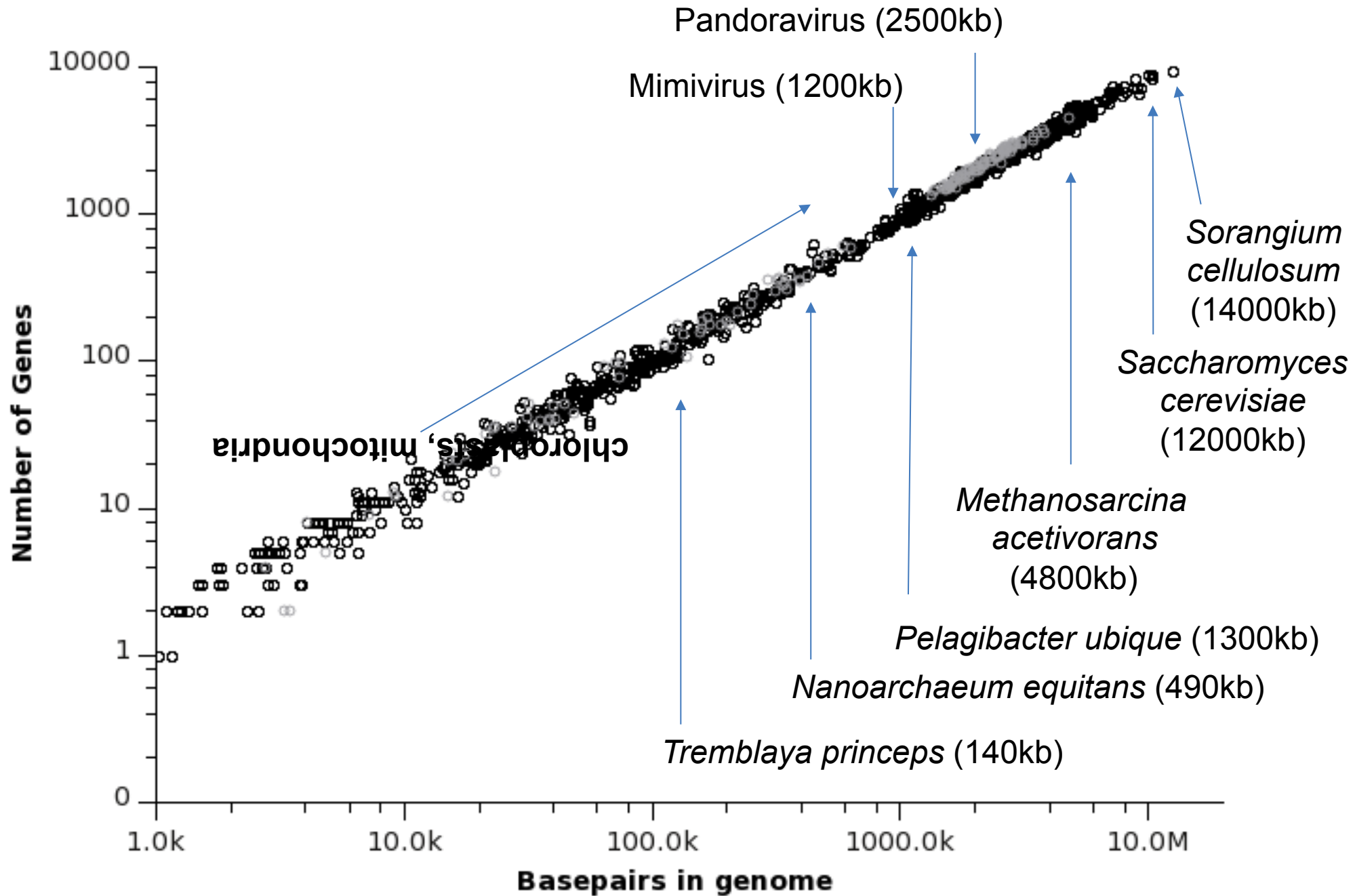
Smallest eukaryote



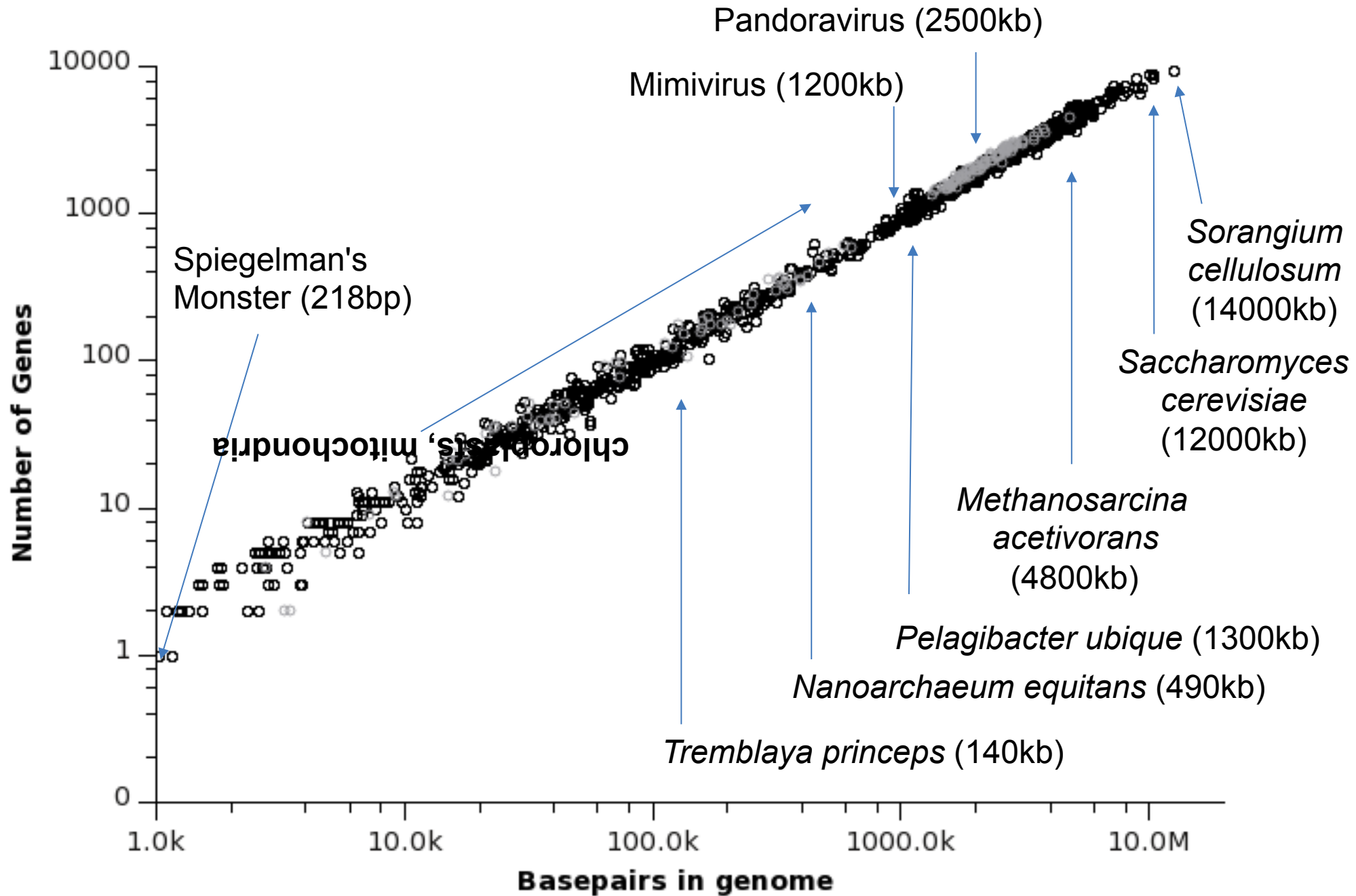
Largest virus



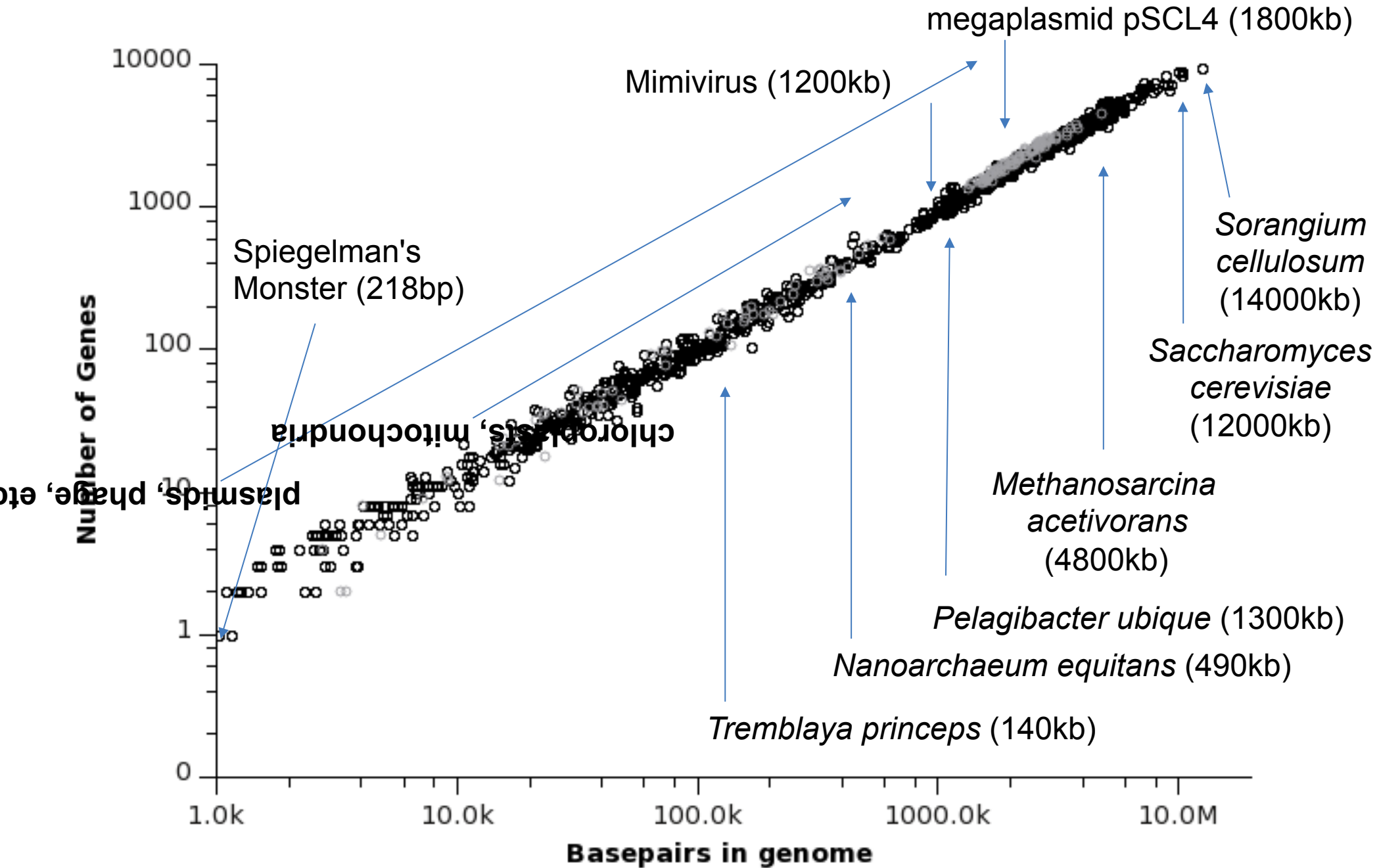
Largest virus (Oct 2013)



Smallest biological entity



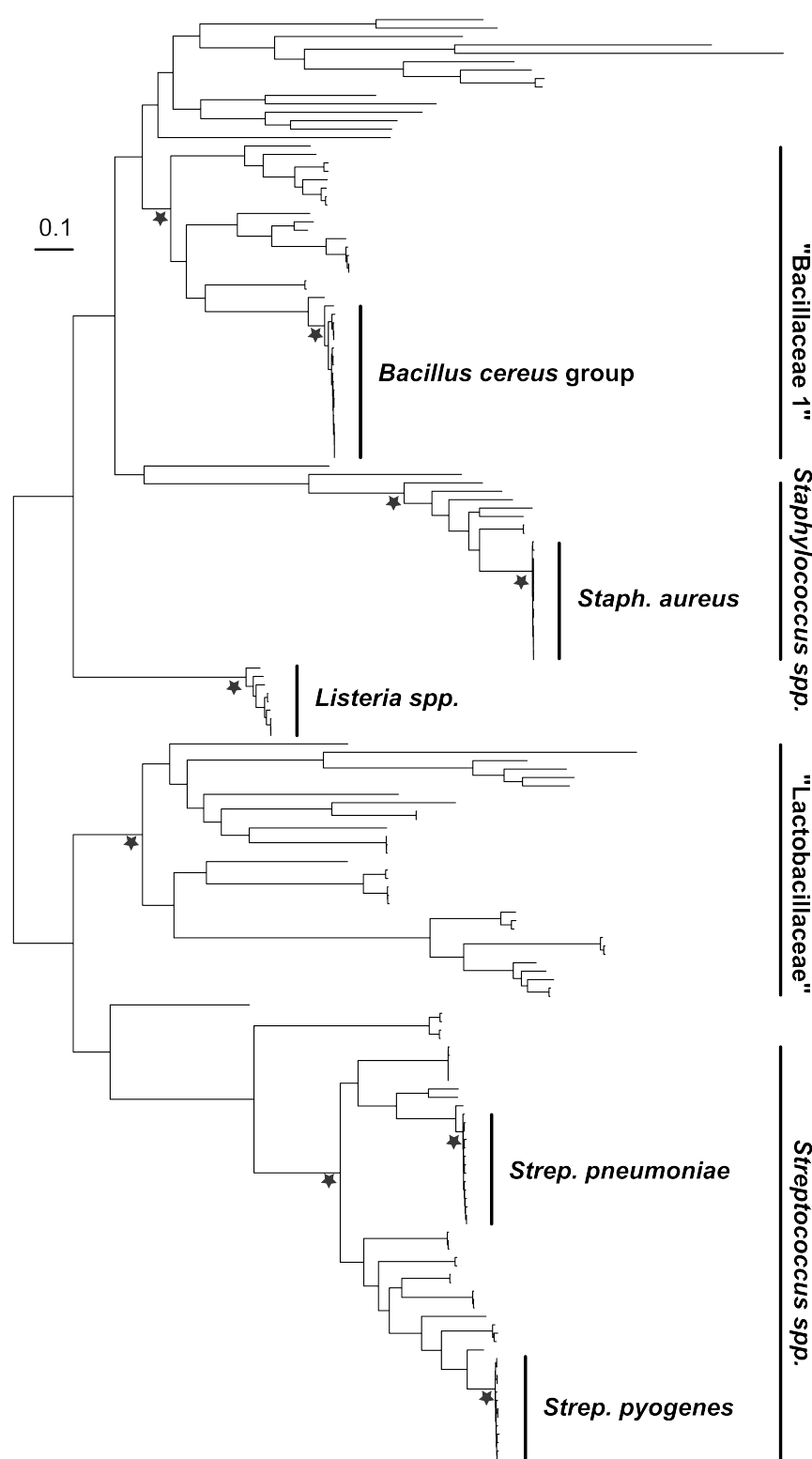
Largest plasmid



λ Genome Modeling

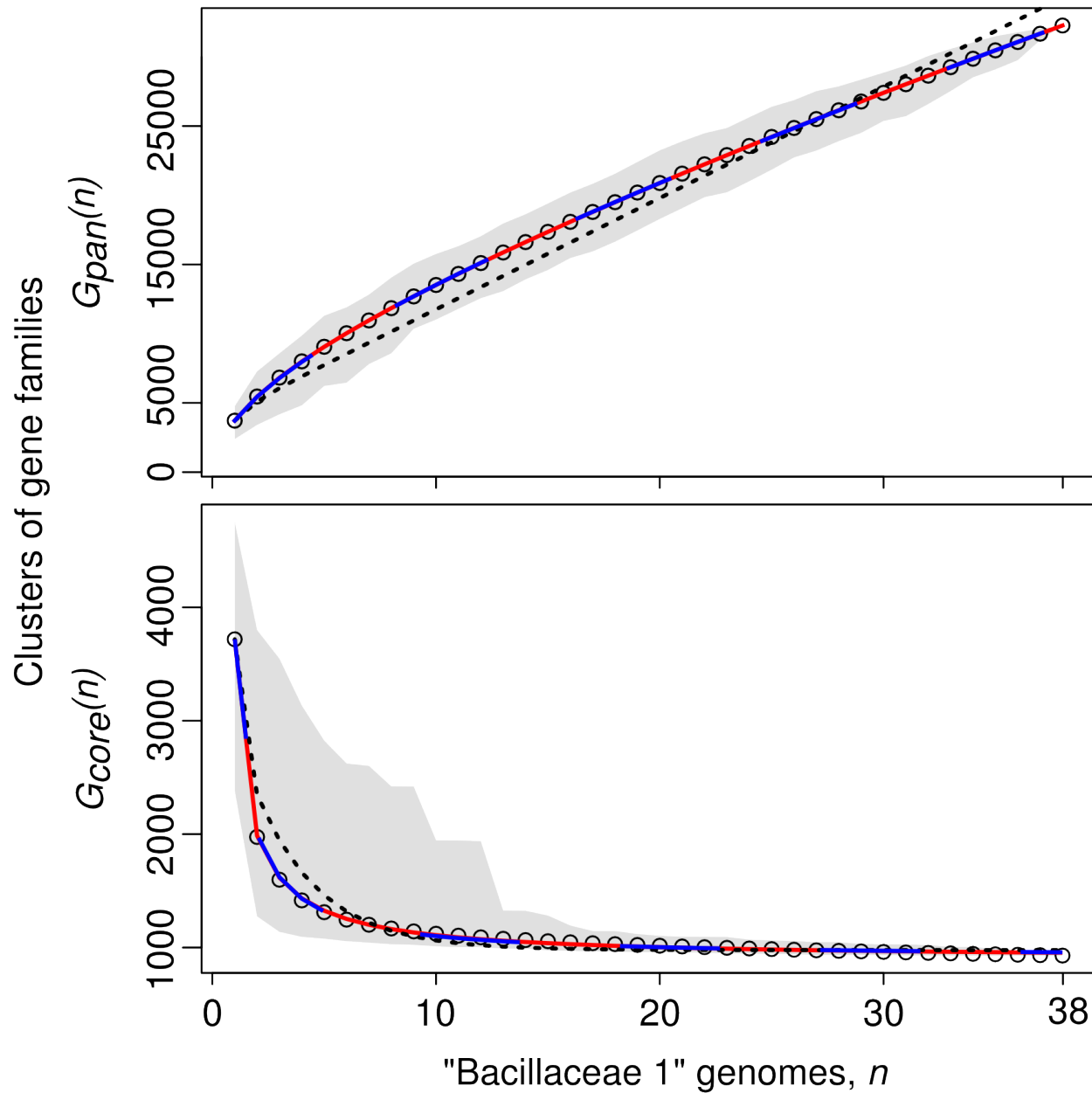


Bacterial Pangenome

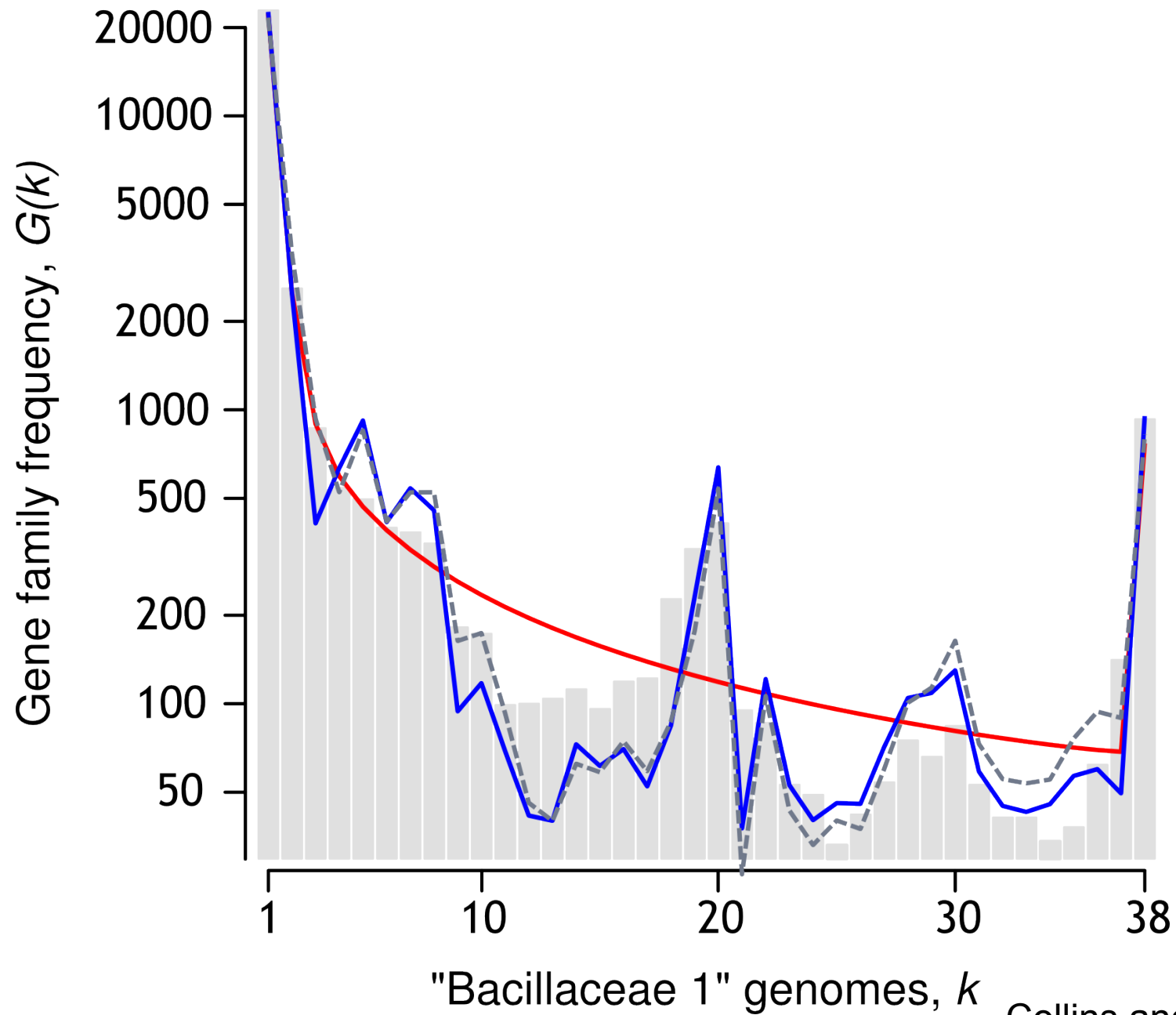


- sequence many isolates of the same species
- compare them to find out what has changed

Core and Pangenomes



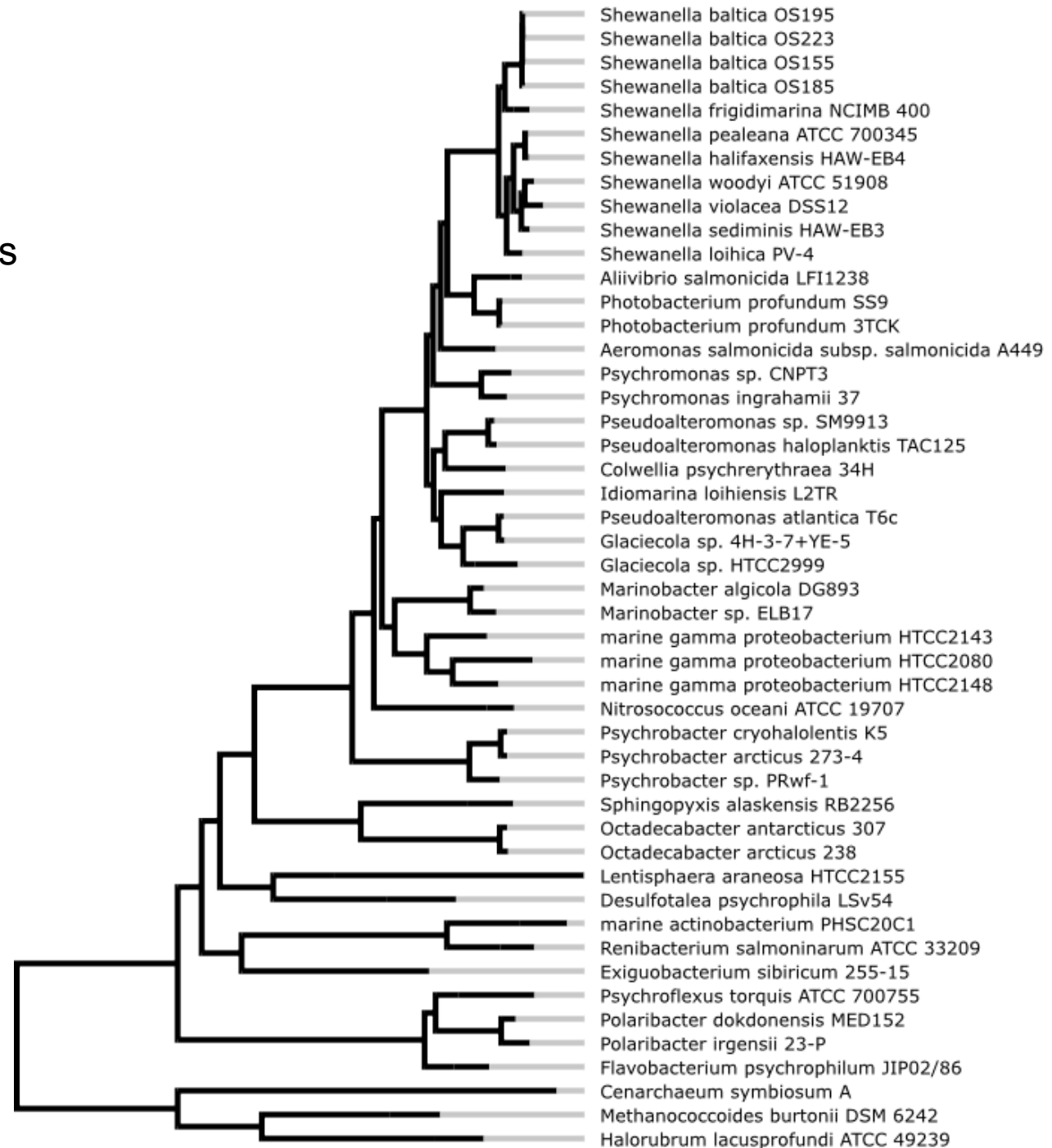
Core and Pangenomes



λ hima/ddmp



hima v1
genomes
(2011)



[illegible]

Which Microbes Grow At Low Temperature?



*Bergey's
Manual
of*

Systematic Bac

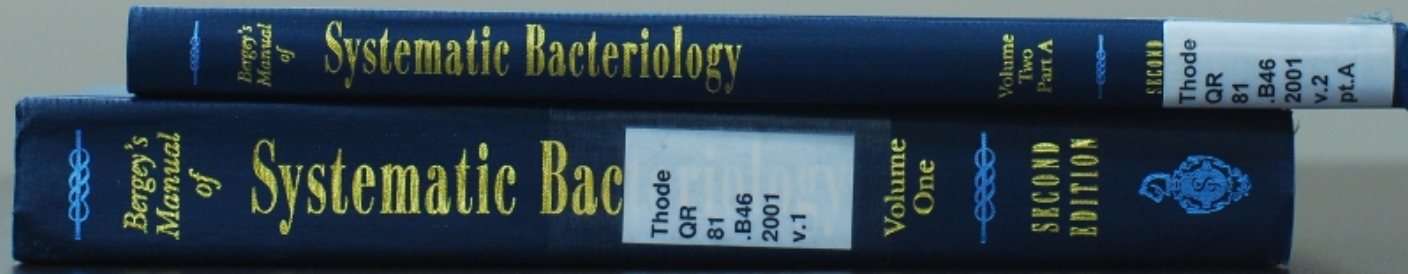
Thode
QR
81
.B46
2001
v.1

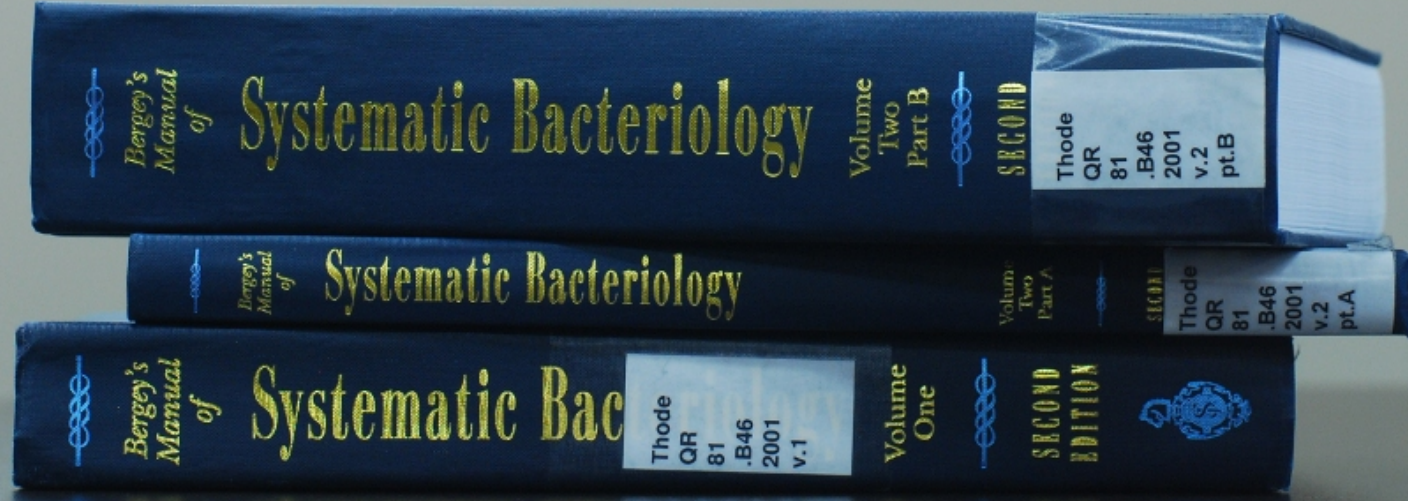
Volume
One

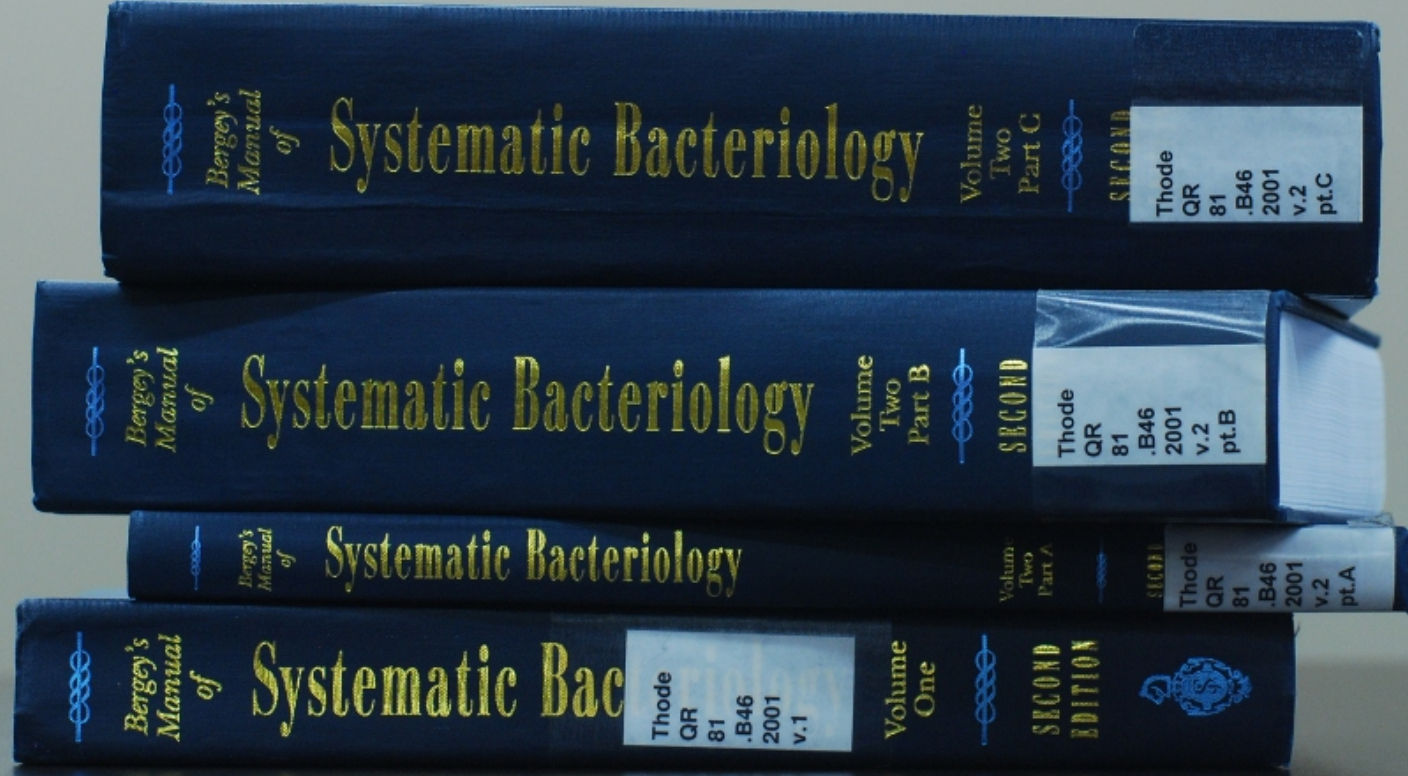


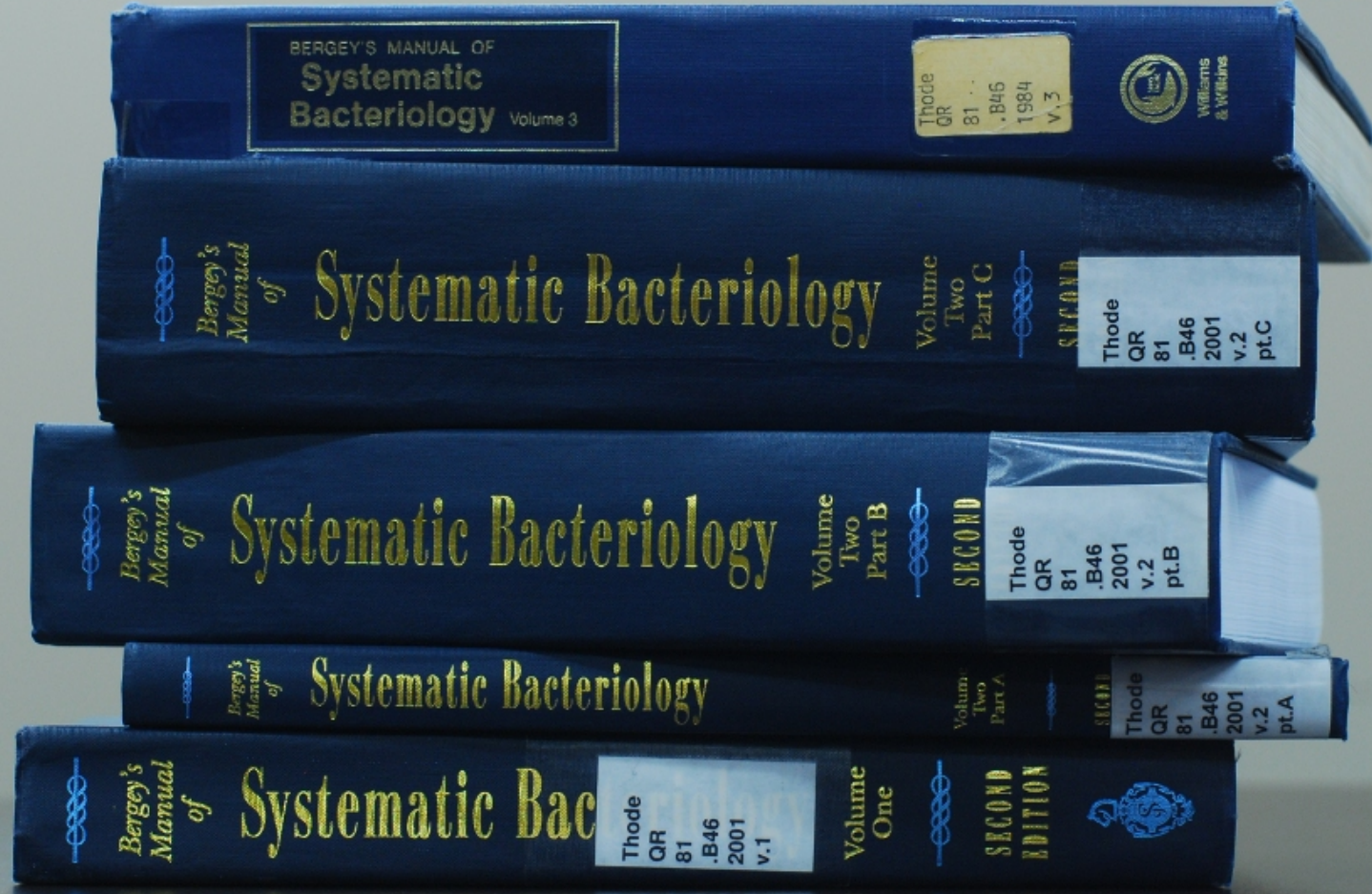
SECOND
EDITION

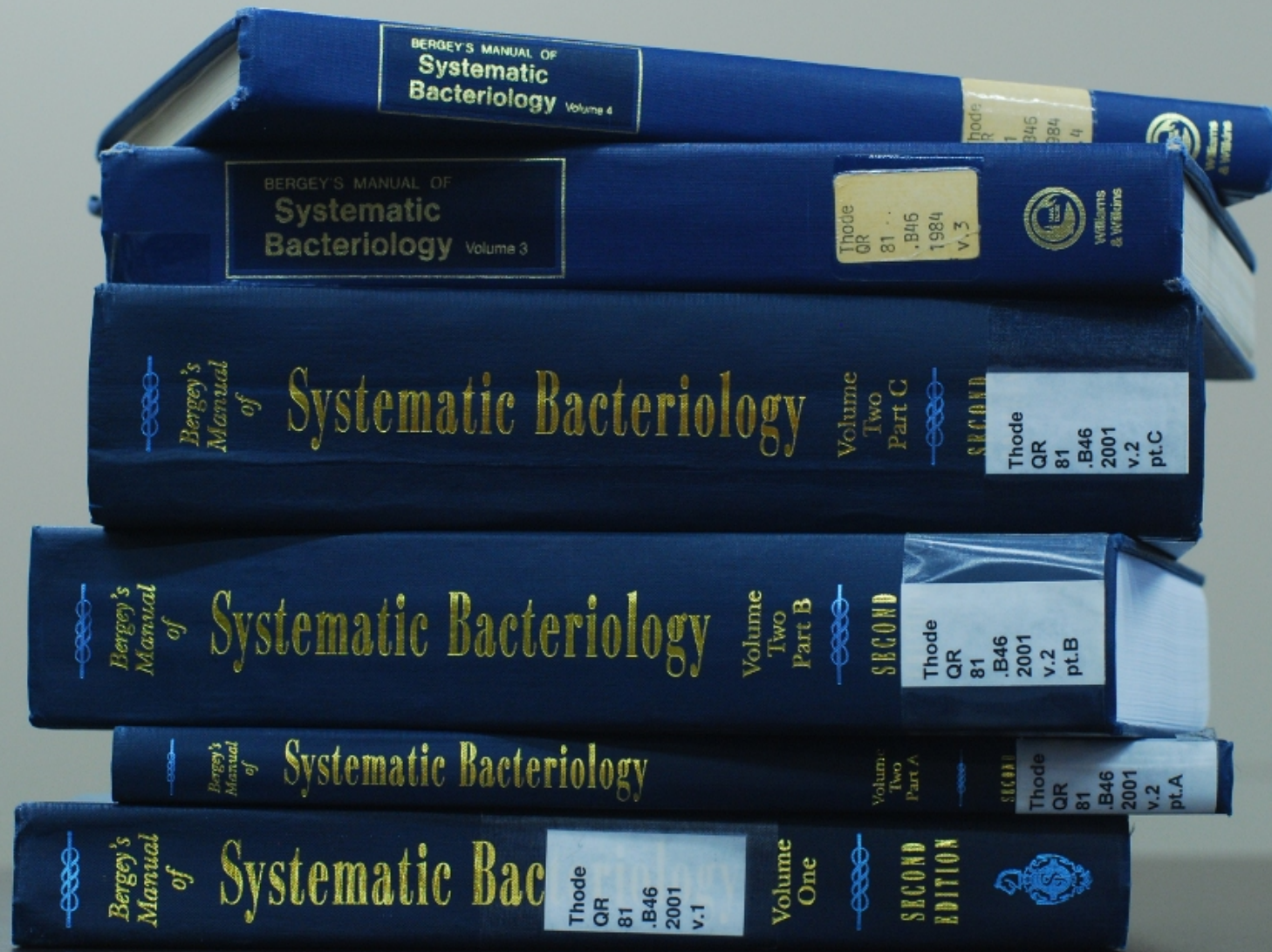












ddmp: a digital database of microbial phenotypes

- λ 100 data tables
- λ 450 taxa
- λ 1000 traits
- λ 22,000 entries
- λ *<10% of the total*
- λ *<http://reric.org/wo>*



ddmp: species



ddmp: traits



Bergey's, etc.

dump

Find data for genera or higher ranks:

Smithella
Stanierella
Stenothermobacter
Subsaxibacter
Subsaximicrobium
Syntrophobacter
Syntrophus
Tenacibaculum
Thermodesulforhabdus
Tolumonas

Find data for species or lower ranks:

Desulfuromusa bakii
Desulfuromusa kysingii
Desulfuromusa succinoxidans
Exiguobacterium acetylicum
Exiguobacterium antarcticum
Exiguobacterium aurantiacum
Exiguobacterium undae
Flavobacterium antarcticum
Flavobacterium aquatile
Flavobacterium aquidurense

Having phenotypes:

growth: salinity; optimum (g/l)
growth: salt tolerance
growth: sea salts; tolerates (200%)
growth: shigella-salmonella agar
growth: sulfur; required
growth: tcbs medium
growth: temperature; maximum
growth: temperature; minimum
growth: temperature; optimum
growth: tryptic soy agar

☒ Raw Data

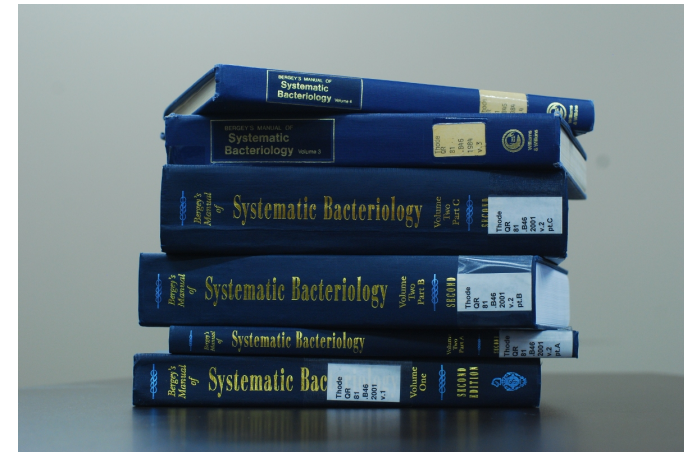
☒ HTML

Get Table

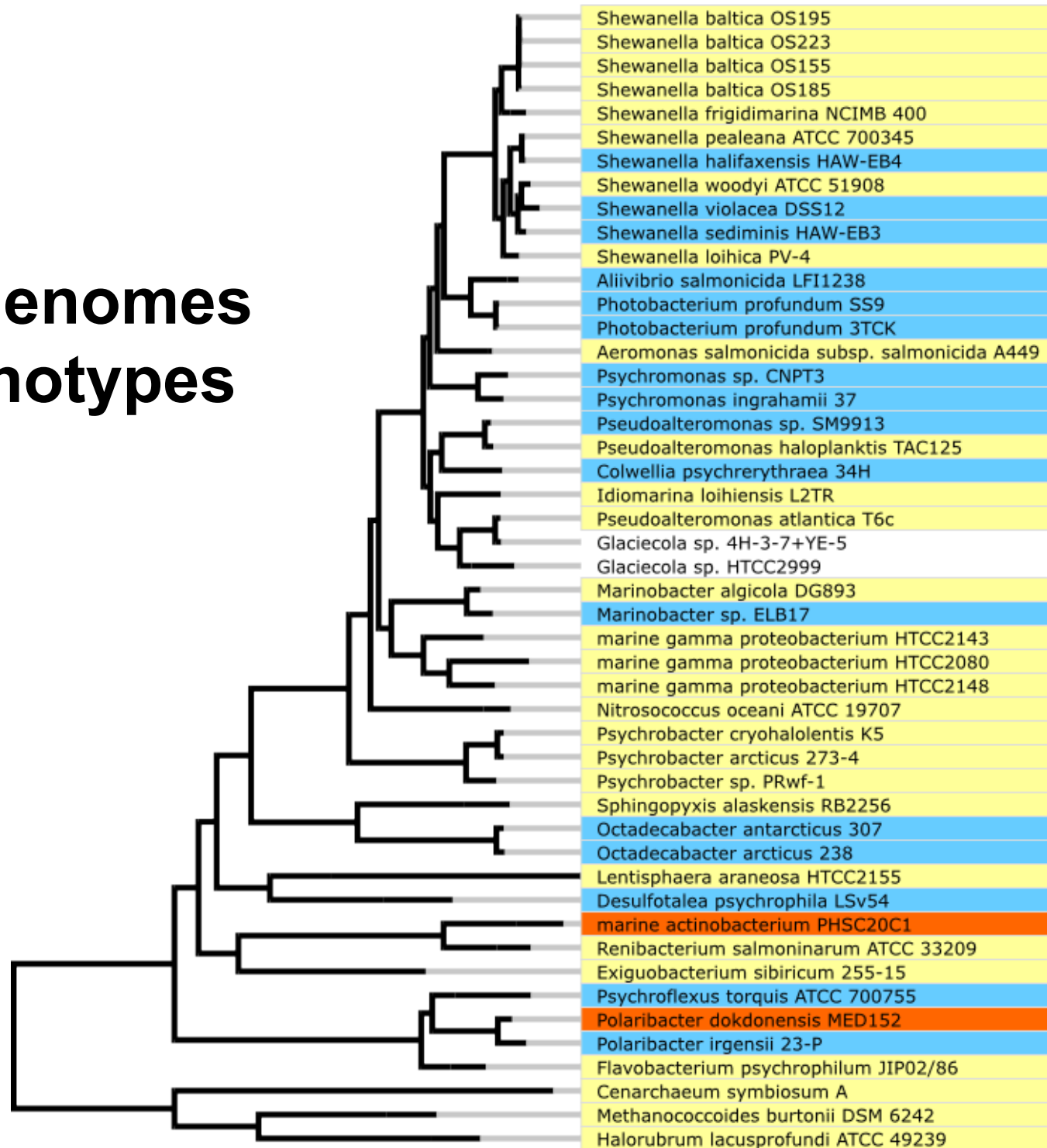
☐ Nominal Data

☐ CSV

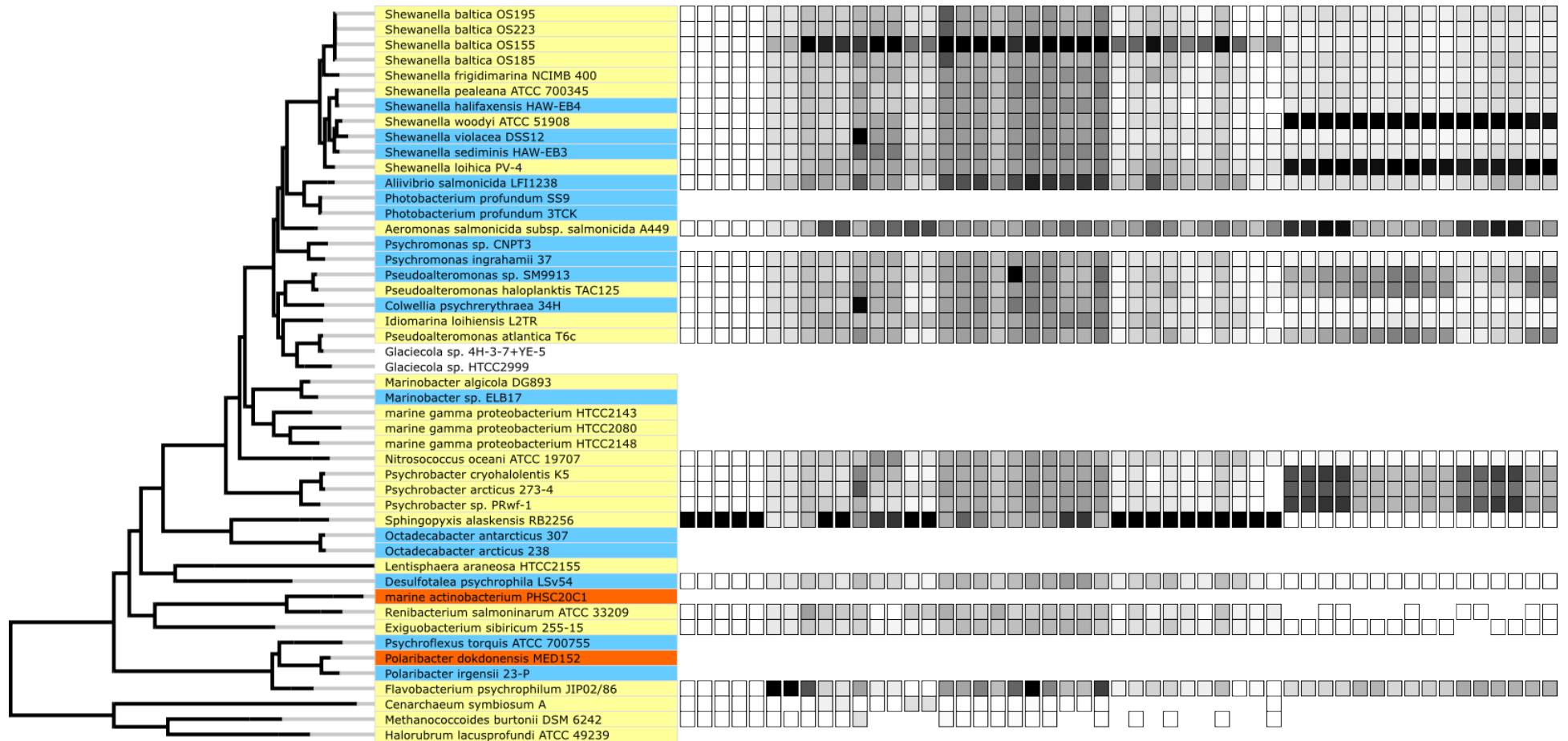
	growth	growth	growth
	temperature; maximum	temperature; minimum	temperature; optimum
Flavobacterium antarcticum	24	5	21
Flavobacterium aquatile	>30	<15	nd
Flavobacterium aquidurense	30	13	19-28
Flavobacterium branchiophilum	30	5	18-25
Flavobacterium columnare	37	15	25-30
Flavobacterium croceum	45	10	25-35



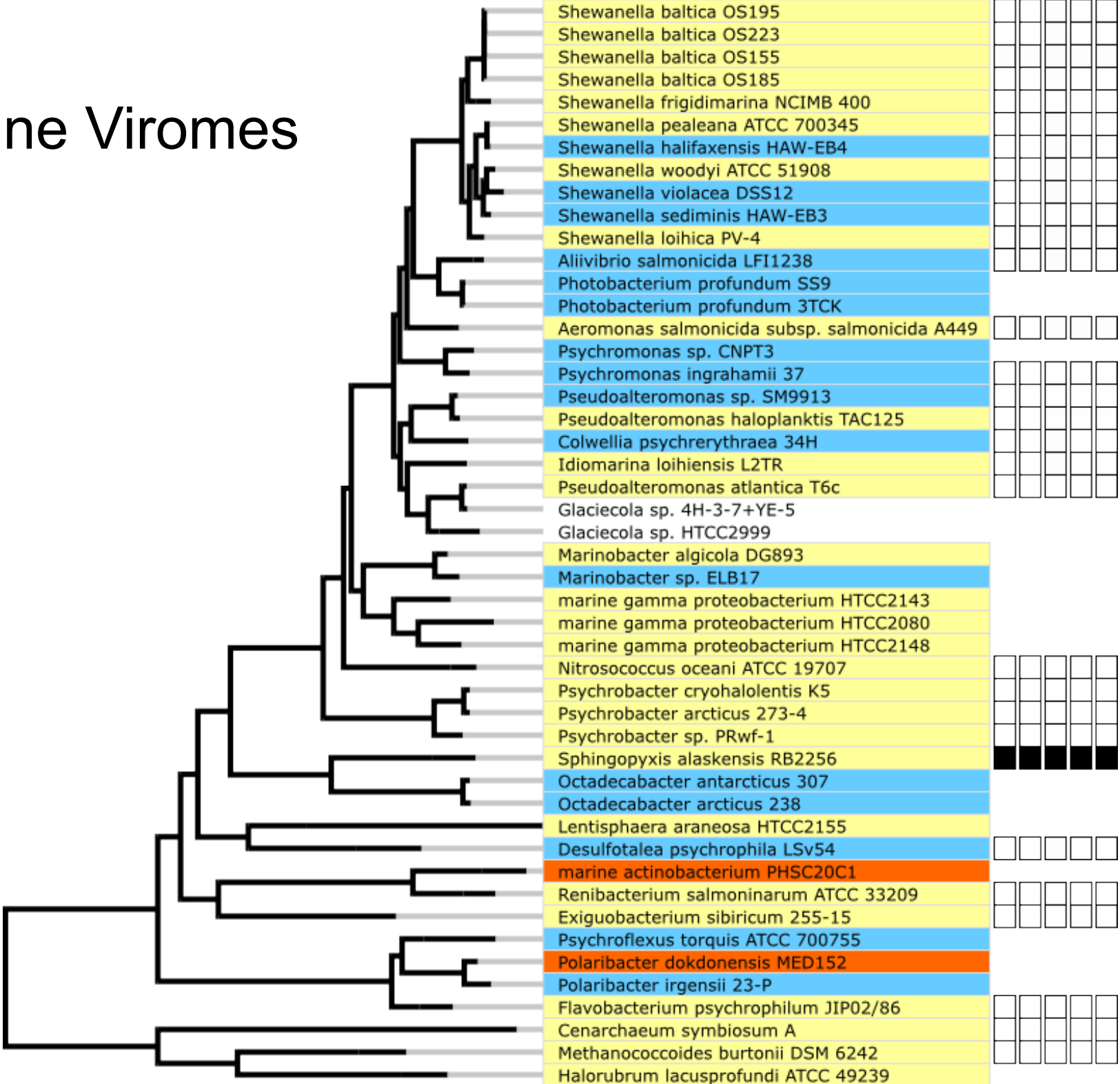
hima v1 genomes with phenotypes



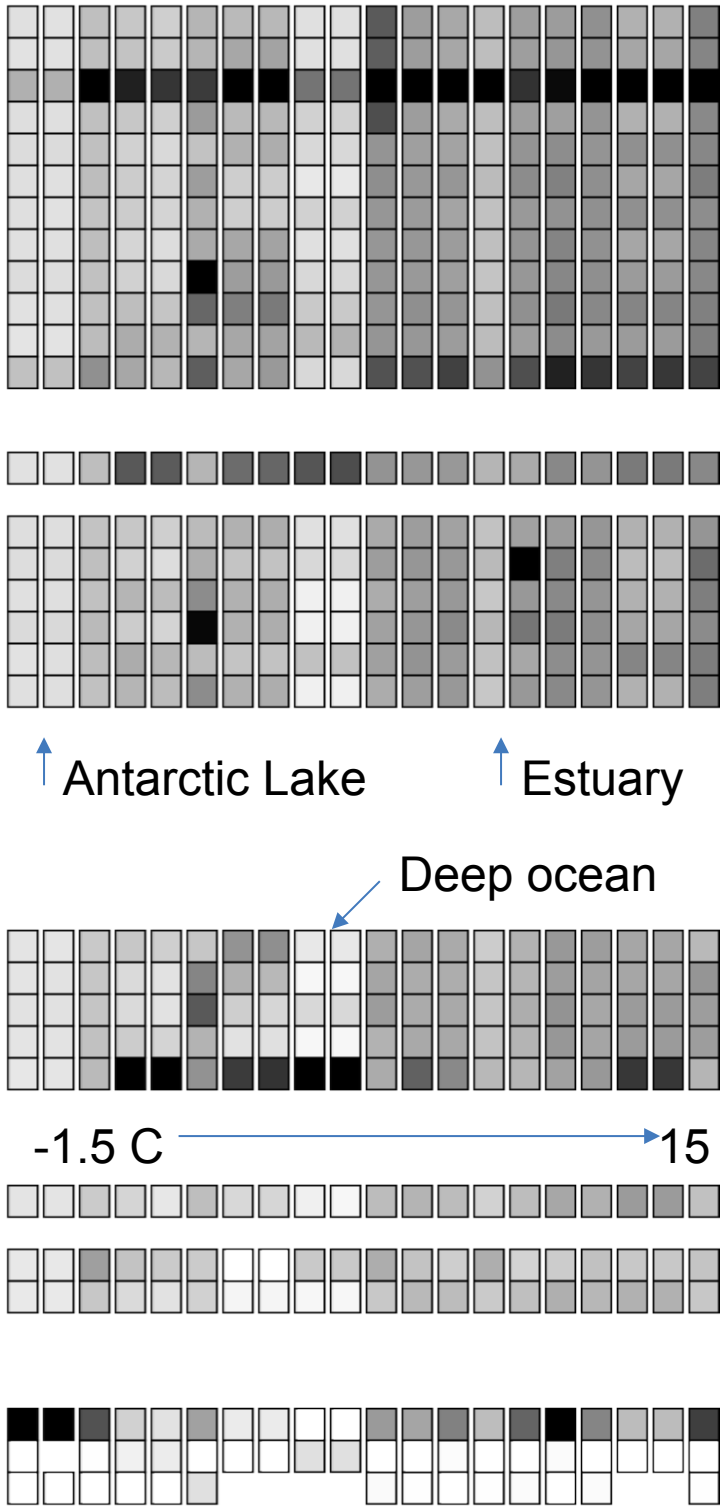
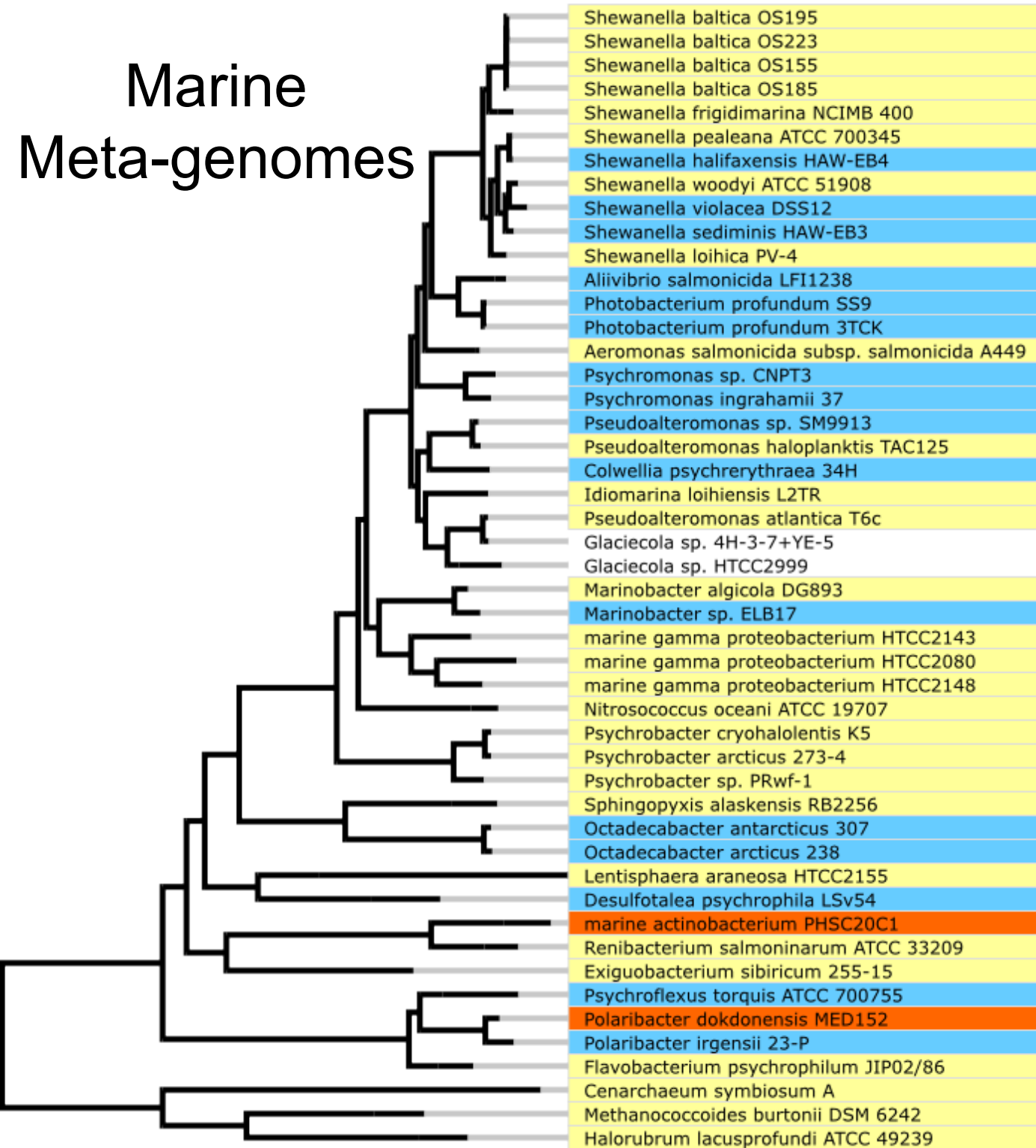
-15 to +15°C



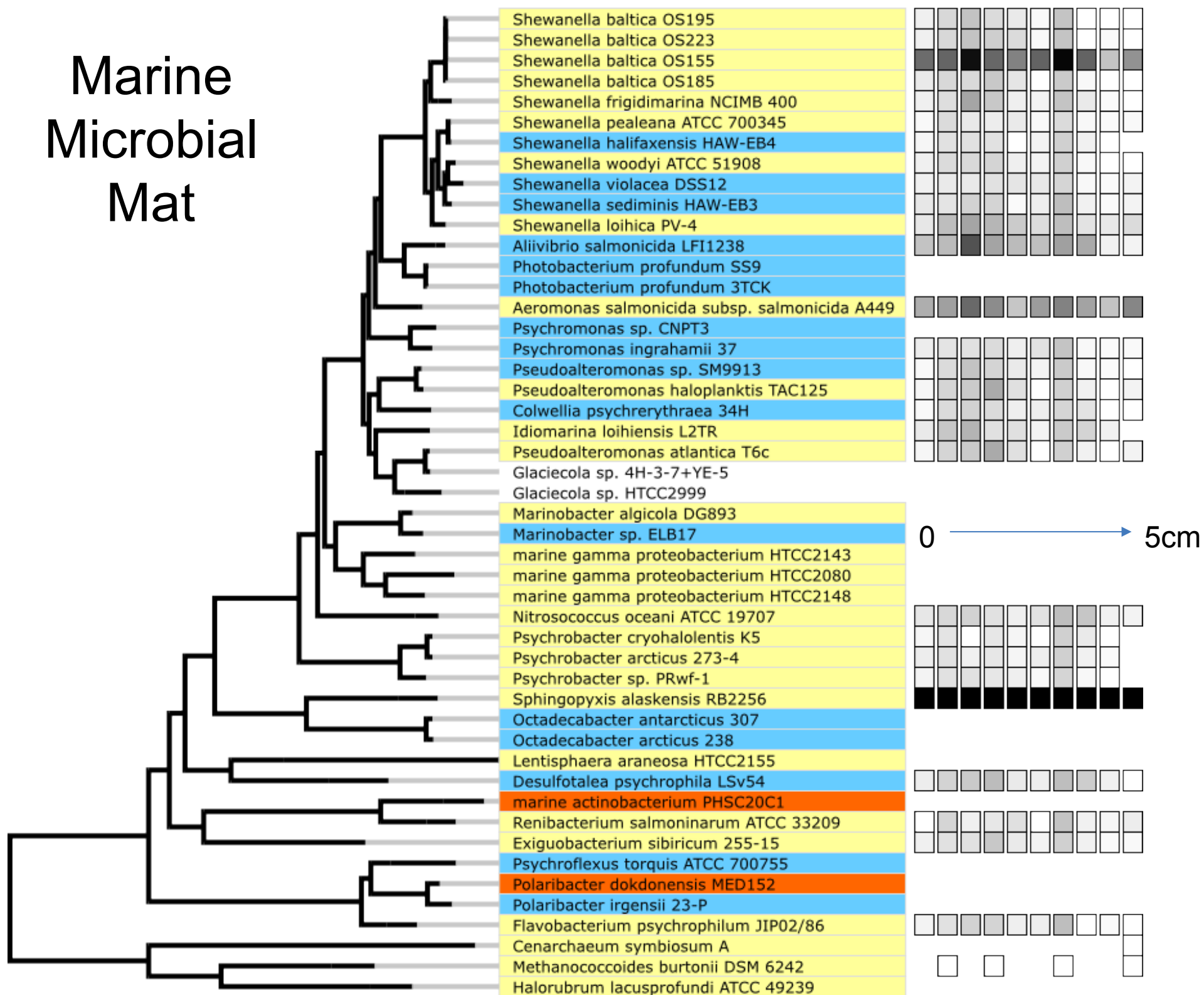
Marine Viromes



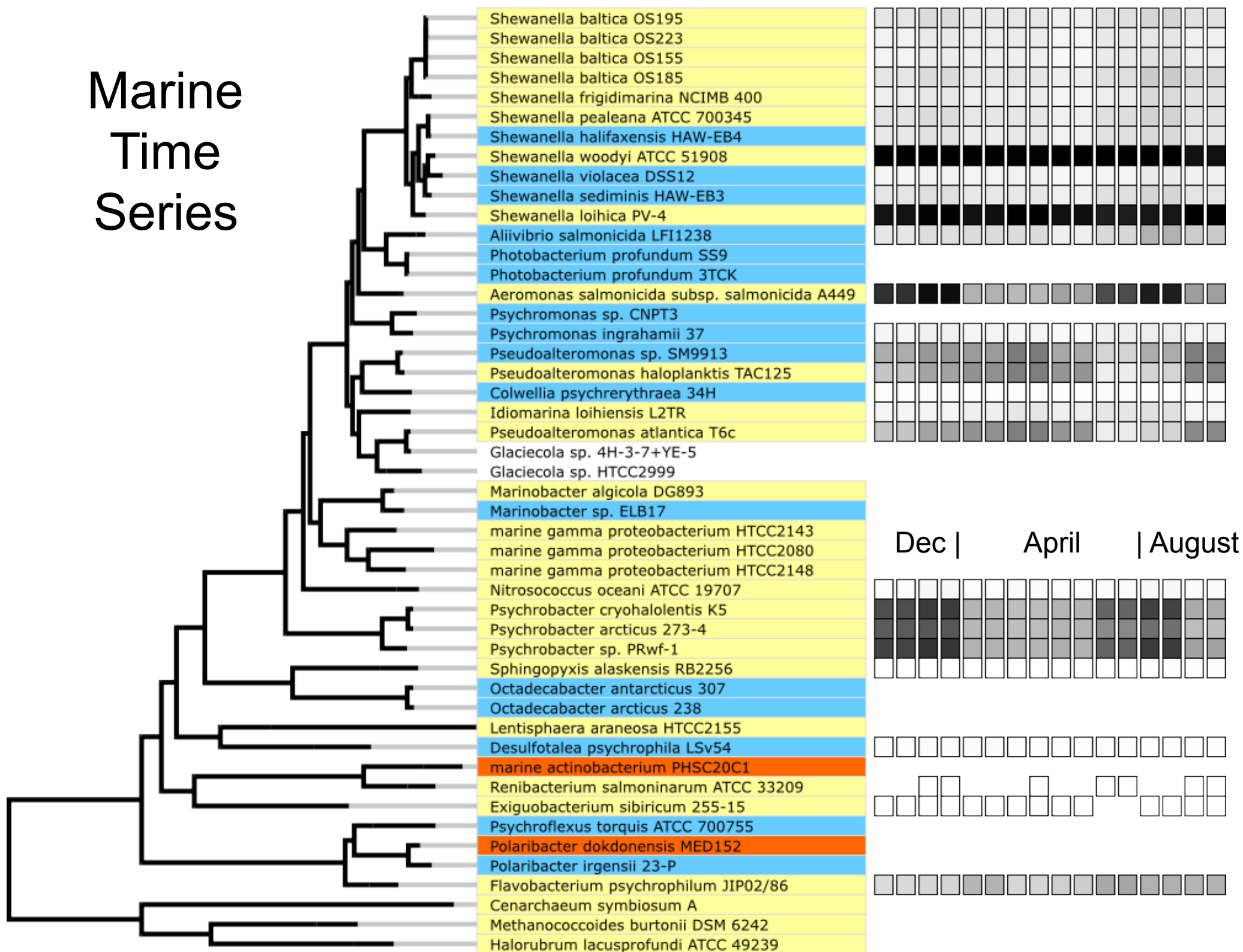
Marine Meta-genomes

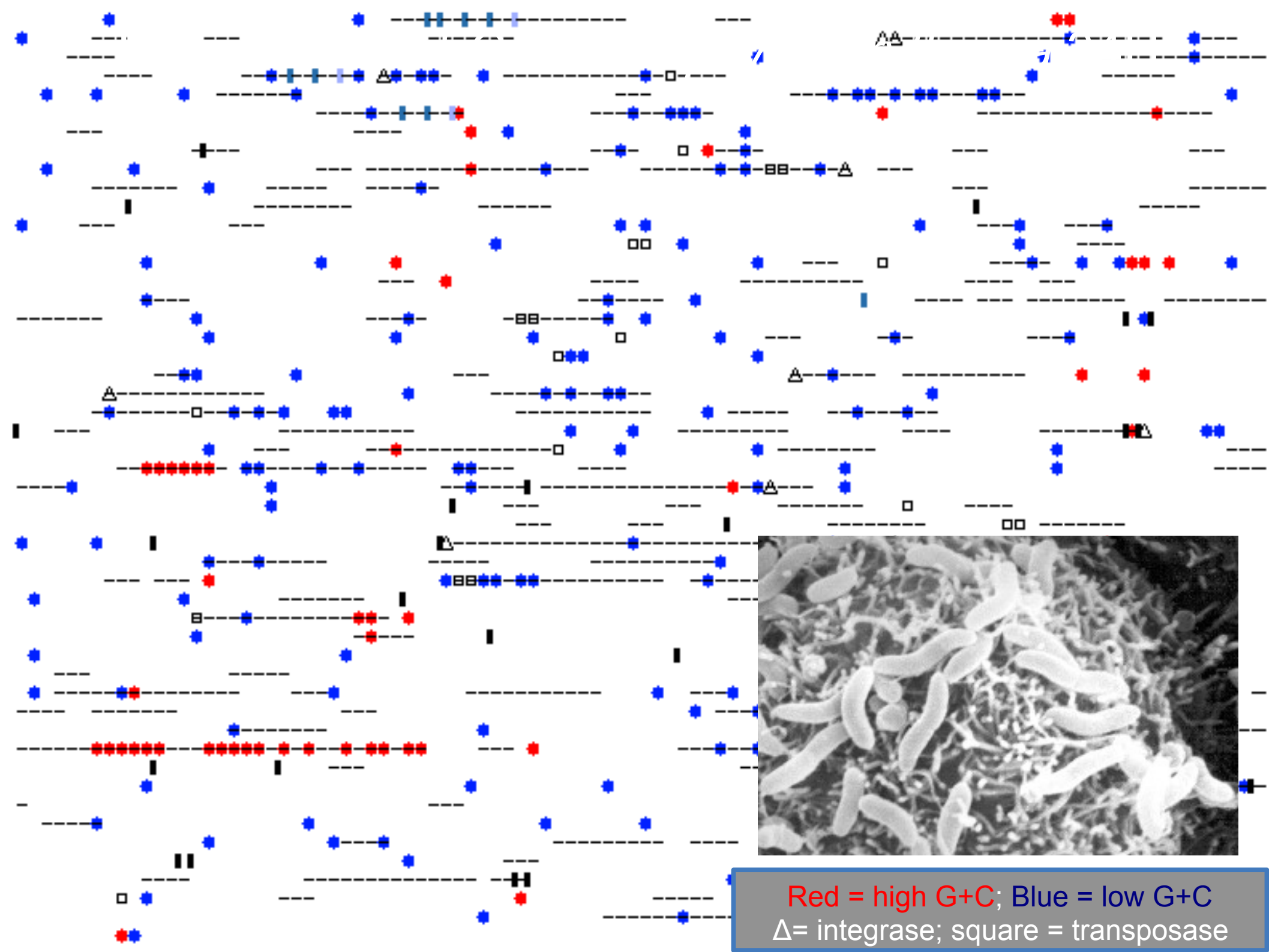


Marine Microbial Mat



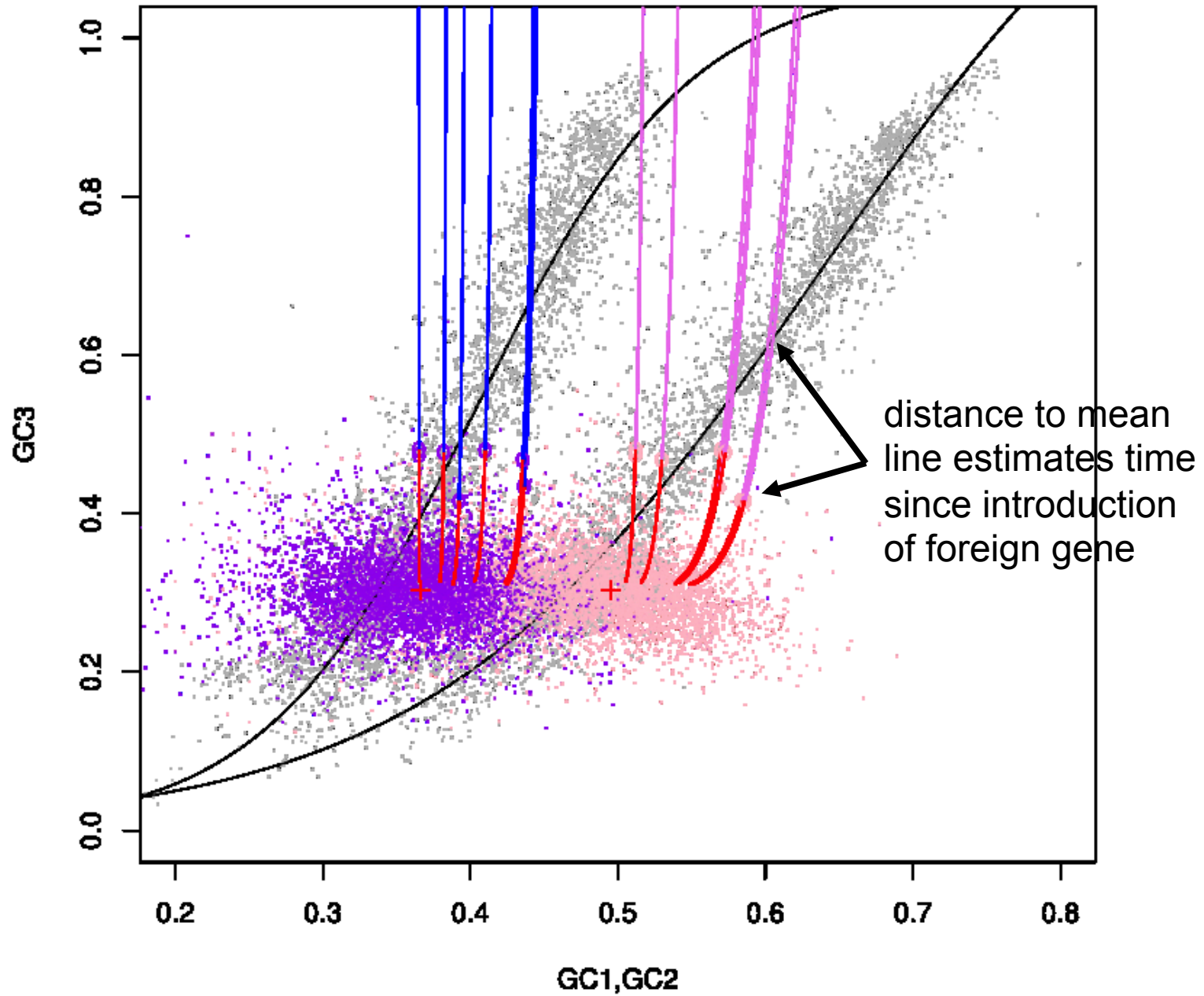
Marine Time Series



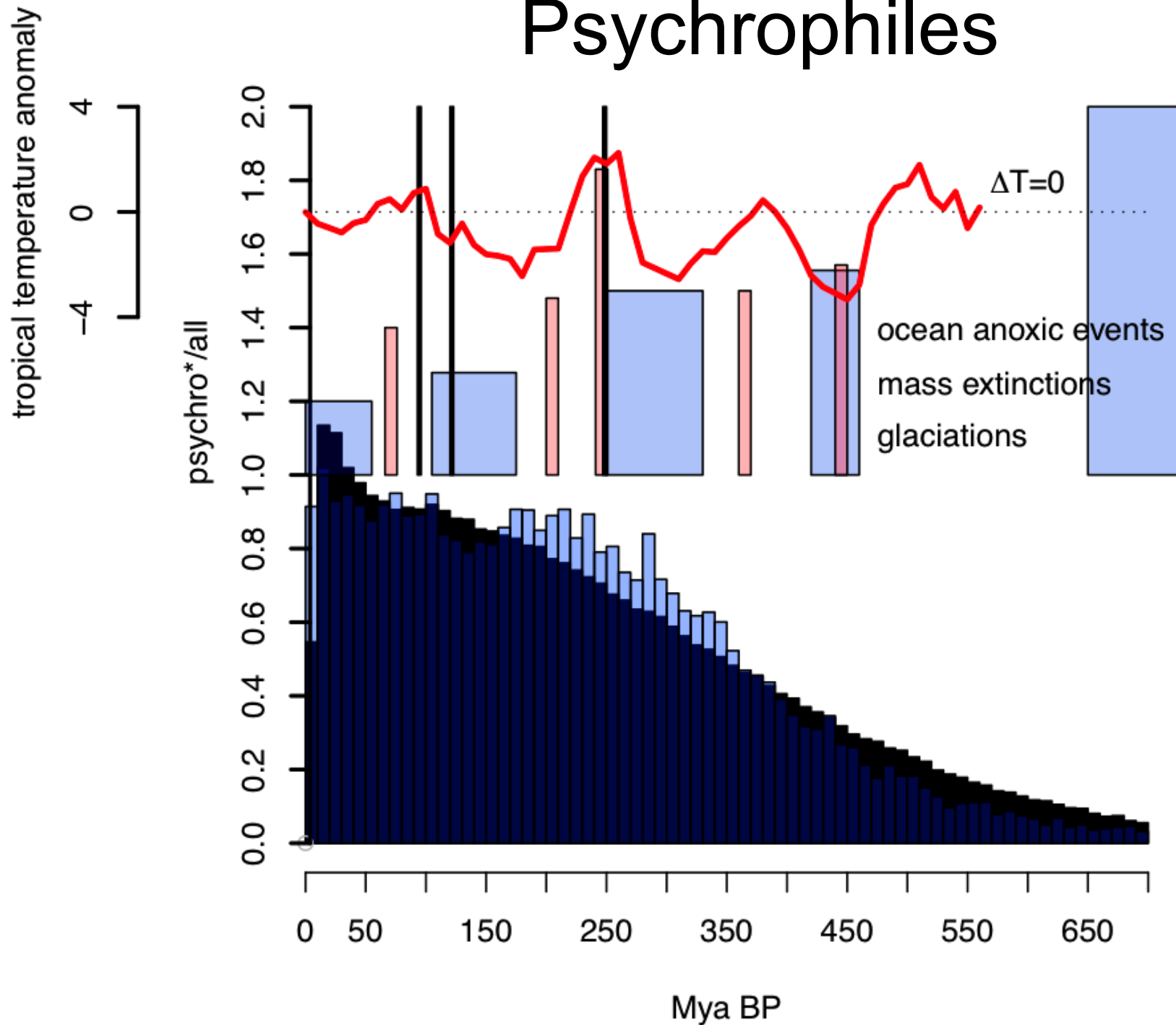


Red = high G+C; Blue = low G+C
Δ = integrase; square = transposase

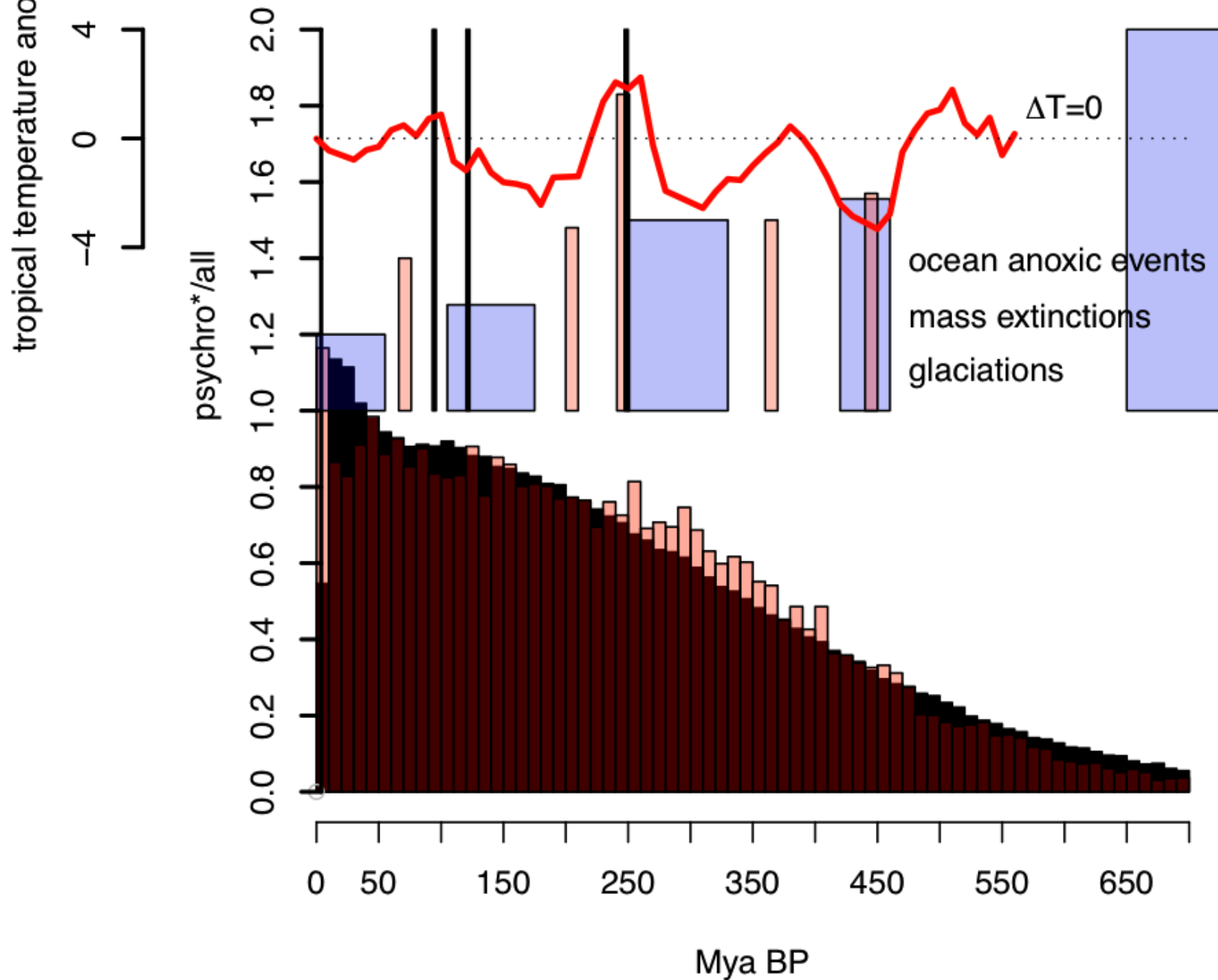
GC content amelioration



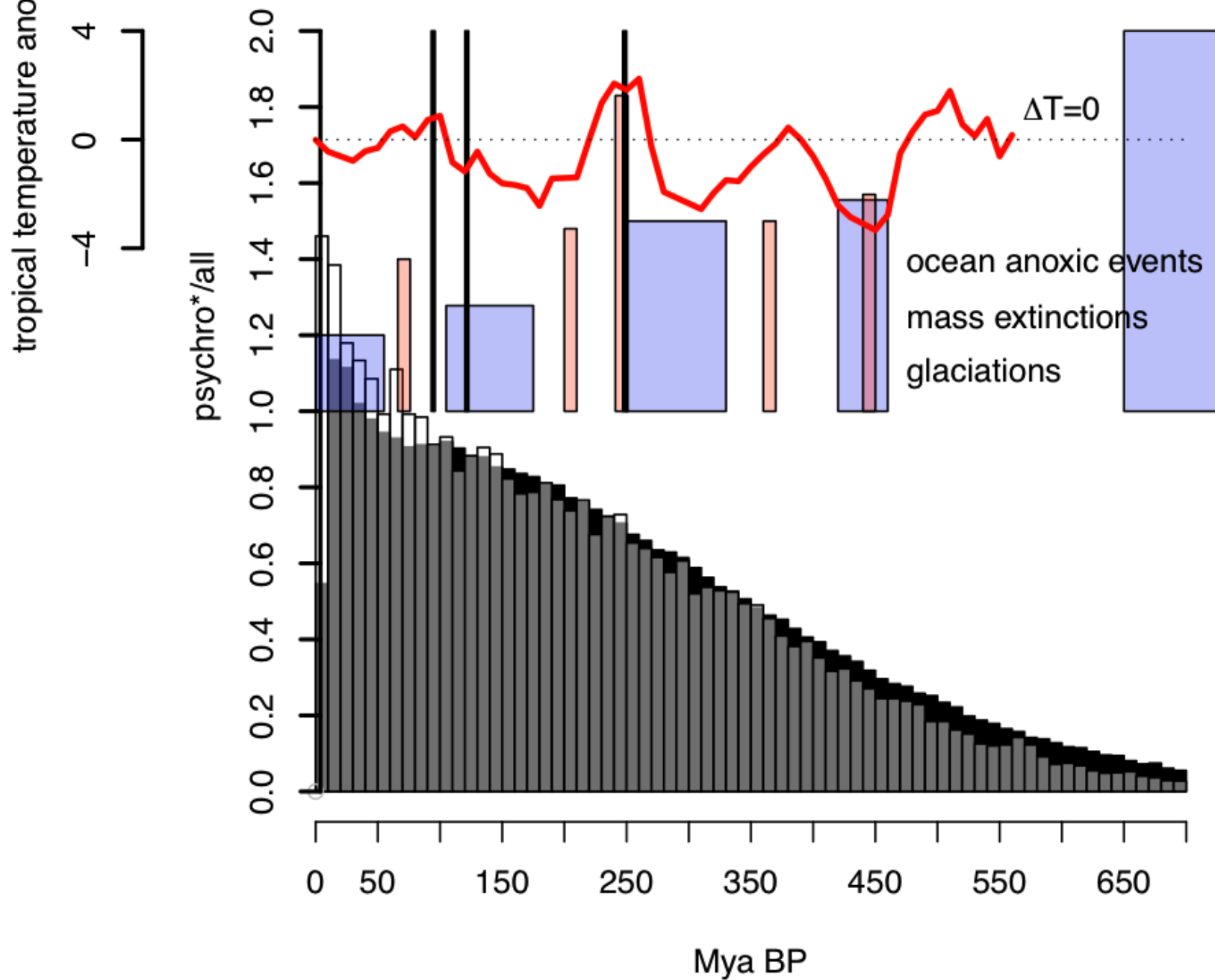
Psychrophiles



Thermophiles

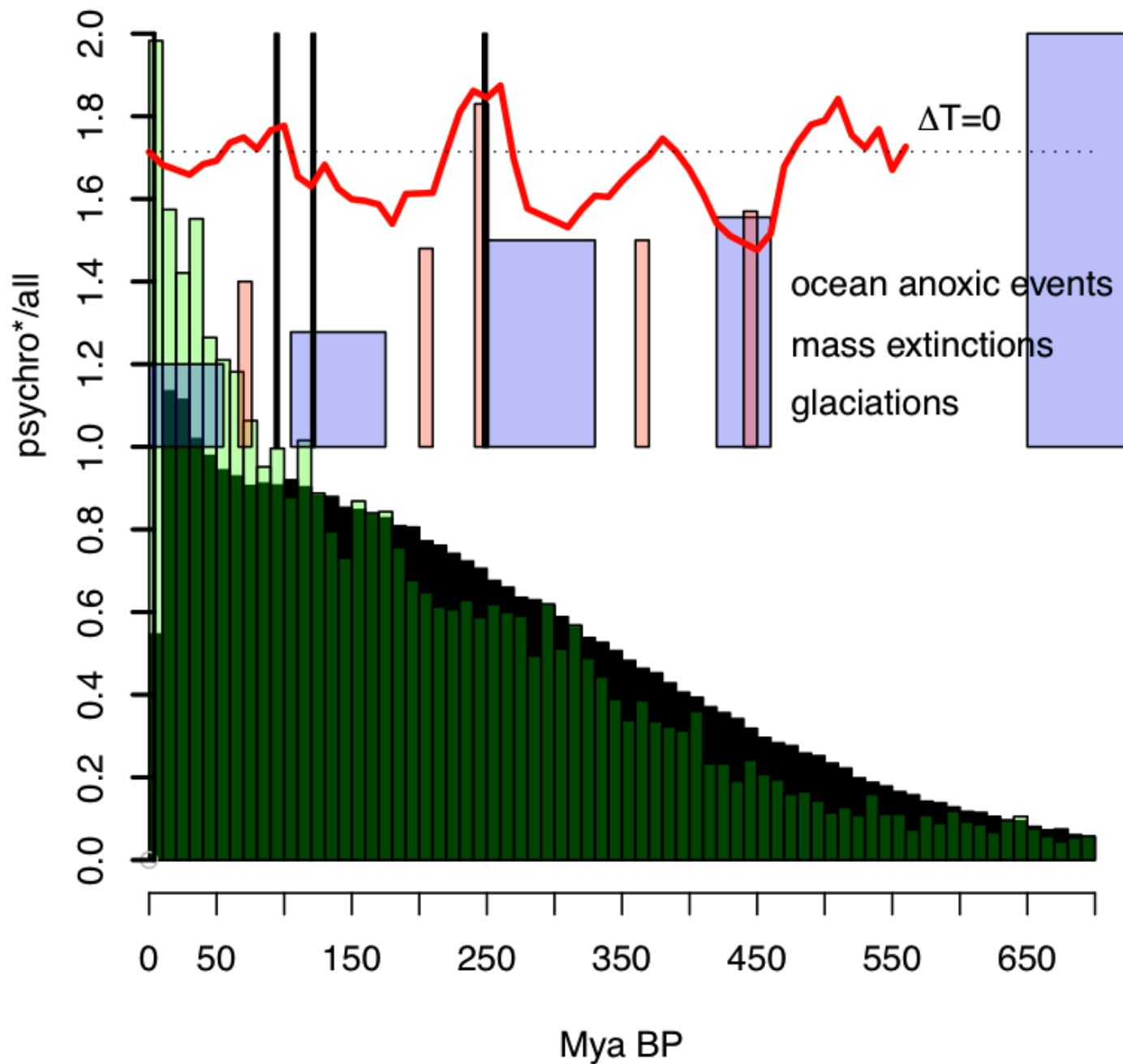
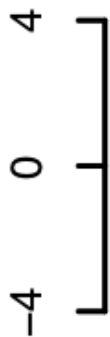


Sulfate Reducers

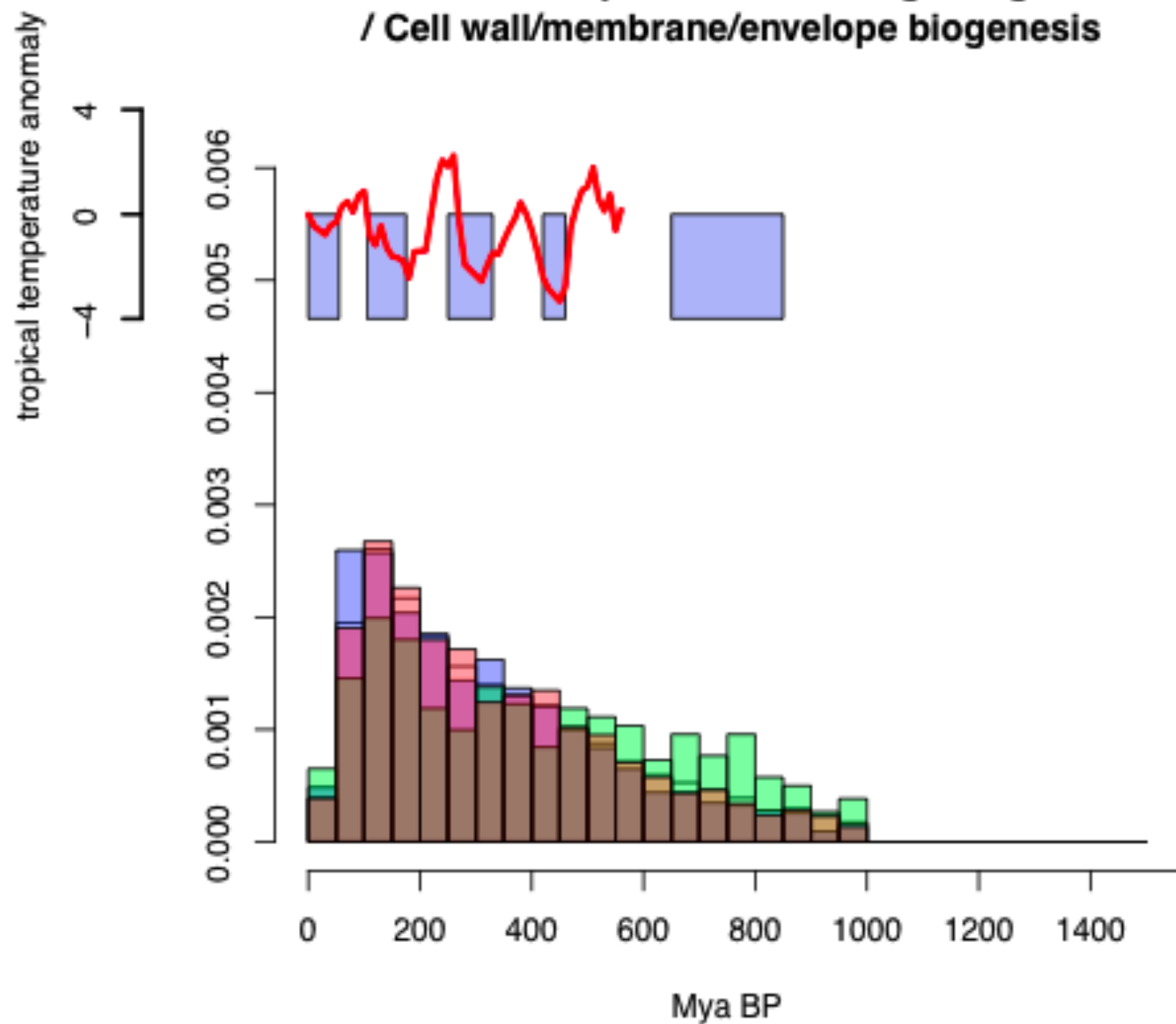


Cyanobacteria

tropical temperature anomaly



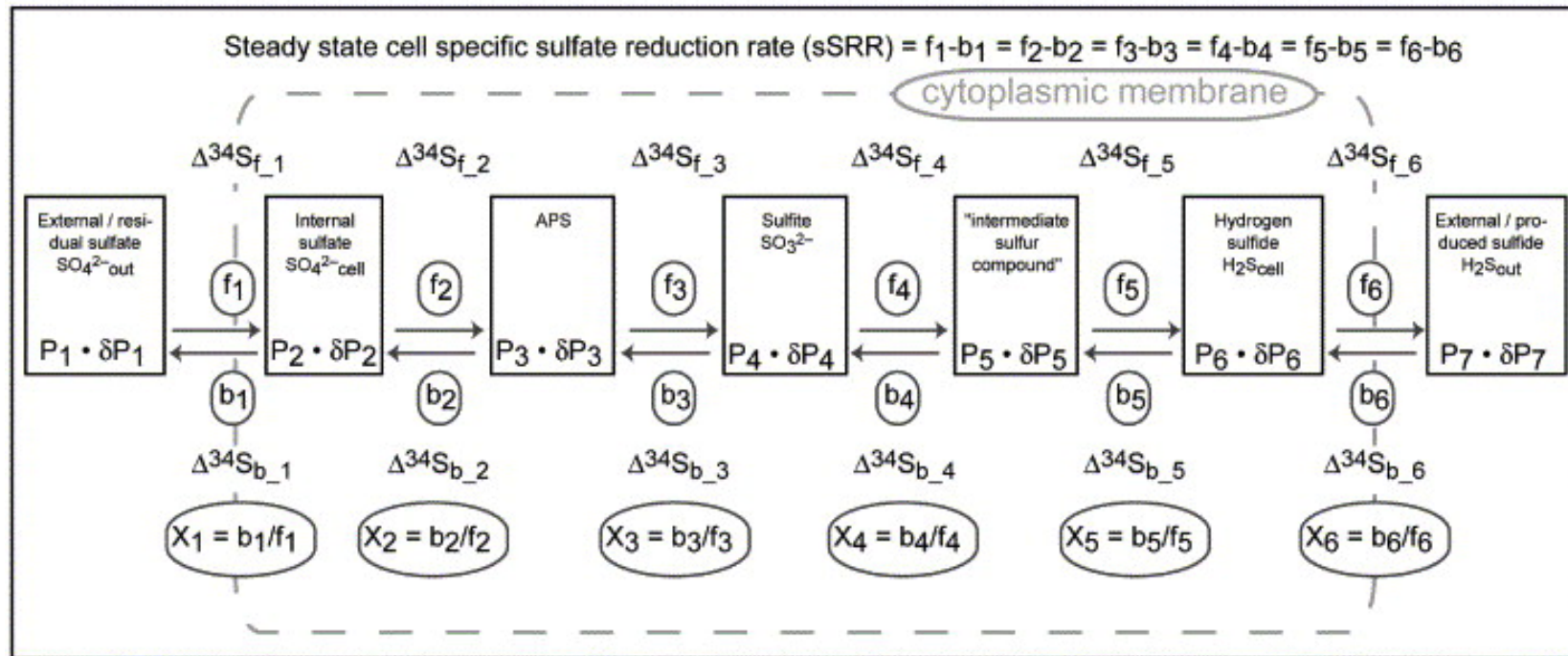
**M / Cellular processes and signaling /
/ Cell wall/membrane/envelope biogenesis**



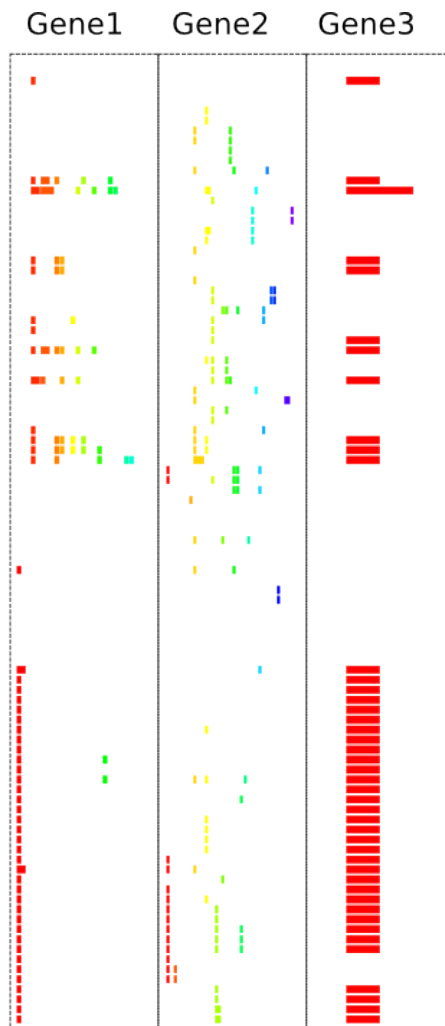
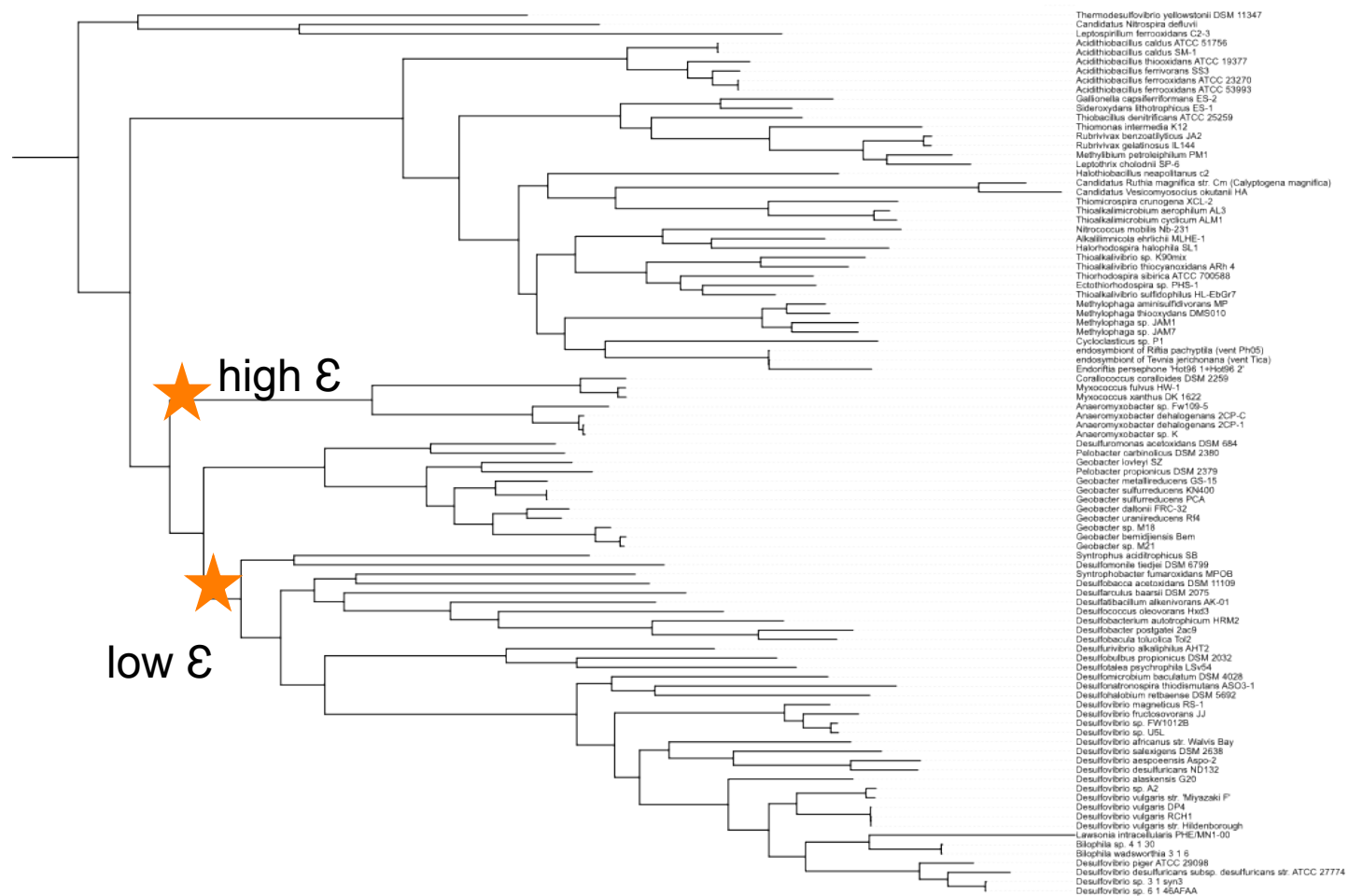
λ SRB



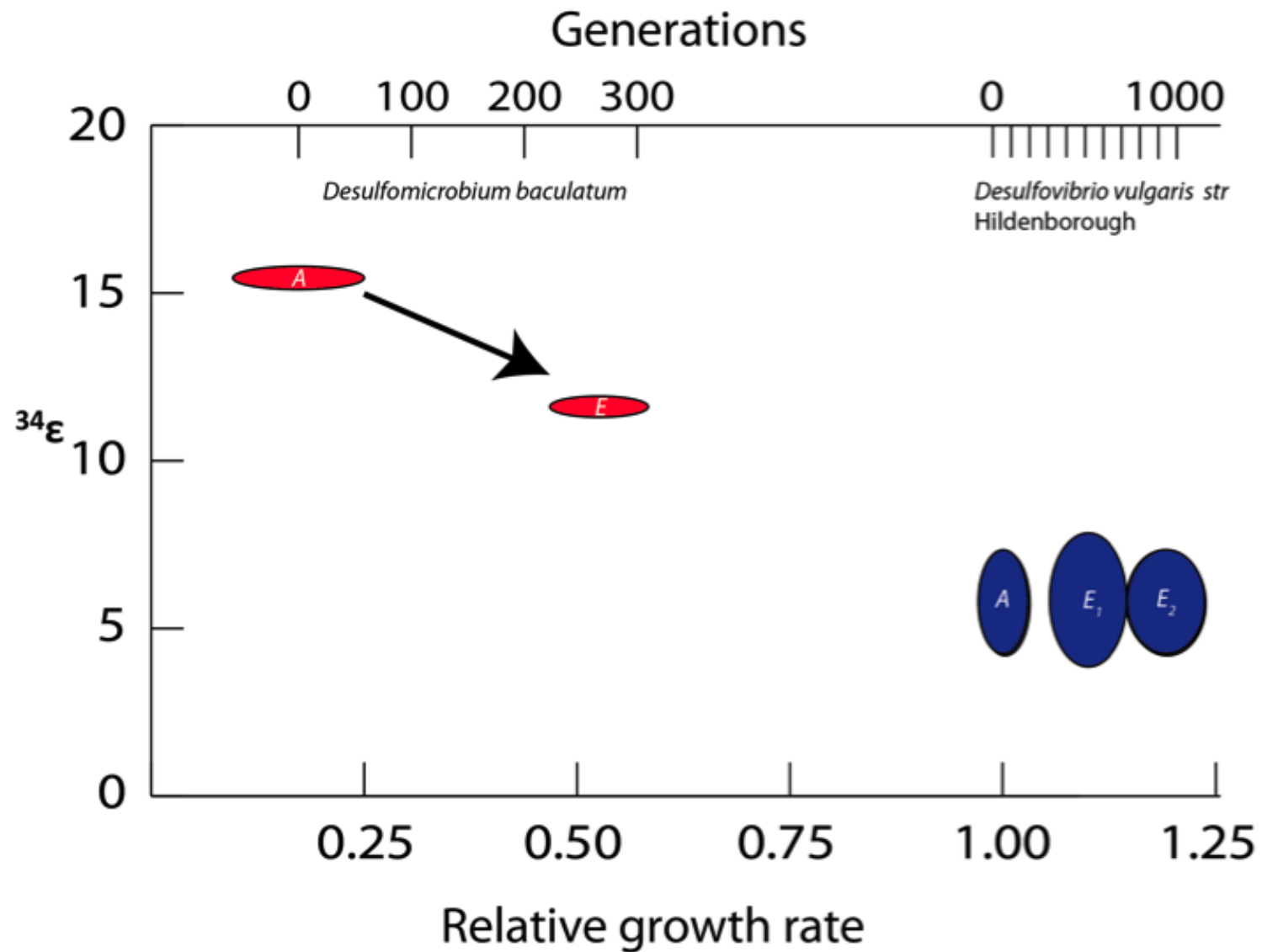
Sulfur Isotope Fractionation by sulfate reducing bacteria



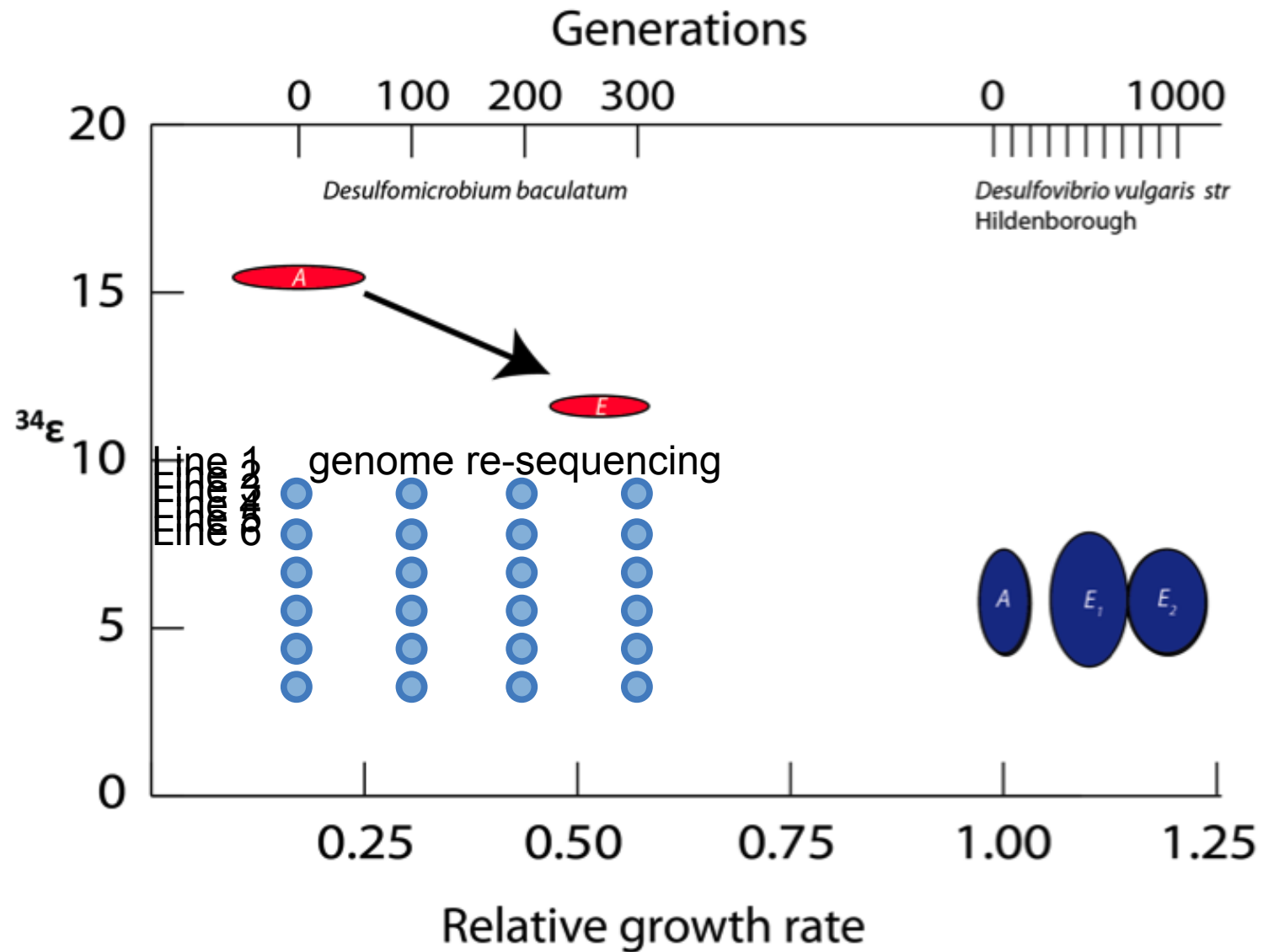
Genome sequencing of sulfate reducing bacteria



Directed Evolution Experiment



Directed Evolution Experiment



λ Upcoming Projects



Barrow 2014

Chukchi 2015?

Cambridge
Bay 2014

Svalbard?

Daneborg 2014

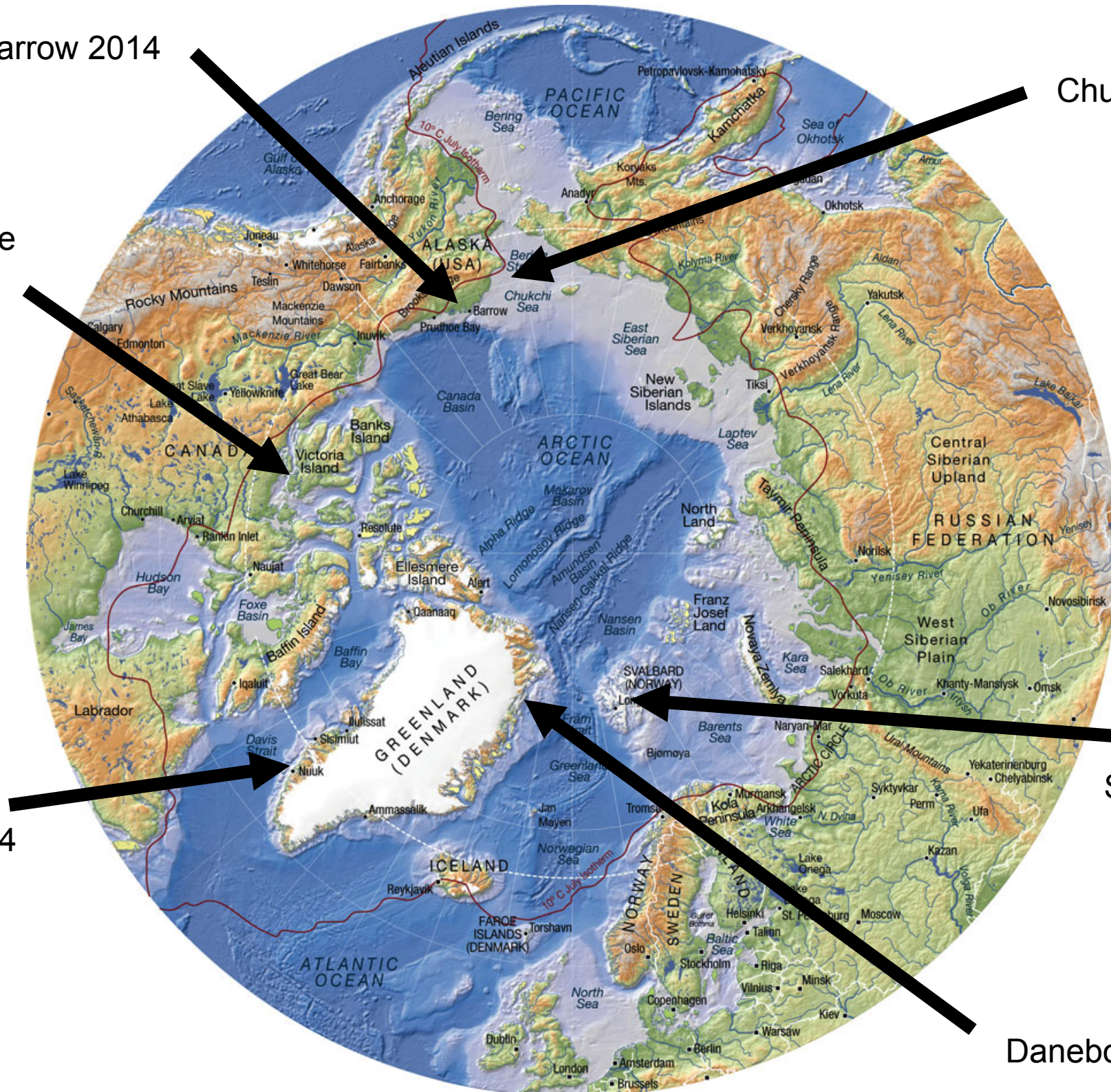




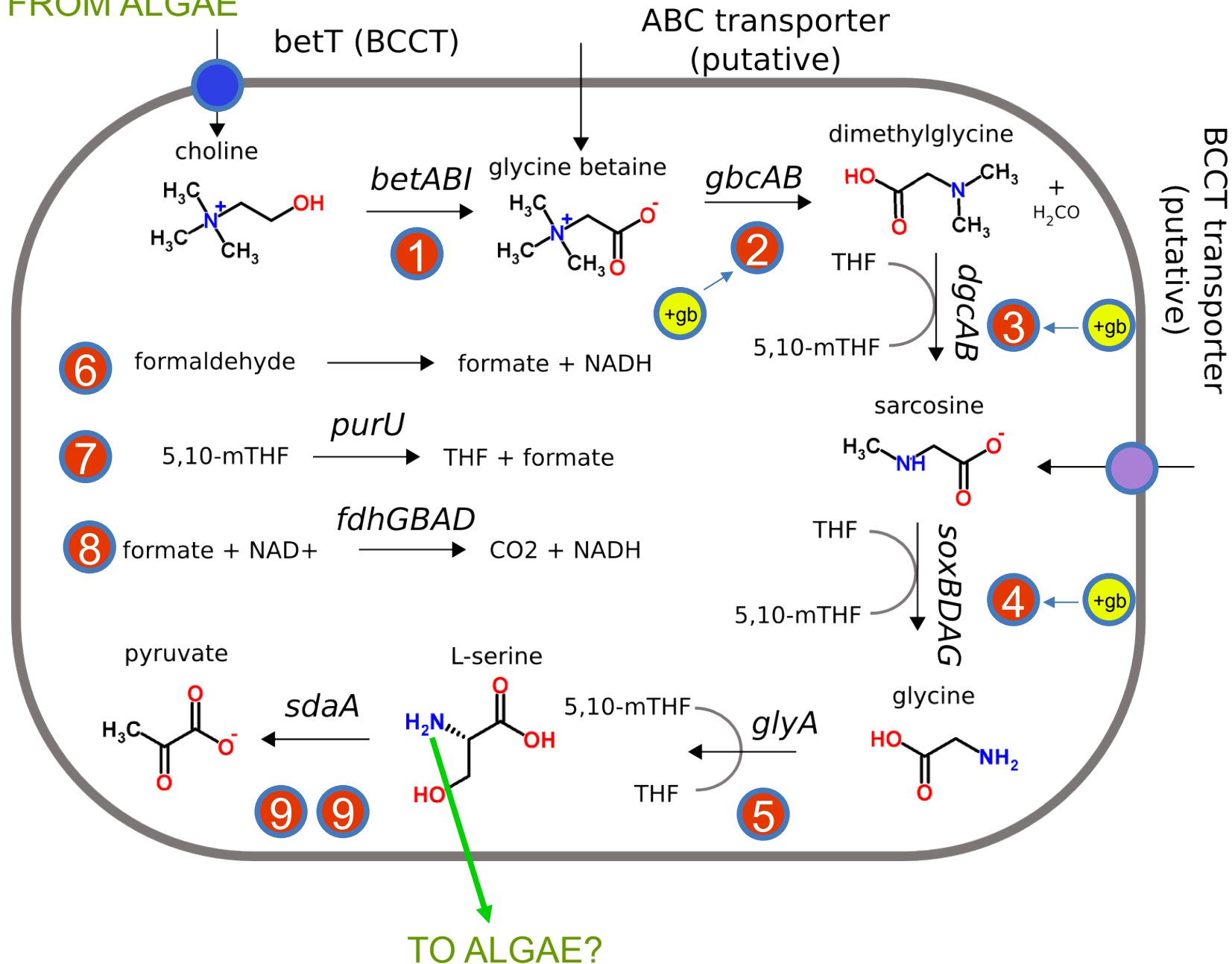




photo by anna

Compatible solute degradation pathway

FROM ALGAE



Thanks!



Jody Deming, Evan Firth & Nuuk team
Rolf Gradinger, Hajo Eicken & sea ice course
Boz Wing, Andre Pellerin & McGill team
Jeff Bowman
Paul Higgs
Funding from NSF OPP, UAF