

Supplementary information

Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells

Jean-François Mangot^{1*}, Ramiro Logares¹, Pablo Sanchez¹, Fran Latorre¹, Yoann Seeleuthner^{2,3,4}, Samuel Mondy^{2,3,4}, Michael E. Sieracki^{5,6}, Olivier Jaillon^{2,3,4}, Patrick Wincker^{2,3,4}, Colomban de Vargas^{7,8}, Ramon Massana^{1*}

Author affiliations

¹ Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)–CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E-08003, Spain.

² CEA, Institut de Génomique, Génoscope, 2 Rue Gaston Crémieux, Evry F-91000, France.

³ CNRS, UMR 8030, CP5706, Evry, F-91000, France.

⁴ Université d'Evry, UMR 8030, CP5706, Evry, F-91000, France.

⁵ National Science Foundation, 4201 Wilson Boulevard, Arlington, VA 22230, USA.

⁶ Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, East Boothbay, ME 04544, USA.

⁷ CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, F-29680, France.

⁸ Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, F-29680, France.

* Correspondence and requests for materials should be addressed to J-F.M. (jean-francois.mangot@wanadoo.fr) or R.M. (ramonm@icm.csic.es).

Tel: (+34) 93 2309500; Fax: (+34) 93 2309555.

Supplementary Tables

Table S1. Main physical-chemical characteristics of the two sampled stations.

Stations	Coordinates	Sampling date	Depth (m)	Temperature (°C)	Oxygen ($\mu\text{mol kg}^{-1}$)	Salinity (psu)	Chlorophyll (mg Chl.m ⁻³)
23	42° 10' 12" N, 17° 43' 12" E	18/11/2009	55	15.7	224.3	38.4	0.06
41	14° 33' 36" N, 70° 0' 36" E	30/03/2010	58	27.1	148.3	36.5	0.47

Table S2. General sampling and sequencing characteristics of the different individual SAGs of MAST-4A and MAST-4E.

Species	SAGs ID	Stations *	Sequencing depth (Gbp)	Sequencing platforms	CV**
MAST-4A	AA538-M19	23	6.9	Hiseq (Genoscope)	
	AA538-N22	23	8.5	Hiseq (Genoscope)	
	AA538-F10	23	6.0	Hiseq (Genoscope)	
	AA538-G04	23	4.7	Hiseq (Genoscope)	
	AA538-G20 [†]	23	4.6	Hiseq (Genoscope)	
	AA538-K07	23	4.0	Hiseq (Genoscope)	
	AA538-E21	23	5.7	Hiseq (Genoscope)	
	AA538-C11	23	2.7	Hiseq (Oregon)	
	AA538-E15	23	6.4	Hiseq (Genoscope)	
	AB537-A17	41	4.0	Hiseq (Oregon)	
	AA538-E19	23	2.4	Hiseq (Oregon)	
	AA538-G20_bis [†]	23	6.8	Hiseq (Oregon)	
	AA538-J18	23	4.8	Hiseq (Genoscope)	
	AB537-K04	41	3.5	Hiseq (Oregon)	
MAST-4E	AA538-A02	23	4.5	Hiseq (Genoscope)	
	AA538-A03	23	4.5	Hiseq (Genoscope)	
	AA538-C05	23	4.6	Hiseq (Genoscope)	
	AA538-F08	23	4.0	Hiseq (Genoscope)	
	AA538-J09	23	4.7	Hiseq (Genoscope)	
	AA538-A11	23	6.8	Hiseq (Genoscope)	
	AA538-L23	23	4.4	Hiseq (Genoscope)	
	AA538-M11	23	4.2	Hiseq (Genoscope)	
	AA538-N16	23	4.9	Hiseq (Genoscope)	
Mean (SE)	all MAST-4A SAGs		5.1 (1.7)		34.0%
	all MAST-4E SAGs		4.7 (0.8)		17.3%
	all SAGs		4.9 (1.4)		28.8%

* Stations 23 and 41 are located in the Mediterranean sea (Adriatic Sea) and Indian Ocean (Arabic Sea), respectively (<http://taraoceans.sb-roscoff.fr/EukDiv/#figureW1>).

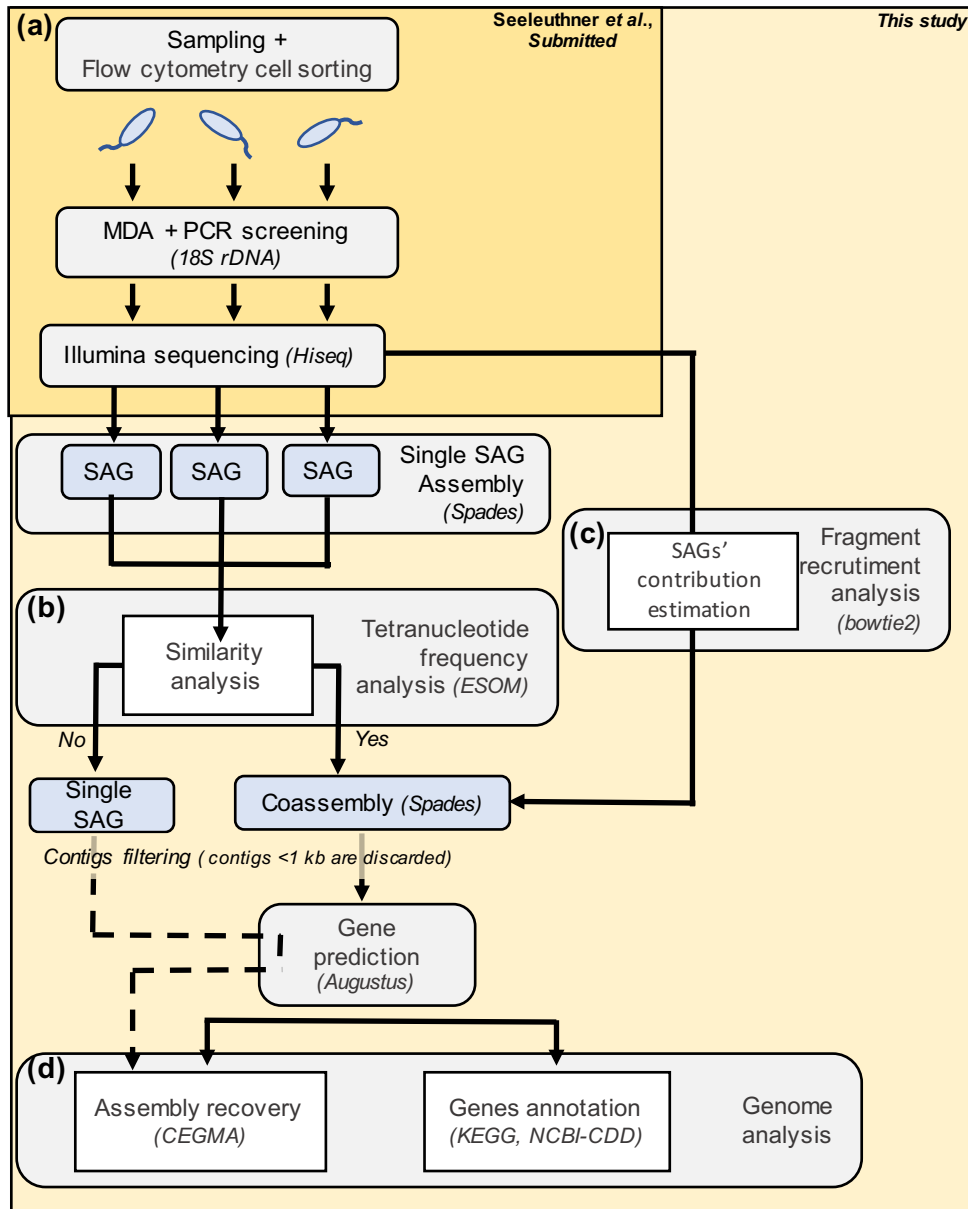
** CV: Coefficient of variation = Standard error/mean.

[†] SAG sequenced by two different sequencing centers, the two sequencing replicates (AA538_G20 and AA538_G20_bis) were kept for further analysis.

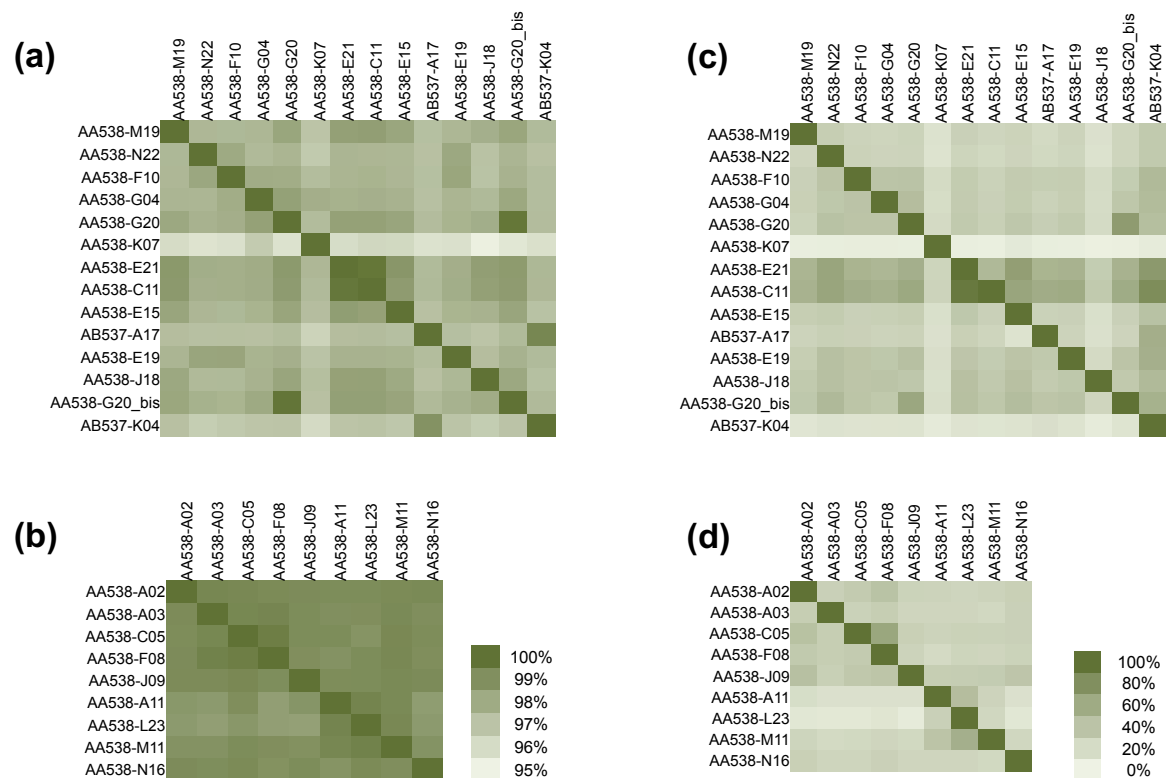
Table S3. General functions present in MAST-4A and MAST-4E genomes based on protein classification according to KOGs of the 248 universal CEGMA eukaryotic genes.

Functioning Process	General Functions	Number of COGs expected	Number of COGs in MAST-4A	Number of COGs in MAST-4E
Information storage and processing	Translation, ribosomal structure and biogenesis	34	29	25
	RNA processing and modification	23	18	16
	Transcription	13	6	9
	Replication, recombination and repair	10	9	7
	Chromatin structure and dynamics	0	0	0
	Shared functions	6	5	4
Cellular processes and signalling	Cell cycle control, cell division, chromosome partitioning	2	2	2
	Nuclear structure	0	0	0
	Defence mechanisms	0	0	0
	Signal transduction mechanisms	4	3	3
	Cell wall/membrane/envelope biogenesis	1	0	0
	Cell motility	0	0	0
	Cytoskeleton	3	2	2
	Extracellular structures	0	0	0
	Intracellular trafficking, secretion, and vesicular transport	17	13	13
	Posttranslational modification, protein turnover, chaperones	42	30	31
Metabolism	Shared functions	6	3	2
	Energy production and conversion	22	13	12
	Carbohydrate transport and metabolism	11	8	9
	Amino acid transport and metabolism	2	2	2
	Nucleotide transport and metabolism	6	4	4
	Coenzyme transport and metabolism	2	2	1
	Lipid transport and metabolism	5	3	2
	Inorganic ion transport and metabolism	3	3	0
	Secondary metabolites biosynthesis, transport and catabolism	1	1	1
Poorly characterized	Shared functions	4	4	4
	General function prediction only	19	15	13
Shared between different functioning process	Function unknown	6	4	3
		6	5	4

Supplementary Figures

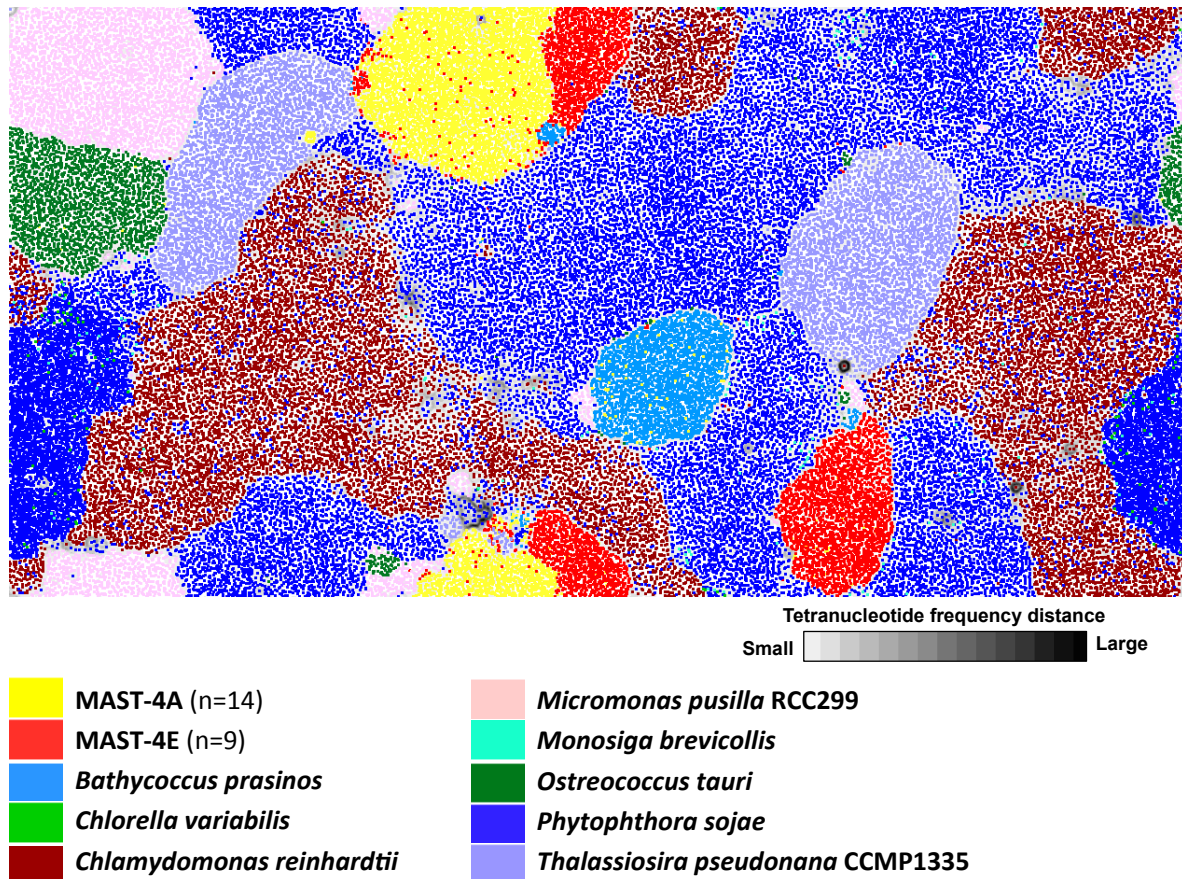


Supplementary Fig. S1. Schematic pipeline of a single-cells co-assembly performed in this study. Details on the sampling, single-cell sorting and SAG sequencing will be available in a concomitant study (Seeleuthner *et al.*, submitted). The rest of the main steps of our co-assembly strategy are described in this paper.



Supplementary Fig. S2. Cross-SAG Blast analysis between MAST-4A and MAST-4E

SAGs. Mean pairwise genomic similarity of MAST-4A **(a)** and MAST-4E **(b)** SAGs are represented, together with the percentage of shared nucleotidic regions for MAST-4A **(c)** and MAST-4E **(d)**. Values derive from blasting full-length contigs of each SAG against full-length contigs of each sister SAG. Query and subject SAGs are listed in the left and top of each heatmap, respectively.



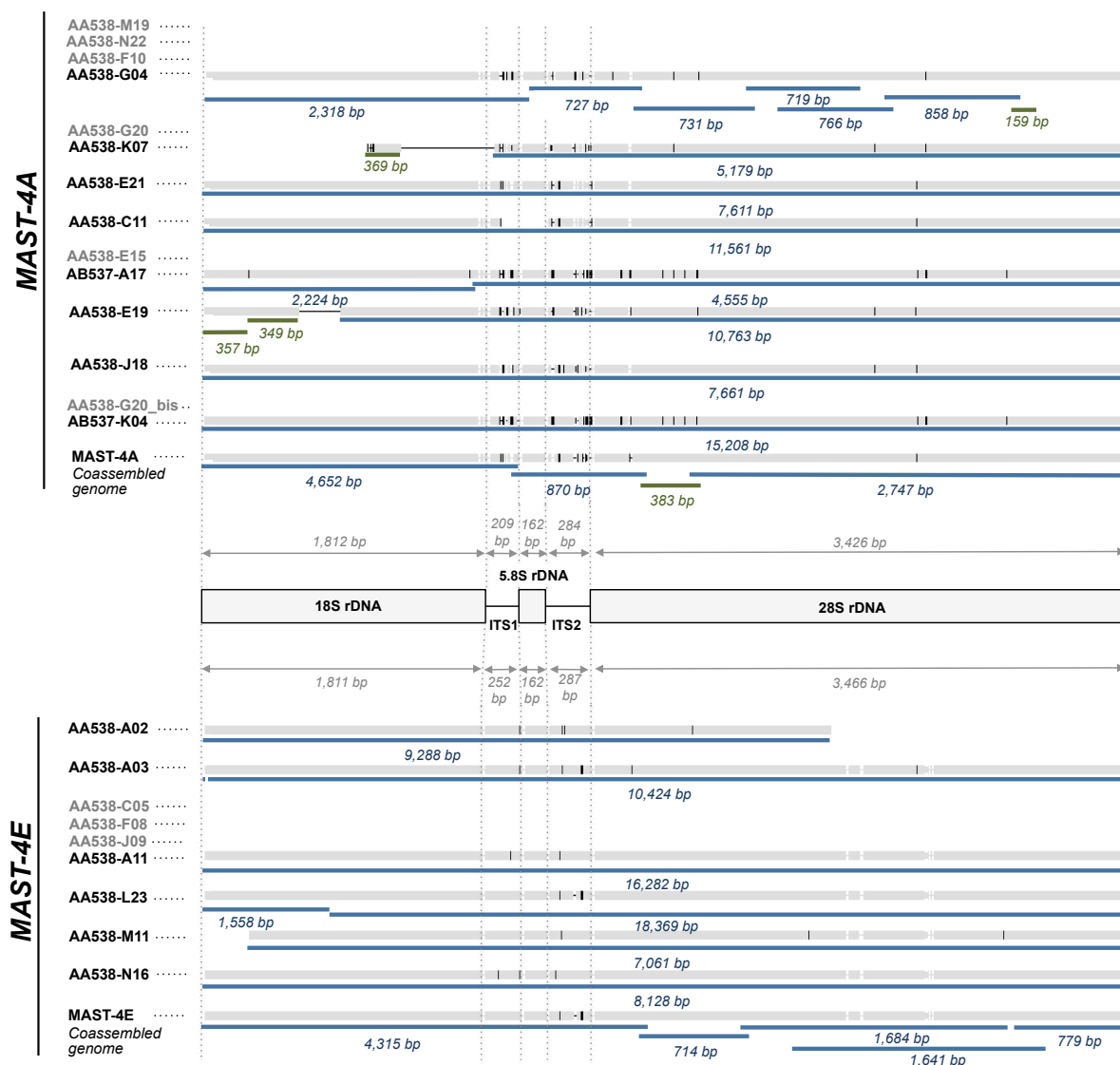
Supplementary Fig. S3. Comparing tetranucleotide frequencies among selected genomes in an ESOM map. Published protist genomes belonging to separate supergroups are combined together with MAST-4A and MAST-4E SAGs. Bestmatches of contigs of 2.5-5 kbp in size are represented by individual points, coloured according to their provenance as MAST-4A (yellow), MAST-4E (red), *Ostreococcus tauri* (dark green), *Micromonas pusilla* (light blue), *Bathycoccus prasinos* (blue), *Chlorella variabilis* (light green), *Chlamydomonas reinhardtii* (dark red), *Thalassiosira pseudonana* (purple), *Phytophthora sojae* (dark blue) and *Monosiga brevicollis* (pink). Large differences in tetranucleotide frequencies represent natural divisions between taxonomic groups.

		MAST-4A										MAST-4E									
		SAGs_ID										SAGs_ID									

[†] Abs : protein sequences of the predicted CEGs were not retrieved in the file generated by CEGMA.

among SAGs (dark grey) or co-assembly (black) are here listed. The mean amino acid

identities of the retrieved CEGs were calculated among SAGs (“SAGs vs SAGs”) and between SAGs and co-assembly (“SAGs vs Coass.”).



Supplementary Fig. S5. Retrieval of the rDNA operon in SAGs of the two MAST-4 lineages. Sequences of MAST-4A (top) and MAST-4E (bottom) individual SAGs containing the rDNA operon were aligned with their corresponding co-assembled genomes. SAGs without rDNA operon contigs are shown in grey. The position and length of contigs from each SAG and co-assemblies necessary to reconstruct the rDNA operon are shown (contigs <500 bp in green and >500 bp in blue). Differences in individual SAGs against the consensus sequence are marked.