Comparative genomics of unculturable marine flagellates (MAST-4) using Single-Amplified Genomes

Fran Latorre[1], Aurélie Labarre[1], Jean-François Mangot[1], Olivier Jaillon[2], Ramon Massana[1] and Ramiro Logares[1]

INSTITUT DE CIÈNCIES DEL MAR
www.icm.csic.es

INTRODUCTION

Marine heterotrophic flagellates (HF) have major importance in the organic carbon transfer to upper trophic levels thanks to their grazing activities on bacteria. Among the HF, Marine Stramenopiles (MASTs) are a highly represented group of organisms constituted by 18 different subgroups. Within the MASTs, there is one group, MAST-4, that due to its small size (2-3 µm), high relative abundance within both MASTs and HFs as well as worldwide distribution, is considered a good model to study marine HF[1].

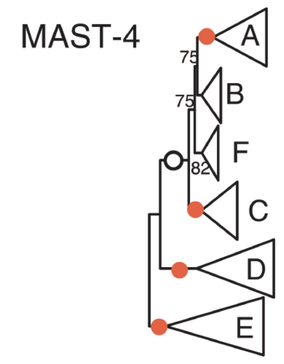

Figure 1. Specific phylogenetic tree of MAST-4 ribogroup, divided into 6 different subclades. (Massana *et al*, 2014)

METHODS

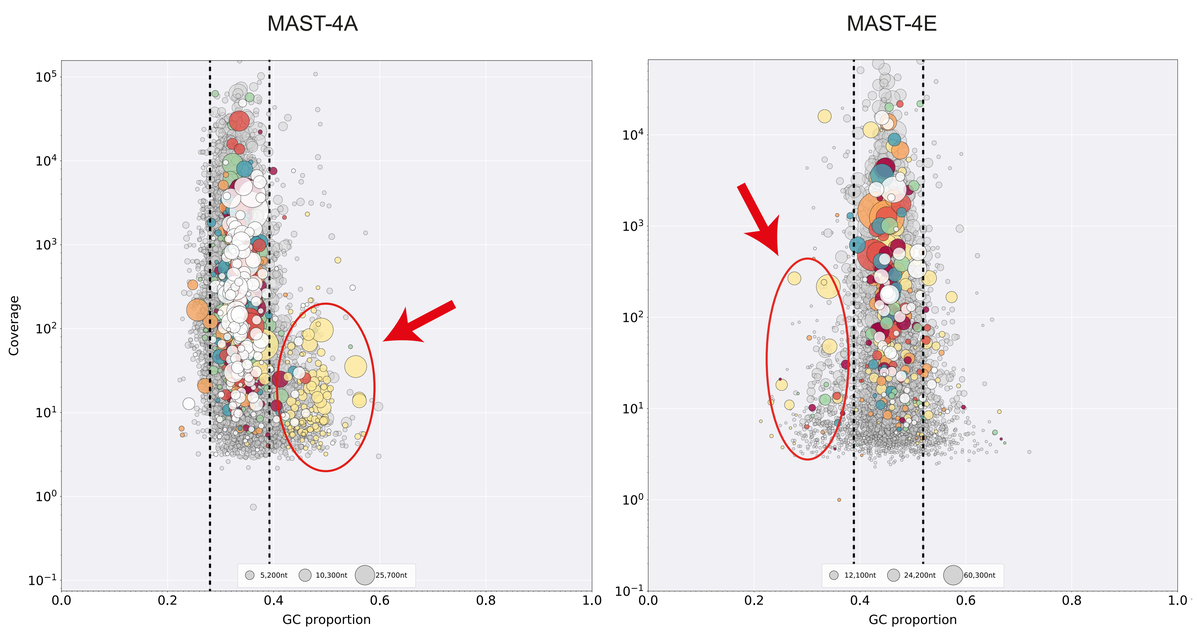

Figure 2. Blobplots. Each point in the plot represents a contig of the respective co-assembly, the length of which is proportional to the point's size. They are colored by taxonomical classification of the NCBI database (not shown for simplicity). Circled in red are the contigs which are assumed to be contaminants, e.g. MAST-4A circle is for *Bathycoccus prasinos* DNA. Blurred lines in black highlight the thresholds used to clean the datasets, every contig outside these regions is discarded.

1. Sample collection; a total of 23 Single Amplified Genomes (SAGs), 14 for MAST-4A and 9 for MAST-4E, were sequenced from Multiple Displacement Amplification reaction (MDA) products of plankton samples collected during the circumglobal Tara Oceans expedition.

2. Assembly and re-scaffolding; orignal Illumina sequencing data were assembled using SPAdes 3.1 and 3.5, discarding contigs shorter than 1,000 bp. MAST-4A and MAST-4E SAGs were Co-assembled to produce higher genome recovery[2]. To improve the quality of the two data-sets a re-scaffolding step was added using SSPACE.

3. Cleaning datasets; due to the grazing activity of the MAST-4 group, it is possible to recover traces of foreign DNA in a SAG. Considering the results of Blobology (Fig. 2) and ESOM (Fig. 3), contigs with a GC content value very different from the mean were removed.

4. Completeness and gene prediction; the proteins recovered from BUSCO v3 to assess the completeness of each clean dataset were used as a training set for AUGUSTUS to make the gene prediction.

5. Single SAG Assemblies (SSA); as an alternative to the co-assembly approach, a gene prediction for each single SAG was performed, and the resulting predicted proteins were merged together and clustered at 95% of identity using Usearch.

6. Functional annotation of predicted genes was done by mapping against the KEGG Orthology database (KO) with BLASTp.

RESULTS

Even though the 18S rDNA phylogeny shows that all subclades within MAST-4 are evolutionary closely related (Fig. 1), we found a significant degree of divergence between them. In particular, we found differences in terms of GC content (33% [A] and 45% [E]) as well as protein and coding sequence composition.

The two clean sets of co-assemblies are composed of 11,146 and 4,060 contigs for MAST-4A and MAST-4E respectively with a 79,9% and 76,6% of genome completeness. From these, a total number of 18,258 and 10,528 proteins were predicted. Comparing the DNA sequences of the predicted genes showed considerable differences between them, which was expected since the mean GC content for each genome is different, agreeing with their tetranucleotide frequencies profiles (Fig. 3). Despite this divergence at the DNA level, amino acid gene-sequences from MA and ME could still be aligned (average gene identity 38%). Functional analyses showed similarities in M4A and M4E composition when compared in terms of Ecological Relevant Genes (ERG) (Table 1).

Gene predictions of SSA recovered almost two times more genes than the co-assemblies (34,020 MAST-4A and 19,992 MAST-4E). Still, neither approach recovered functions more consistently than the other (Table 2); both methods had a few unique predicted genes.

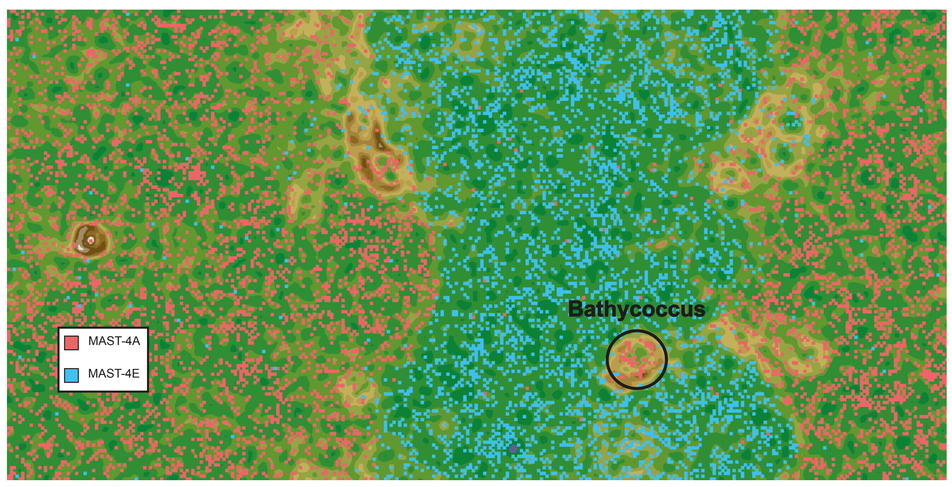

Figure 3. ESOM 2D toroid representation of MAST-4A and MAST-4E contigs clustered by their tetranucleotde frequencies.

CONCLUSIONS

Comparative genomics of MAST-4 clades A and E based on co-assemblies of single-cell MDA-based data from Tara Oceans allow us to start understanding the genomic differences between these poorly known flagellates as well as linking gene composition and ecological role. Specifically, this approach allowed determining the degree of genomic differentiation between clades: functional analyses of both MAST-4 indicated that they are similar in terms of composition of ERG. Nonetheless, genomes are different at the DNA level (that is, MAST-4A/E code differently for similar functions), suggesting that MAST-4 A and E clades are evolutionary more distantly related than previously expected.

Table 1. KOs found in MAST-4A and MAST-4E for pathways of interest.

Pathway	M4A	M4E	Shared	A unique	E unique
Lysosome	49	34	32	17	2
Endocytosis	41	34	29	12	5
Phagosome	22	17	14	8	3
ABC transporters	14	11	9	5	2

Table 2. Number of KOs recovered for each pathway for MAST-4A

Pathway	MAST-4A SSA	MAST-4A CoA
Lysosome	50	49
Endocytosis	38	41
Phagosome	23	22
ABC transporters	13	14

Fran Latorre, latorre@icm.csic.es

[1] **Institut de Ciències del Mar (CSIC),**
Department of Marine Biology and Oceanography,
Pg. Marítim de la Barceloneta, 37-49,
08003, Barcelona, Spain.
[2] **CEA-Institut de Genomique, Genoscope,**
Centre National de Sequencage,
Evry Cedex, France.

References

[1.] Massana, R *et a*l. Exploring the uncultured microeykaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. ISME J. 8, 854-66 (2014).
[2.] Mangot, J.-F. *et al*. Approaching complete genomes of uncultured picoeukaryotes by combining sequences from several single cells. (2017).