A first large scale characterization of eukaryotic unknown genes of plankton

Olivier Jaillon [1,2,3], Thomas Vannier[1,2,3], *Tara* Oceans coordinators & Patrick Wincker[1,2,3]

[1] CEA-Institut de Génomique, GENOSCOPE, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706, 91057 Evry Cedex, France.

[2] Université d'Evry, UMR 8030, CP5706, 91057 Evry Cedex, France.

[3] Centre National de la Recherche Scientifique (CNRS), UMR 8030, CP5706, 91057 Evry Cedex, France.

The extent and the diversity of the repertoire of eukaryotic genes remain undetermined as most gene sequences in the public databases are from animals, plants and fungi phyla which represent only a tiny fraction of the eukaryotic diversity. Because life likely evolved in the ocean, and because highly diverse representatives of almost all eukaryotic lineages abound in oceanic plankton, this community may provide an estimate of the extent of eukaryotic gene repertoire.

Here, a size fractionation of plankton samples coupled with deep sequencing allowed us to analyze by unbiased metagenomics three Mediterranean sites of the *Tara* oceans expedition, each spanning the size range from bacteria to small metazoans. We obtained 4.2 Gb of assembled sequences, and annotated 7.6 million genes. One third of these genes, named *unknown genes* bear no similarity, even distantly, to known sequences. Furthermore, we obtained Single Amplified Genomes (SAGs) which together with ribosomal marker sequences enable to decipher the likely phylogenetic origin of *unknown genes* and to analyze their dynamics across different marine zones. These analyses confirm that organisms holding *unknown genes* likely belong to abundant and poorly explored taxa (MAST, rhizaria, alveolata). We present evidences from metagenomic and metatranscriptomic data from all oceans of their transcriptional activities and analyses of their conservation between divergent strains.