Identifying losses and expansions of selected gene families in incomplete genomic datasets

Arnaud Di Franco[1,2], **Marc Hanikenne**[2,3], **Denis Baurain**[1,2]

[1]Eukaryotic Phylogenomics, Dept of Life Sciences, University of Liège, B-4000 Liège, Belgium
[2]PhytoSYSTEMS, University of Liège, B-4000 Liège, Belgium
[3]Functional Genomics and Plant Molecular Imaging, Center for Protein Engineering (CIP), Dept of Life Sciences, University of Liège, B-4000 Liège, Belgium

Introduction

DURING evolution, some eukaryotic lineages have acquired the ability to photosynthesize through endosymbiosis and conversion of the endosymbiont into a plastid. The first endosymbiotic event has involved a phagotrophic eukaryote and a cyanobacterium and has given rise to three monophyletic groups: glaucophytes, green plants and red algae. Red algae have generally smaller genomes with less coding genes compared to the two others (especially green plants). Because the earliest-branching group among red algae (Cyanidiales) is composed of thermoacidophilic organisms, we hypothesize that these missing genes might have been lost by the common ancestor of extant red algal during its adaptation to life in extremophilic conditions.

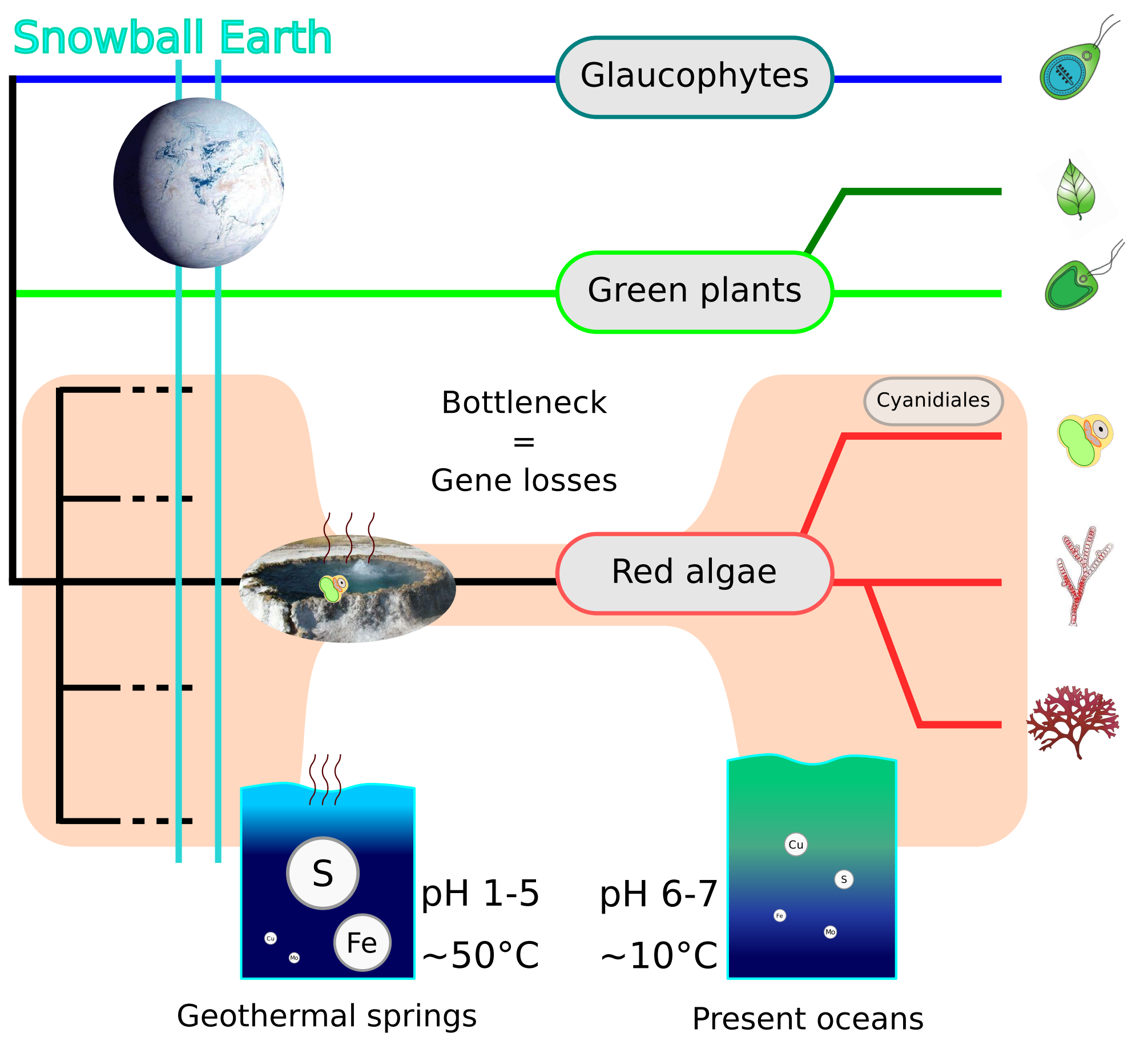

Figure 1: ***Hypothesis of the bottleneck in red algal evolution.*** *Due to a past environmental perturbation (e.g., Snowball Earth), red algal diversity could have been reduced to extremophilic organisms. The adaptation of red algae to the peculiar thermoacidophilic environment would have led to massive losses of unnecessary genes. We speculate that such an hypothesis might be demonstrated by observing specific gene losses shared by current red algae.*

Results

USING the pipeline described to the right, we examined in red algae the conservation of thousands of gene families defined in green plants, among which about 200 families corresponding to manually selected genes related to differential metabolism in geothermal springs and present oceans. We found that several ancestral gene families were absent in all available red algal genomes (family losses), whereas, for other gene families, the number of genes composing the family had been increased in red algal genomes (family expansion). These observations carried out on our inventory were compared to those obtained from all gene families.

Family losses

IN figure 2, we compared the losses in our inventory and in all gene families. The Venn diagrams show that the losses in each group are proportionally equivalent. This observation can also be made in the bar chart, even if the latter points out that losses are less abundant among the genes of our inventory. This suggests that we selected genes (and gene families) that are generally more conserved in red algae than the average gene (or gene family).

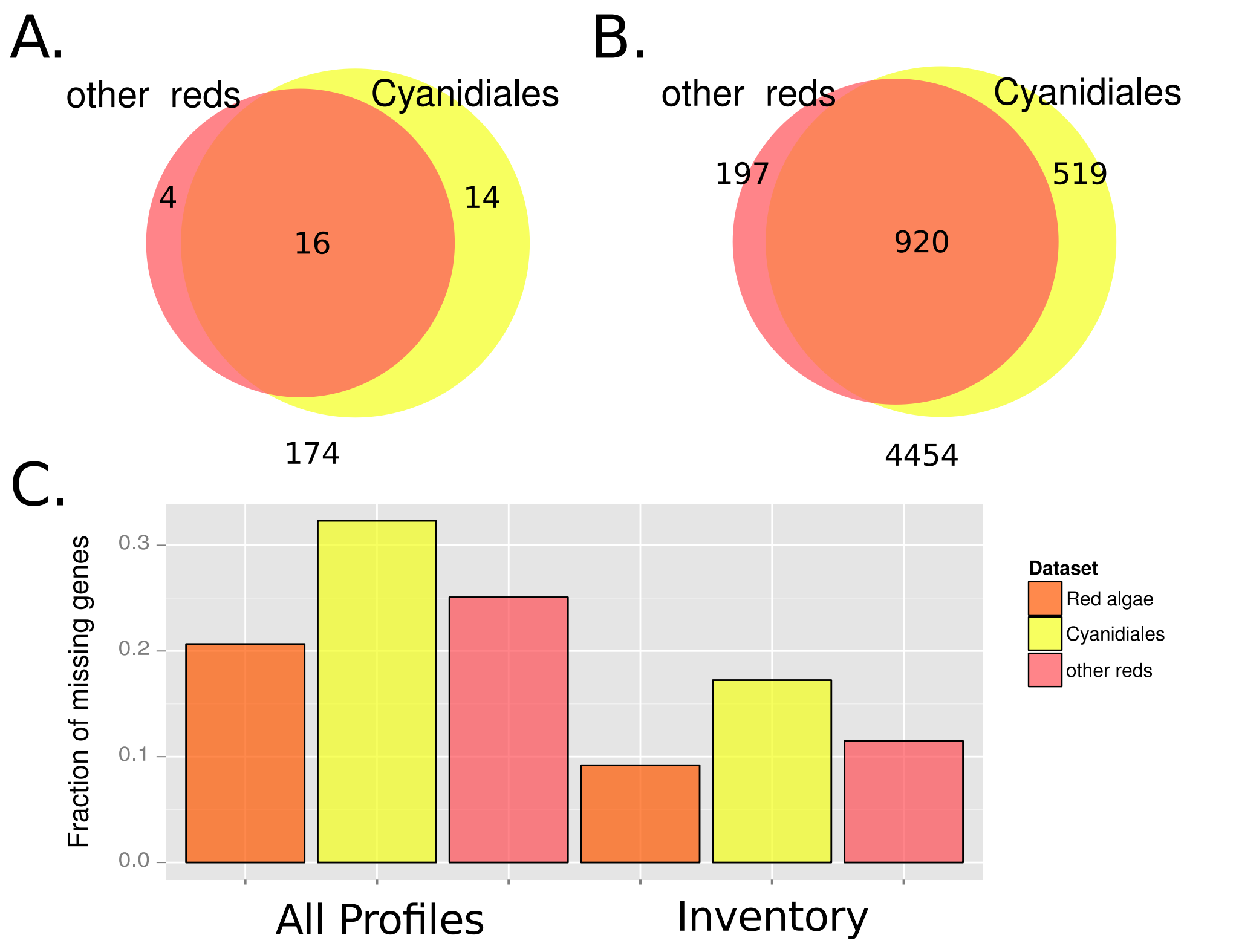

Figure 2: ***Comparison of gene family losses between the set of genes in our inventory and in all gene families.*** *Shared losses were studied in two groups of red algal species:* **Cyanidiales** *and* **other reds**. *Cyanidiales are composed of two species, Cyanidioschyzon merolae and Galdieria sulphuraria, whereas other reds include three species, Porphyridium purpureum, Pyropia yezoensis and Chondrus crispus. Therefore, genes lost in both groups are absent from all red algal genomes currently available. A. Venn diagram of gene family losses in our inventory. B. Venn diagram of gene family losses in all gene families. C. Bar chart comparing the fraction of missing genes in our inventory and in all gene families. Regarding our inventory, 16 gene families are absent from all red algal genomes (among which APG7, COPT1, CTP2, CUTC, CYP737A1, FPN1, MDHM, MOT1, PKL1, SOUL1, THI1), whereas 14 more gene families are missing only in Cyanidiales (among which ARD1, CDJ3, CYC4, FSD1, NAR1.1, LciA, NCR2, PAO1, TIC55, pfl-AE, Sac3) for only 4 gene families in other reds (COX17, CYP746A1, SULTR2, SULTR3).*

Family expansions

WE determined gene family expansions by comparing the size of gene families in green plants and red algae. In figure 3, the comparison is based on the following ratio: $e_{p,cy} = \log_{10}\left(\frac{c_{p,cy}}{c_{p,gp}}\right)$ where $c_{p,g} = \frac{\#\ \text{orthologs for profile } p}{\#\ \text{genomes for group } g}$. $e_{p,cy} > 0$ indicates a more abundant family in red algae (here, Cyanidiales) relative to green plants, whereas $e_{p,cy} < 0$ suggests the opposite. The different plots show that our inventory is enriched in expanded families and also in really small red algal gene families compared to green plants. In figure 4, we observe that gene families related to copper and iron transport seem to be often expanded in non-Cyanidiales (other) red algae.

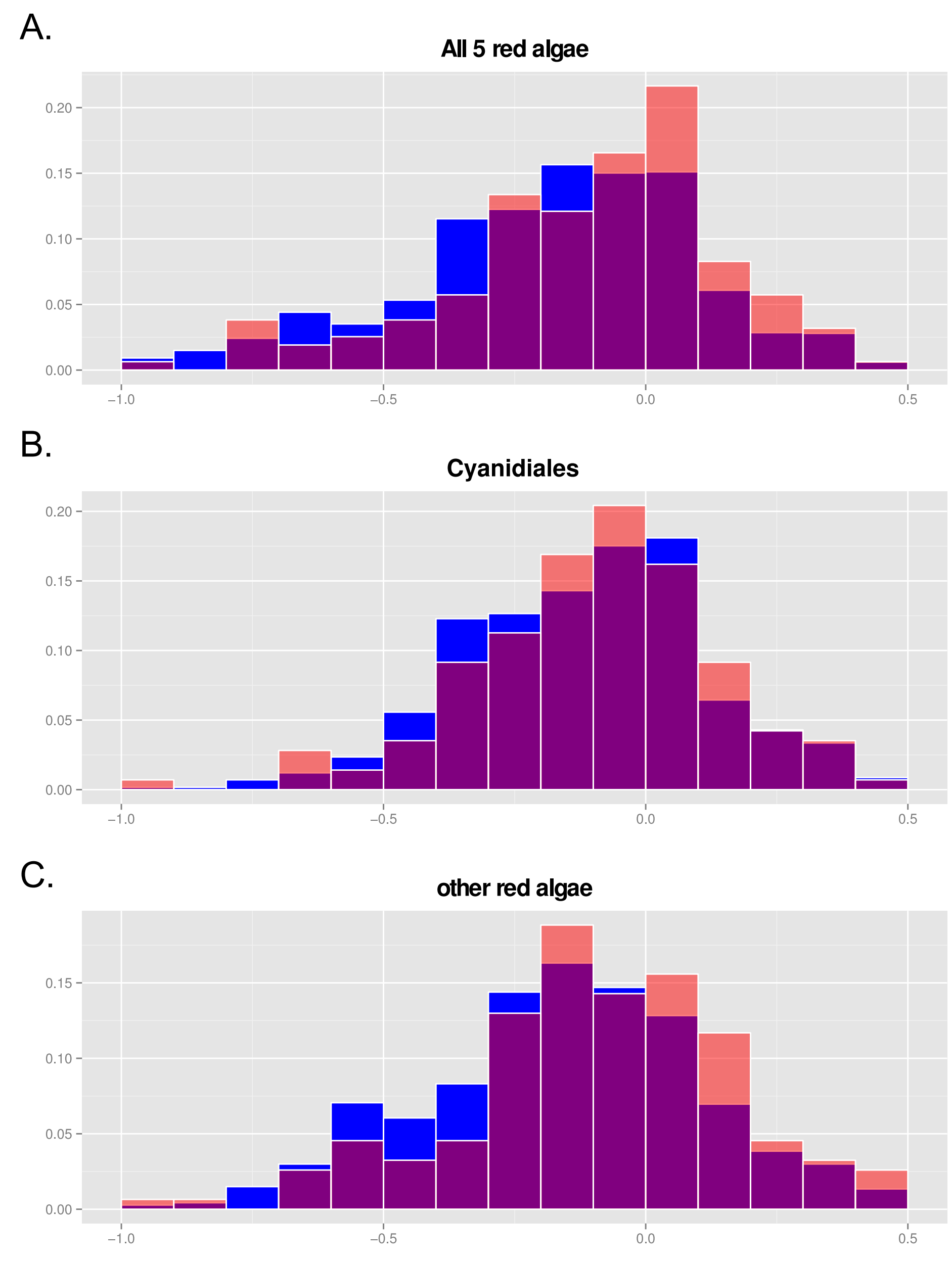

Figure 3: ***Comparison of gene family sizes between green plants and red algae.*** *The distribution of the $e_{p,group}$ ratio for all gene families is depicted by the blue bar chart, while the distribution for the genes of our inventory is overlaid in red. A. $e_{p,all}$ (all five red algae). B. $e_{p,cy}$ (two Cyanidiales). C. $e_{p,oth}$ (three other red algae).*

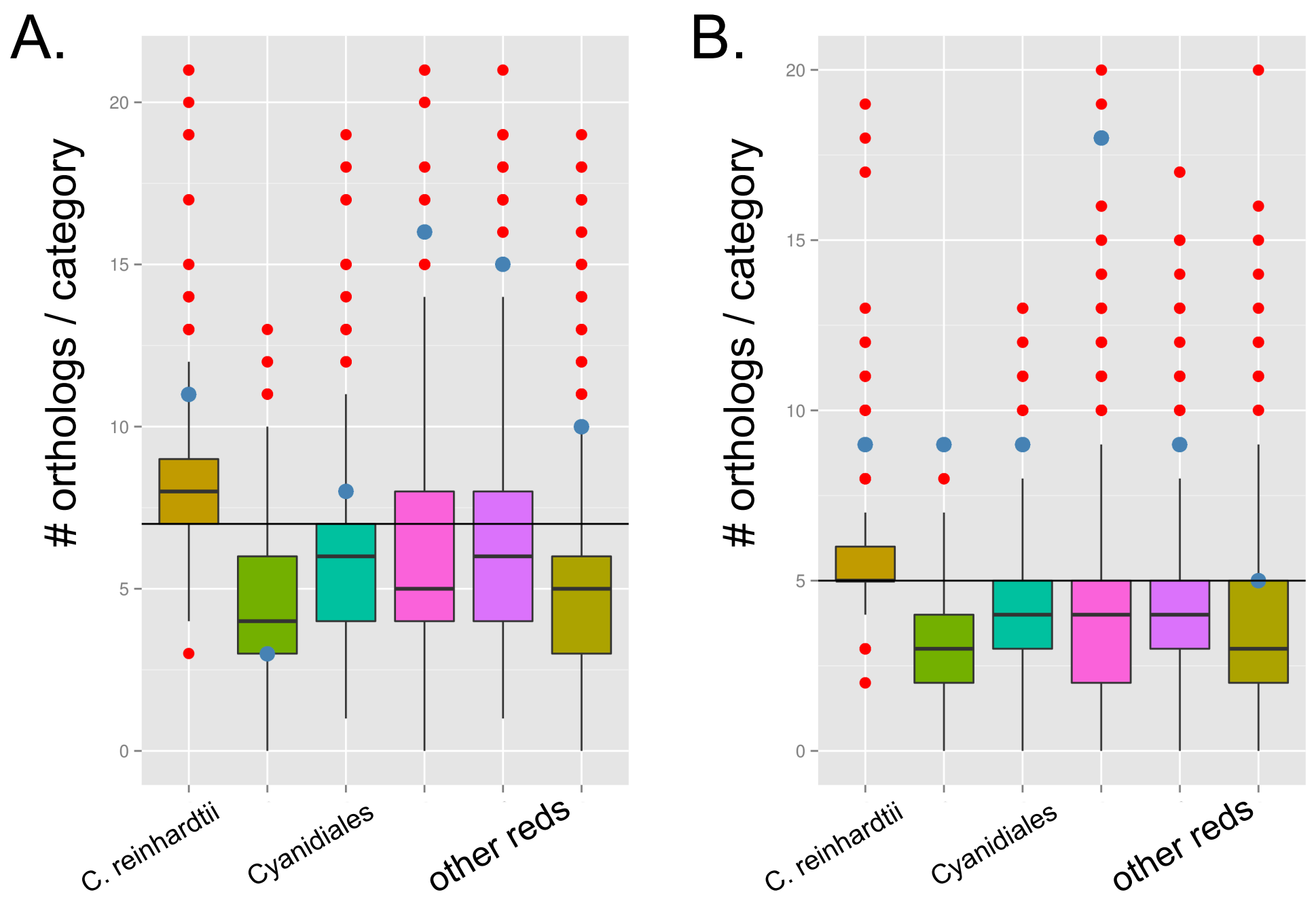

Figure 4: ***Gene family expansions for copper and iron transport-related gene families.*** *For each of the two categories, the corresponding set of pHMMs defined on green plants was selected and the total count of orthologs plotted for each red algal species (blue dot). The box plot summarizes the distribution of control counts obtained from 1000 random sets of pHMMs mimicking the statistical properties of the real sets (see figure 6.F). A. Copper-transport gene families. B. Iron-transport gene families.*

Example

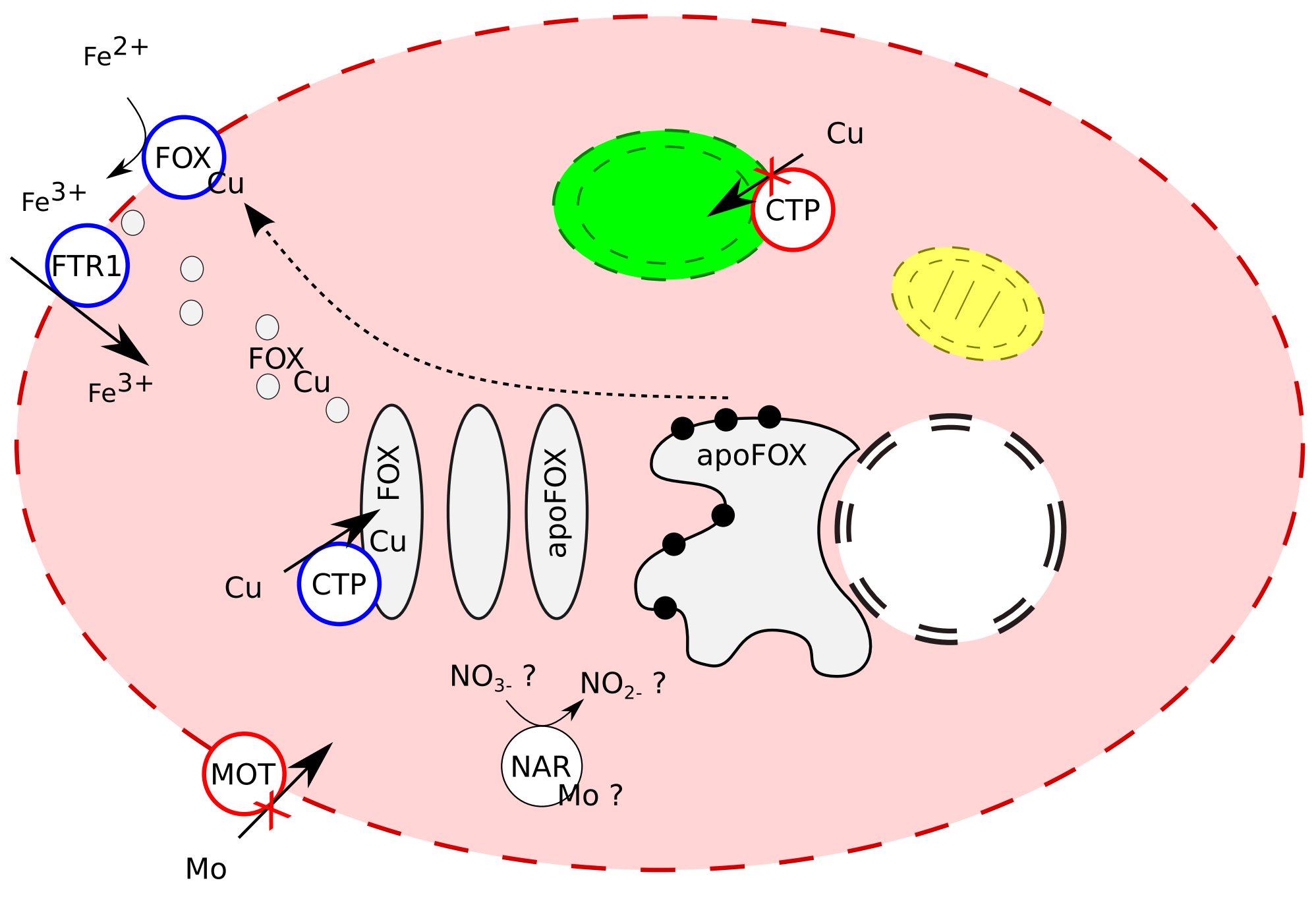

Figure 5: ***Main expansions and losses in Pyropia yezoensis.*** *We observed that the genes of three expanded families interact for a better assimilation of iron. CTP1 and CTP3 transport copper into Golgi vesicles. FOX1 and FOX2 are copper-dependent ferroxidases located in the plasma membrane. FRE1 is an iron permease also located in the plasma membrane. In contrast, the specific molybdenum transporter MOT1 (regulated by nitrate) is absent, as well as CTP2, another copper transporter related to CTP1/CTP3 and located in the plastid membrane.*

Methods

WE computed orthology relationships between green plant genes and red algal genes using a combination of Hidden Markov Model profiles (pHMMs) and (BLAST) Best Reciprocal Hit criterion (BRH). These relationships were uploaded into a SQL database to facilitate downstream analyses. Querying of the database was carried out directly from R to take advantage of the statistic and graphical functions of the latter.

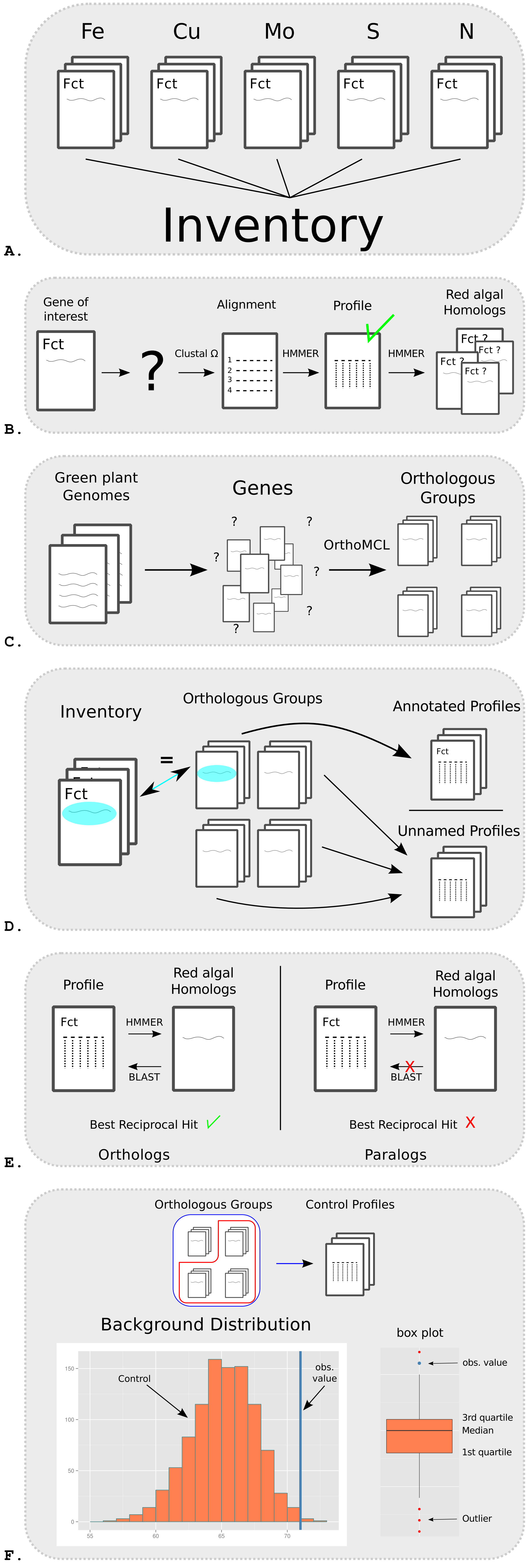

Figure 6: ***Pipeline.*** *A. An inventory of genes related to copper, iron, molybdenum, nitrogen and sulphur is established using Chlamydomonas reinhardtii as a genomic source. B. Red algal homologs are identified by similarity to pHMMs built from green plant orthologs. C. To this end, orthologous groups (OGs) are computed for ten genomes of green plants using OrthoMCL. D. OGs are annotated by the genes of our inventory using BLAST. E. OGs are aligned, converted to pHMMs and used to identify red algal homologs with HMMER3. Orthologs are separated from paralogs using reciprocal BLAST searches and the BRH criterion. F. Unknown pHMMs and annotated pHMMs are merged into a pool of control pHMMs used to generate a background distribution for the observed counts.*

BBC2013, the eighth Benelux Bioinformatics Conference, Brussels, December 9--10, 2013